

# Proceedings of the XVIII EURALEX International Congress

## Lexicography in Global Contexts

17-21 July 2018, Ljubljana

Edited by Jaka Čibej, Vojko Gorjanc,  
Iztok Kosem and Simon Krek

**EURALEX** 



Univerza v Ljubljani  
**FILOZOFSKA  
FAKULTETA**

# Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts

Edited by: Jaka Čižej, Vojko Gorjanc, Iztok Kosem and Simon Krek

Reviewers: Andrea Abel, Zoe Gavriilidou, Robert Lew and Tinatin Margalitadze

English language proofreading: Paul Steed

Technical editor: Aleš Cimprič



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press, Faculty of Arts

Issued by: University of Ljubljana, Centre for language resources and technologies

For the publisher: Roman Kuhar, Dean of the Faculty of Arts, University of Ljubljana

Ljubljana, 2018

First edition, e-edition

Publication is free of charge.

The editors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0215).

This text was written using the ZRCola input system (ZRCola.zrc-sazu.si), developed at the Research Centre of the Slovenian Academy of Sciences and Arts in Ljubljana (www.zrc.sazu.si) by Peter Weiss.

DOI: 10.4312/9789610600961

E-book is available at: <https://e-knjige.ff.uni-lj.si/>

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani

[COBISS.SI-ID=24619494](https://nuk.ub.uni-lj.si/COBISS.SI-ID=24619494)

ISBN 978-961-06-0097-8 (epub)

ISBN 978-961-06-0096-1 (pdf)



## Acknowledgements

We would like to thank all those who have made the XVIII EURALEX International Congress possible, by contributing to the reviewing, to the logistics and by financially supporting the event. In particular, we would like to thank our sponsoring partners and patrons:

A.S. Hornby Educational Trust

Ingenierie Diffusion Multimedia Inc.

Oxford University Press

ELEXIS – European Lexicographic Infrastructure

CLARIN.SI – Common Language Resources and Technology Infrastructure, Slovenia

Alpineon d.o.o.

DELIGHT d.o.o.

TshwaneDJe

Faculty of Arts, University of Ljubljana

## Programme Committee

Andrea Abel (European Academy of Bozen/Bolzano, EURAC)

Polona Gantar (University of Ljubljana, Faculty of Arts)

Zoe Gavriilidou (Democritus University of Thrace, Xanthi)

Vojko Gorjanc (University of Ljubljana, Faculty of Arts)

Iztok Kosem (University of Ljubljana, Faculty of Arts / Trojina)

Simon Krek (Chair) (University of Ljubljana, Center for Language Resources and Technologies / Jožef Stefan Institute, Artificial Intelligence Laboratory)

Robert Lew (Adam Mickiewicz University in Poznań, Faculty of English)

Tinatin Margalidze (Ivane Javakhishvili Tbilisi State University)

## Reviewers

Adam Rambousek, Agáta Karčová, Agnes Tutin, Ales Horak, Alexander Geyken, Amália Mendes. Amanda Laugesen, Andrea Abel, Anne Dykstra, Annette Klosa, Antton Gurrutxaga, Arvi Tavast, Carla Marello, Carole Tiberius, Carolin Müller-Spitzer, Chris Mulhall, Christine Moehrs, Corina Forascu, Danie Prinsloo, Edward Finegan, Egon Stemle, Elena Volodina, Francesca Frontini, Francis Bond, Frieda Steurs, Geoffrey Williams, Henrik Lorentzen, Ilan Kernerman, Iztok Kosem, Janet Decesaris, Jelena Kallas, Jette Hedegaard Kristoffersen, John McCrae, Jorge Gracia, Julia Bosque-Gil, Julia Miller, Klaas Ruppel, Kristina Strkalj Despot, Kseniya Egorova, Lars Trap-Jensen, Lionel Nicolas, Lothar Lemnitzer, Lut Colman, Magali Paquot, Margit Langemets, Maria Khokhlova, Marie-Claude L’Homme, Michal Měchura, Michal Kren, Miloš Jakubiček, Monica Monachini, Nataša Logar Berginc, Nicoletta Calzolari, Nikola Ljubešić, Nora Aranberri, Oddrun Grønvik,

Orin Hargraves, Orion Montoya, Patrick Hanks, Patrick Drouin, Paul Cook, Paz Battaner, Philipp Cimiano, Pilar León Araúz, Piotr Zmigrodzki, Pius ten Hacken, Polona Gantar, Radovan Garabík, Robert Lew, Roberto Navigli, Ruben Urizar, Rufus Gouws, Sass Bálint, Sara Može, Simon Krek, Stella Markantonatou, Svetla Koeva, Sylviane Granger, Špela Arhar Holdt, Tamás Váradi, Tanneke Schoonheim, Tatjana Gornostaja, Thierry Fontenelle, Tinatin Margalitadze, Tomaž Erjavec, Ulrich Heid, Valentina Apresjan, Vincent Ooi, Vojko Gorjanc, Xabier Artola Zubillaga, Xabier Saralegi, Yongwei Gao, Yukio Tono, Zoe Gavriilidou







# Contents

<b>Foreword</b>	<b>13</b>
<b>PLENARY LECTURES</b>	<b>15</b>
Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution? <i>Sylviane Granger</i>	17
Lexicography between NLP and Linguistics: Aspects of Theory and Practice <i>Lars Trap-Jensen</i>	25
<b>PAPERS</b>	<b>39</b>
<b>RESEARCH INTO DICTIONARY USE</b>	<b>41</b>
Investigating the Dictionary Use Strategies of Greek-speaking Pupils <i>Elina Chadjipapa</i>	43
Everything You Always Wanted to Know about Dictionaries (But Were Afraid to Ask): A Massive Open Online Course <i>Sharon Creese, Barbara McGillivray, Hilary Nesi, Michael Rundell, Katalin Sule</i>	59
Researching Dictionary Needs of Language Users Through Social Media: A Semi-Automatic Approach <i>Jaka Čibej, Špela Arhar Holdt</i>	67
The DHmine Dictionary Work-flow: Creating a Knowledge-based Author's Dictionary <i>Tamás Mészáros, Margit Kiss</i>	77
Analyzing User Behavior with Matomo in the Online Information System Grammis <i>Saskia Ripp, Stefan Falke</i>	87
Combining Quantitative and Qualitative Methods in a Study on Dictionary Use <i>Sascha Wolfer, Martina Nied Curcio, Idalete Maria Silva Dias, Carolin Müller-Spitzer, María José Domínguez Vázquez<sup>4</sup></i>	101
<b>DICTIONARY-MAKING PROCESS</b>	<b>113</b>
Nathanaël Duez lexicographe : l'art de (re)travailler les sources <i>Antonella AmatuZZi</i>	115
A Workflow for Supplementing a Latvian-English Dictionary with Data from Parallel Corpora and a Reversed English-Latvian Dictionary <i>Daiga Dekšne, Andrejs Veisbergs</i>	127
Towards a Representation of Citations in Linked Data Lexical Resources <i>Anas Fahad Khan, Federico Boschetti</i>	137
The Sounds of a Dictionary: Description of Onomatopoeic Words in the Academic Dictionary of Contemporary Czech <i>Magdalena Kroupová, Barbora Štěpánková, Veronika Vodrážková</i>	149
Comparing Orthographies in Space and Time through Lexicographic Resources <i>Christian-Emil Smith Ore, Oddrun Grønvik</i>	159
A Universal Classification of Lexical Categories and Grammatical Distinctions for Lexicographic and Processing Purposes <i>Roser Saurí, Ashleigh Alderslade, Richard Shapiro</i>	173

<b>Commonly Confused Words in Contrastive and Dynamic Dictionary Entries</b> <i>Petra Storjohann</i>	187
<b>Slovenian Lexicographers at Work</b> <i>Alenka Vrbinc, Donna M. T. Cr. Farina, Marjeta Vrbinc</i>	199
<b>Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish</b> <i>Piotr Żmigrodzki</i>	209
<b>LEXICOGRAPHICAL PROJECTS AND PHRASEOLOGY</b>	221
<b>Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement</b> <i>Michal Boleslav Měchura</i>	223
<b>A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined</b> <i>Lut Colman, Carole Tiberius</i>	233
<b>ColloCaid: A Real-time Tool to Help Academic Writers with English Collocations</b> <i>Robert Lew, Ana Frankenberg-Garcia, Geraint Paul Rees, Jonathan C. Roberts, Nirwan Sharma</i>	247
<b>Looking for a Needle in a Haystack: Semi-automatic Creation of a Latvian Multi-word Dictionary from Small Monolingual Corpora</b> <i>Inguna Skadiņa</i>	255
<b>TERMINOLOGY, TERMINOGRAPHY AND SPECIALISED LEXICOGRAPHY</b>	267
<b>Semantic-based Retrieval of Complex Nominals in Terminographic Resources</b> <i>Melania Cabezas-García, Juan Carlos Gil-Berrozpe</i>	269
<b>Towards a Glossary of Rum Making and Rum Tasting</b> <i>Cristiano Furiassi</i>	283
<b>Russian Borrowings in Greek and Their Presence in Two Greek Dictionaries</b> <i>Zoe Gavriilidou</i>	297
<b>Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary</b> <i>Laura Giacomini</i>	309
<b>Dictionaries of Linguistics and Communication Science / Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK)</b> <i>Stefan J. Schierholz</i>	319
<b>When Learners Produce Specialized L2 Texts: Specialized Lexicography between Communication and Knowledge</b> <i>Patrick Leroyer, Henrik Køhler Simonsen</i>	329
<b>New Platform for Georgian Online Terminological Dictionaries and Multilingual Dictionary Management System</b> <i>Tinatin Margalitadze</i>	339
<b>Building a Portuguese Oenological Dictionary: from Corpus to Terminology via Co-occurrence Networks</b> <i>William Martinez, Sílvia Barbosa</i>	351
<b>Using Diachronic Corpora of Scientific Journal Articles for Complementing English Corpus-based Dictionaries and Lexicographical Resources for Specialized Languages</b> <i>Katrin Menzel</i>	363

<b>ELeFyS: A Greek Illustrated Science Dictionary for School</b>	<b>373</b>
<i>Maria Mitsiaki, Ioannis Lefkos</i>	
<b>Terms Embraced by the General Public: How to Cope with Determinologization in the Dictionary?</b>	<b>387</b>
<i>Jana Nová</i>	
<b>REPORTS ON LEXICOGRAPHICAL PROJECTS</b>	<b>399</b>
<b>Thesaurus of Modern Slovene: By the Community for the Community</b>	<b>401</b>
<i>Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja</i>	
<b>Dictionary of Verbal Contexts for the Romanian Language</b>	<b>411</b>
<i>Ana-Maria Barbu</i>	
<b>A Sample French-Serbian Dictionary Entry based on the ParCoLab Parallel Corpus</b>	<b>423</b>
<i>Saša Marjanović, Dejan Stosic, Aleksandra Miletic</i>	
<b>HISTORICAL LEXICOGRAPHY, ETYMOLOGY</b>	<b>437</b>
<b>Lexicography in the Eighteenth-century Gran Chaco: the Old Zamuco Dictionary by Ignace Chomé</b>	<b>439</b>
<i>Luca Ciucci</i>	
<b>Historical Corpus and Historical Dictionary: Merging Two Ongoing Projects of Old French by Integrating their Editing Systems</b>	<b>453</b>
<i>Sabine Tittel</i>	
<b>Heritage Dictionaries, Historical Corpora and other Sources: Essential And Negligible Information</b>	<b>467</b>
<i>Alina Villalva</i>	
<b>SIGN LANGUAGE LEXICOGRAPHY</b>	<b>481</b>
<b>Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How</b>	<b>483</b>
<i>Gabriele Langer, Anke Müller, Sabrina Wähl, Julian Bleicken</i>	
<b>Multimodal Corpus Lexicography: Compiling a Corpus-based Bilingual Modern Greek – Greek Sign Language Dictionary</b>	<b>499</b>
<i>Anna Vacalopoulou, Eleni Efthimiou, Kiki Vasilaki</i>	
<b>PHRASEOLOGY AND COLLOCATION</b>	<b>509</b>
<b>Bilingual Corpus Lexicography: New English-Russian Dictionary of Idioms</b>	<b>511</b>
<i>Guzel Gizatova</i>	
<b>Computer-aided Analysis of Idiom Modifications in German</b>	<b>523</b>
<i>Elena Krotova</i>	
<b>NEOLOGISMS</b>	<b>533</b>
<b>On the Detection of Neologism Candidates as a Basis for Language Observation and Lexicographic Endeavors: the STyrLogism Project</b>	<b>535</b>
<i>Andrea Abel, Egon W. Stemle</i>	
<b>Neologisms in Online British-English versus American-English Dictionaries</b>	<b>545</b>
<i>Sharon Creese</i>	

<b>New German Words: Detection and Description</b>	<b>559</b>
<i>Annette Klosa, Harald Lungen</i>	
<b>“Brexit means Brexit”: A Corpus Analysis of Irish-language BREXIT Neologisms in The Corpus of Contemporary Irish</b>	<b>571</b>
<i>Katie Ni Loingsigh</i>	
<b>LEXICOGRAPHY OF LESSER USED LANGUAGES</b>	<b>583</b>
<b>Synonymy in Modern Tatar reflected by the Tatar-Russian Socio-Political Thesaurus</b>	<b>585</b>
<i>Alfia Galieva</i>	
<b>Revision and Extension of the OIM Database – The Italianisms in German</b>	<b>595</b>
<i>Anne-Kathrin Gärtig</i>	
<b>The Treatment of Politeness Elements in French-Korean Bilingual Dictionaries</b>	<b>607</b>
<i>Hae-Yun Jung, Jun Choi</i>	
<b>Lexicography in the French Caribbean: An Assessment of Future Opportunities</b>	<b>619</b>
<i>Jason F. Siegel</i>	
<b>VARIOUS TOPICS</b>	<b>629</b>
<b>The Dictionary of the Learned Level of Modern Greek</b>	<b>631</b>
<i>Anna Anastassiadis-Symeonidis, Asimakis Fliatouras, Georgia Nikolaou</i>	
<b>In Praise of Simplicity: Lexicographic Lightweight Markup Language</b>	<b>641</b>
<i>Vladimír Benko</i>	
<b>Corpus-based Cognitive Lexicography: Insights into the Meaning and Use of the Verb Stagger</b>	<b>649</b>
<i>Thomai Dalpanagioti</i>	
<b>Polysemy and Sense Extension in Bilingual Lexicography</b>	<b>663</b>
<i>Janet DeCesaris</i>	
<b>Associative Experiments as a Tool to Construct Dictionary Entries</b>	<b>675</b>
<i>Ksenia S. Kardanova-Biryukova</i>	
<b>Lexicographic Potential of the Syntactic Properties of Verbs: The Case of Reciprocity in Czech</b>	<b>685</b>
<i>Václava Kettnerová, Markéta Lopatková</i>	
<b>LexBib: A Corpus and Bibliography of Metalexicographical Publications</b>	<b>699</b>
<i>David Lindemann, Fritz Kliche, Ulrich Heid</i>	
<b>Process Nouns in Dictionaries: A Comparison of Slovak and Dutch</b>	<b>713</b>
<i>Renáta Panocová, Pius ten Hacken</i>	
<b>Definitions of Words in Everyday Communication: Associative Meaning from the Pragmatic Point of View</b>	<b>723</b>
<i>Svitlana Pereplotchykova</i>	
<b>Verifying the General Academic Status of Academic Verbs: An Analysis of co-occurrence and Recurrence in Business, Linguistics and Medical Research Articles</b>	<b>735</b>
<i>Natassia Schutz</i>	
<b>Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX</b>	<b>749</b>
<i>Arvi Tavast, Margit Langemets, Jelena Kallas, Kristina Koppel</i>	



<b>On the Interpretation of Etymologies in Dictionaries</b>	<b>763</b>
<i>Pius ten Hacken</i>	

<b>The Virtual Research Environment of VerbaAlpina and its Lexicographic Function</b>	<b>775</b>
<i>Christina Mutter, Aleksander Wiatr</i>	

<b>POSTER PRESENTATIONS</b>	<b>787</b>
-----------------------------	------------

<b>Lexicographie et terminologie au XIX<sup>e</sup> siècle :</b>	
<i>Vocabularu romano-francesu [Vocabulaire roumain-français], de Ion Costinescu (1870)</i>	<b>789</b>
<i>Maria Aldea</i>	

<b>Developing a Russian Database of Regular Semantic Relations Based on Word Embeddings</b>	<b>799</b>
<i>Ekaterina Enikeeva, Andrey Popov</i>	

<b>Semantic Classification of Tatar Verbs: Selecting Relevant Parameters</b>	<b>811</b>
<i>Alfiia Galieva, Ayrat Gatiatullin, Zhanna Vavilova</i>	

<b>Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings</b>	<b>819</b>
<i>Nicolai Hartvig Sørensen, Sanni Nimb</i>	

<b>Building a Lexico-Semantic Resource Collaboratively</b>	<b>827</b>
<i>Mercedes Huertas-Migueláñez, Natascia Leonardi, Fausto Giunchiglia</i>	

<b>The CPLP Corpus: A Pluricentric Corpus for the Common Portuguese Spelling Dictionary (VOC)</b>	<b>835</b>
<i>Maarten Janssen, Tanara Zingano Kuhn, José Pedro Ferreira, Margarita Correia</i>	

<b>Málið.is: A Web Portal for Information on the Icelandic Language</b>	<b>841</b>
<i>Halldóra Jónsdóttir, Ari Páll Kristinsson, Steingrímur Steingrímsson</i>	

<b>Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantic Knowledge from Corpora (MultiGenera)</b>	<b>847</b>
<i>María José Domínguez Vázquez, Carlos Valcárcel Riveiro, David Lindemann</i>	

<b>A Lexicon of Albanian for Natural Language Processing</b>	<b>855</b>
<i>Besim Kabashi</i>	

<b>Building a Gold Standard for a Russian Collocations Database</b>	<b>863</b>
<i>Maria Khokhlova</i>	

<b>Rethinking the role of digital author's dictionaries in humanities research</b>	<b>871</b>
<i>Margit Kiss, Tamás Mészáros</i>	

<b>European Lexicographic Infrastructure (ELEXIS)</b>	<b>881</b>
<i>Simon Krek, Iztok Kosem, John P. McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, Tanja Wissik</i>	

<b>The EcoLexicon English Corpus as an Open Corpus in Sketch Engine</b>	<b>893</b>
<i>Pilar León-Araúz, Antonio San Martín, Arianne Reimerink</i>	

<b>A Call for a Corpus-Based Sign Language Dictionary: An Overview of Croatian Sign Language Lexicography in the Early 21st Century</b>	<b>903</b>
<i>Klara Majetić, Petra Bago</i>	

<b>Exploring the Frequency and the Type of Users' Digital Skills Using S.I.E.D.U.</b>	<b>909</b>
<i>Stavroula Mavrommatidou</i>	

**From Standalone Thesaurus to Integrated Related Words in The Danish Dictionary** 915  
*Sanni Nimb, Nicolai H. Sørensen, Thomas Troelsgård*

**Exploratory and Text Searching Support in the Dictionary of the Spanish Language** 925  
*Jordi Porta-Zamorano*

**Interactive Visualization of Dialectal Lexis Perspective of Research Using the Example of Georgian Electronic Dialect Atlas** 931  
*Marine Beridze, Zakharia Pourtskhvanidze, Lia Bakuradze, David Nadaraia*

**The Dictionary of the Serbian Academy: from the Text to the Lexical Database** 941  
*Ranka Stanković, Rada Stijović, Duško Vitas, Cvetana Krstev, Olga Sabo*

## **SOFTWARE DEMONSTRATIONS** 951

**An Overview of FieldWorks and Related Programs for Collaborative Lexicography and Publishing Online or as a Mobile App** 953  
*David Baines*

**Wortschatz und Kollokationen in „Allgemeine Reisebedingungen“. Eine intralinguale und interlinguale Studie zum fachsprachlich-lexikographischen Projekt „Tourlex“.** 959  
*Carolina Flinz, Rainer Perkuhn*

**Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages** 967  
*Mika Hämmäläinen, Jack Rueter*

**Linking Corpus Data to an Excerpt-based Historical Dictionary** 979  
*Tarrin Wills, Ellert Þór Jóhannsson, Simonetta Battista*

**Collocations Dictionary of Modern Slovene** 989  
*Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Cyprian Laskowski*

**Computerized Dynamic Assessment of Dictionary Use Ability** 999  
*Osamu Matsumoto*

**Creating a List of Headwords for a Lexical Resource of Spoken German** 1009  
*Meike Meliss, Christine Möhrs, Dolores Batinić, Rainer Perkuhn*

**fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data** 1017  
*Peter Meyer, Mirjam Eppinger*

**Wordnet Consistency Checking via Crowdsourcing** 1023  
*Aleš Horák, Adam Rambousek*

## Foreword

EURALEX, European Association for Lexicography was founded in 1983 and the year 2018 marks its thirty-fifth anniversary. From its second congress in 1986, the association organises a biannual congress series. Its 18<sup>th</sup> edition, EURALEX 2018 International Congress, was held between 17<sup>th</sup>-21<sup>st</sup> July 2018 in Ljubljana, Slovenia. It was organised jointly by the Centre for Language Resources and Technologies (CLRT) at the University of Ljubljana, and Trojina Institute for Applied Slovene Studies. Both institutions are dedicated to scientific research, and the development and maintenance of digital language resources and language technology applications for contemporary Slovene. Trojina Institute was founded in 2004 with the primary objective of promoting contemporary, goal-oriented research of the Slovene language, and the University of Ljubljana founded the Centre in 2015 to ensure a systematic long-term development of technologies, resources and tools for Slovene.

The motto of EURALEX 2018 was “Lexicography in global contexts”, emphasising changes in the field of lexicography related to digital transformation, and the associated need to bring together lexicographic efforts on a global level. This has been done in recent years through the Globalex initiative, a constellation of lexicographic associations that includes representatives from all continental associations of lexicography: Afrilex, Asialex, Australex, Dictionary Society of North America, and Euralex. Similar development can be witnessed in the decision of European Commission in 2017 to fund a four-year project dedicated to the establishment of the European Lexicographic Infrastructure (ELEXIS), which was also presented at the congress.

This volume of proceedings includes congress papers submitted in three categories: papers, posters, and software demonstrations. During the review process each submitted contribution was evaluated by two independent blind referees. In case of doubt, a third independent opinion was involved. Similar to previous congresses, contributions were submitted on various topics of lexicography, including, but not limited to, the following fields:

- The Dictionary-Making Process
- Research on Dictionary Use
- Lexicography and Language Technologies
- Lexicography and Corpus Linguistics
- Bi- and Multilingual Lexicography
- Lexicography for Specialised Languages, Terminology and Terminography
- Lexicography of Lesser Used languages
- Phraseology and Collocation
- Historical Lexicography and Etymology
- Lexicological Issues of Lexicographical Relevance
- Reports on Lexicographical and Lexicological Projects

Four plenary lectures were given at the congress, with two plenary papers also included in this volume. In the Hornby lecture and paper, Sylviane Granger from the Centre for English Corpus Linguistics, Université catholique de Louvain, discusses the value of adding learner corpus data to the lexicographer’s monolingual and bilingual corpus base. Plenary lecture and paper by Lars Trap-Jensen from Danish Society of Language and Literature, also former president of Euralex, discusses three major revolutions that lexicography has witnessed in the last hundred years. The remaining two plenary lectures were presented by Judy Pearsall, Dictionaries Director at Oxford University Press, titled “One model, many languages? An approach to developing global language content” and Edward Finegan, professor emeritus of linguistics and law at the University of Southern California, on “Legal Interpretation via Corpora: Are Judges Failing Lexicography 101?”

The organising committee would like to thank all plenary speakers for setting the tone of the congress, and to other contributors for submitting very interesting work. We would also like to thank all the colleagues who reviewed the papers and the colleagues who participated in the work of the EURALEX 2018 programme committee. As in past EURALEX editions, the Hornby Trust generously sponsored one of the plenary lectures in honour of A.S. Hornby, a pioneering figure in learner's dictionaries for non-native speakers. All patrons and sponsors who supported us for this edition are listed on a dedicated page within these proceedings.

As the chair of the congress, I would like to acknowledge precious work of the members of the organising committee who joined efforts with me to make EURALEX 2018 a successful event: Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Iztok Kosem and Nataša Logar.

Simon Krek  
Chair, XVIII EURALEX International Congress  
July 5, 2018



# PLENARY LECTURES



# Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution?

**Sylviane Granger**

*Centre for English Corpus Linguistics, Université catholique de Louvain*

*E-mail: sylviane.granger@uclouvain.be*

## Abstract

In 1992 Rundell and Stock wrote an extended three-part article on the “corpus revolution”, in which they describe the rise of corpora and their impact on lexicography. Rundell (2008) revisited the topic and focused on the arrival of the web, which triggered a second stage in the corpus revolution. The purpose of my presentation is to look at some aspects of the current lexicographic scene and assess whether the corpus revolution has really fulfilled its promise. I will show that, while this is largely true of monolingual learners’ dictionaries (especially in the case of English), the situation is much less favourable when it comes to general bilingual dictionaries, a particularly worrying fact given that bilingual dictionaries have been proved to be learners’ favourite reference tool. The lack of representative translation corpora is partly responsible for this failure to keep pace. However, I will show that the judicious use of currently available monolingual and bilingual corpus resources, limited though they may be, can already bring about substantial improvements to bilingual dictionaries, in particular to one of their weakest aspects: their phraseological coverage. I will also highlight the value of adding learner corpus data to the lexicographer’s monolingual and bilingual corpus base and illustrate the benefit with the *Louvain English for Academic Purposes Dictionary*, a customisable web-based tool designed to help non-native speakers of English write academic texts.

**Keywords:** phraseology, learner corpora, bilingual corpora, bilingual dictionaries

## 1 The Corpus Revolution

In their three-part 1992 article entitled *The corpus revolution* Michael Rundell & Penny Stock described the “exciting and unnerving changes” to lexicography brought about by the emergence of a new type of lexicographic evidence, namely large collections of texts in electronic format which can be stored and processed by computers. In *The corpus revolution revisited*, written in 2008, Michael Rundell described a second stage in the corpus revolution triggered by the accessibility of much larger quantities of language data from the web. Ten years on it seems timely to take stock of the situation, highlighting both successes and weaknesses and pointing to possible avenues of improvement. In particular, as foreign language learning and teaching was at the heart of A.S. Hornby’s work, it seems appropriate, in a lecture in his honour, to assess the impact of learner corpora, a corpus resource that emerged in the early 1990s and which I, perhaps somewhat prematurely, presented as heralding another revolution, this time in the wider field of applied linguistics (Granger 1994). A fully comprehensive survey of the field is beyond the scope of this lecture and in any case exceeds my own level of expertise. My focus will be on general (i.e. non-technical) bilingual and pedagogical lexicography.

There are many different types of corpus and each has its own set of lexicographic uses. The corpora that can truly be said to have revolutionised lexicography are monolingual corpora representing native/expert speaker language. In today’s digital world, vast amounts of electronic text of this nature can be collected fairly easily, thereby ensuring that the criteria of size and balance are met. Nowhere is this revolution more in evidence than in monolingual learners’ dictionaries (MLDs) and

more particularly, in their “concern for describing and explaining phraseology” (Rundell 1998:318). Although the pioneering work of Hornby and Palmer paved the way for a more phrasal approach to lexis, it is the advent of the computer corpus that has allowed it truly to materialise. Sinclair’s innovative corpus-driven approach to lexicography, in particular, led to a formidable extension of the field of phraseology, whose scope came to embrace an increasingly large and diversified set of units, including collocations, colligations, lexical bundles and semantic prosodies.

Although the greatest contribution to monolingual learners’ dictionaries is that of native corpus data, a more limited, but highly promising, impact has been made by learner corpus data. In fact, MLDs were the first language resource to make use of this type of corpus. Based on an idea initially put forward by Maingay & Rundell (1987), Longman included error notes based on the *Longman Learner Corpus* in the third edition of the *Longman Dictionary of Contemporary English* (1995). Two other publishing houses – Cambridge and Macmillan – followed suit. In the *Macmillan English Dictionary for Advanced Learners* (2007), the notes, which took the form of ‘Get it right’ boxes, resulted from a close collaboration between Macmillan and the Centre for English Corpus Linguistics at the University of Louvain. Close investigation of the *International Corpus of Learner English* (Granger et al. 2009) resulted in extended error notes designed to warn users against common pitfalls (Rundell & Granger 2007). This was a valuable first step but as I will show in this article, it is possible and desirable to make a much more ambitious use of learner corpus data.

Fully justified as a description of the development of English monolingual dictionaries, the term ‘revolution’ is much less appropriate in the context of bilingual dictionaries. Although there can be no doubt about the potential usefulness of translation corpora in bilingual lexicography, their actual use is still very limited: “One area where one might expect such corpora to be widely used is bilingual lexicography, but in fact such corpora have not been exploited significantly in dictionary compilation – unlike monolingual lexicography, where it would be unthinkable today not to use single-language corpora” (Salkie 2008). As a result, “[t]he extraordinary range of lexical and grammatical information they [MLDs] include is rarely even approached by the best bilingual dictionaries available” (Rundell 1999). Twenty years on Rundell’s observation remains largely true. The main reason is the unavailability of representative, balanced translation corpora. Many translation corpora have been compiled by academic research teams but they are rarely publicly available. Those that are in the public domain often represent specialised registers of English (e.g. parliamentary debates in *EuroParl*) and fail to differentiate between source and target language, a serious weakness in view of the possibility of source text influence on translated texts. Other factors play a part in the – hopefully temporary – failure of corpus-based bilingual lexicography to keep pace, notably the lack of adequate tools and training to facilitate bilingual corpus exploration by lexicographers. This is not to say that the corpus revolution has entirely passed bilingual lexicography by. Some advances have been made but there is still ample room for improvement. It is to be hoped that the development of user-friendly corpus tools, in particular bilingual functionalities such as those included in *Sketch Engine* (Matuska 2018), will lead to a more systematic use of corpus data. This is particularly desirable given that several studies have shown that bilingual dictionaries are learners’ favourite reference tool. Nesi (2014: 38), for example, notes that “[a]lthough it is commonly believed that monolingual dictionaries are superior to bilingual dictionaries in terms of their usefulness as language learning tools, attitude and ownership surveys have found that learners generally prefer to use bilingual dictionaries”. This is because learners, even at an advanced stage of acquisition, still regularly think in their mother tongue (L1) and, when the target word escapes them, have no alternative but to turn to bilingual dictionaries. Even bilingualised dictionaries will not help as they are essentially monolingual dictionaries supplemented with very succinct bilingual information.

Clearly then, while the term ‘corpus revolution’ is certainly not undeserved, there is still scope for more intensive and diversified use of corpus data in lexicography. In the following sections I will



describe two directions that can be taken in order to help achieve that objective: (1) the combined use of monolingual and translation corpora to improve the phraseological coverage of bilingual dictionaries and (2) more extensive use of learner corpus data and its integration into web-based lexical environments.

## 2 Phraseology in General Bilingual Dictionaries

For a long time the phraseological coverage of bilingual dictionaries was restricted to semantically non-compositional, often figurative, units such as idioms and proverbs. The situation has fortunately improved in recent years and many bilingual dictionaries now also contain rich descriptions of collocations. However, the phraseological information provided in bilingual dictionaries is still generally felt to be insufficient (Farina 2009, Gouws 2010, Mogorrón Huerta 2011, Xia 2015). A large proportion of the prefabricated units uncovered by corpus methods, especially those that are semantically compositional, is left on the sidelines. This applies, for example, to Biber et al.'s (1999 "lexical bundles", i.e. recurrent contiguous word sequences that can easily be identified by the n-gram method, which extracts all sequences of a specific number of words (e.g. 4-word sequences) or in a given range (3- to 6-grams). The n-gram method has been commonly used to extract specialised bilingual terminology, but seems to have been underused in general bilingual lexicography. This is unfortunate as these sequences often play a key role in discourse, as textual organisers or stance markers. In addition, many are polyfunctional, which makes them particularly difficult to translate and justifies their inclusion in bilingual dictionaries. Examples of these sequences in French are: *vous n'êtes pas sans savoir que* (you are no doubt aware that, you certainly know that; lit. you are not without knowing that), *en matière de* (regarding, with regard to; lit. in matter of), *ces derniers temps* (lately, recently; lit. these last times), *il n'en reste pas moins que* (the fact remains that; lit. it no less remains that) and *en provenance de* (from; lit. in provenance from).

In view of the importance of these units and the difficulty they represent, in particular for encoding purposes, it seems worthwhile to check for their presence in general bilingual dictionaries, identify the status they are given in the micro- and macro-structure of the dictionary and assess the quality of the translations provided. This has been the purpose of recent research (Granger & Lefer 2012, 2013, 2016, Granger in press) focused on French-English general bilingual dictionaries. The methodology used involved two types of corpus: monolingual corpora to extract lexical bundles and translation corpora to extract authentic translations of the bundles. The corpus-extracted phrases and their translations were subsequently set against the descriptions provided in electronic French-English dictionaries.

In spite of its restricted scope, the research turned out to be highly instructive. First, it showed that it is possible to extract a sizeable number of useful phrases even from a small monolingual corpus of French. In Granger & Lefer (2012) we were able to identify 425 dictionary-worthy sequences of 2 to 5 words from a 1-million word corpus of French. A comparison with two French-English electronic dictionaries (*Le Grand Robert & Collins v2* and *Hachette Oxford*) revealed that 12-15% of these sequences were absent from the French-English part of the dictionaries. A follow-up study centred on the phraseology of high-frequency adverbs such as *encore* in French or *yet* in English (Granger & Lefer 2013) suggests that dictionaries tend to include phrases that are more typical of speech (*et puis quoi encore!* What next!), while the corpus brings to light phrases typical of writing, many of them with linking functions (*l'Italie, l'Espagne ou encore la France*: Italy, Spain or France). This difference is most probably due to the fact that lexicographers still rely mainly or exclusively on introspection and quite naturally tend to think of interactive markers that are more emotionally loaded than the fairly inconspicuous cohesive markers typical of writing, which can only be brought to consciousness

by corpus methods. In this area, especially given the absence of spoken translation corpora, there is clear complementarity between introspection-based and corpus-derived insights.

A second important result is that when present, a large proportion (c. 50%) of the multiword units (MWUs) are included as subentries. Not only does this status make them more difficult to access, even in electronic dictionaries, but more importantly, it accounts for the very cursory treatment they are given, often limited to one translation and/or one example sentence. In the case of a quarter of the units the situation is even worse as they appear only in an example sentence. The proportion of sequences given head-entry status was found to be very limited (4% to 13%), thereby confirming Jackson's (2013) observation that "[m]ultiword expressions [...] are most of the time treated in the microstructure although their fixed status (when they have one, which is not always the case) could justify their presence in the list of headwords". Headword status would be fully justified for word-like units which are in effect what Palmer (1917) called "accidents of graphic continuity", i.e. fixed units that are to all intents and purposes words. This would go some way towards fulfilling Sinclair's (2010: 37) general call to award headphrase status to MWUs: "The evidence from corpora adds up to a strong case for extending the treatment of multi-word units of meaning— a much wider concept than idiom— and giving them the same status as the usual headwords". Headphrase status may not be realistic for all types of phraseological units but it is certainly desirable for word-like units such as multiword prepositions (*due to, instead of, in accordance with*), multiword conjunctions (*even if, provided that, insofar as*) and multiword adverbs (*of course, on the other hand, in my opinion*), which are very numerous.

A third finding, already established by several other corpus-based studies, is that the translations provided by bilingual dictionaries and those extracted from translation corpora are very different. A first observation in this connection is that many translations suggested by bilingual dictionaries are infrequently attested in translation corpora. This is partly due to lexicographers' tendency to translate MWUs literally, a tendency that is reinforced by the fact that literal translations are often possible. For example, *de la même manière/façon* can – in some contexts – be translated by *in the same way*. However, corpus analysis shows that this literal translation, which is promoted in bilingual dictionaries, is only appropriate when the French phrase is used as an adverb of manner. When it is used as a linking adverb, the most usual translations are *similarly* and *likewise*. This example illustrates another typical characteristic of dictionary translations: the tendency to translate an MWU by another MWU, rather than by a single-word equivalent. Bilingual dictionaries also contain ample evidence of what Granger & Lefer (2012) refer to as 'categorical bias', i.e. lexicographers' tendency to translate a given source item exclusively into a word of the same grammatical category in the target language (to translate an adjective into an adjective, an adverb into an adverb, etc.). By contrast, corpus translations often involve changes in grammatical category, resulting in a host of "nice surprises" (Salkie 2008), i.e. excellent translations unattested in bilingual dictionaries. This difference raises challenging questions on the types of equivalence that bilingual dictionaries should include: systemic or textual equivalence (Adamska-Sałaciak 2010). As argued in Granger & Lefer (2016), in a production-oriented dictionary it seems legitimate to give textual equivalence a more prominent place than is currently the case.

Translation corpora have undeniable lexicographic value. However, one important caveat regarding their use must be mentioned: translation corpora are not error-free. Even when the texts have been translated by professional translators, they may contain numerous examples of calques and infelicitous translations. This is especially common in the case of discourse-oriented phrases whose pre-fabricated nature may pass unnoticed in view of their semantic compositionality and the possibility of literal translation. The role of lexicographers remains essential, since weeding out infelicitous or erroneous translations and selecting the best candidates for inclusion in the dictionary requires a level of expertise that only human experts can provide.

### 3 More Extensive Use of Learner Corpus Data

Learner corpora are a relatively new resource in the constellation of corpora that has been built over the years. One of their main benefits is that they make it possible to draw up catalogues of learners' difficulties at any stage in the learning process. These difficulties can be assessed in terms of over- and underuse of particular words, phrases or structures, and/or outright errors. As stated in the introduction, several publishing houses have availed themselves of this type of information as a basis for inserting error notes in MLDs. However, work on this front has been at a standstill for a number of years now. One of the reasons, besides possible financial motives, may well be that lexical errors tend to be L1-specific, and only errors shared by a large number of L1 populations are suitable for inclusion in generic dictionaries like MLDs. The next step forward is clearly to customise MLDs, partly in order to cater for the specific needs of different L1 populations. This trend towards customisation has been advocated by numerous authors. Rundell (2007: 50) foresees the end of "the current globally-marketed one-size-fits-all package" and sees future reference materials "as a set of components which customers can mix and match according to their needs". Individualisation also lies at the heart of function theory which describes the role of dictionaries as that of "meeting the specific types of information which a specific type of users may have in a specific type of situation" (Tarp 2009: 47).

Web-based dictionaries have a high degree of flexibility in the type and quantity of information that can be presented to users and are therefore the ideal resource to implement customisation. Another key advantage is that they allow for a high degree of interconnectivity: they can easily be integrated into wider environments featuring components such as writing aids, exercises and direct corpus access, to name just a few. Although the "one-stop shopping" dreamed of by Bowker (2010: 166-7) is probably unrealistic, at least in the foreseeable future, every effort needs to be made to reduce the number of tools that users have to open when performing a particular task, such as writing. A few tools that combine the functions of dictionary and learning/writing aid, have been developed or are currently under development (see, for example, Verlinde & Peeters 2012, García-Salido et al. 2018). In the following lines I describe the *Louvain English for Academic Purposes Dictionary* (LEAD)<sup>1</sup> (Granger & Paquot 2015), a hybrid tool designed to help non-native writers of English with writing academic texts.

The launch of the LEAD initiative was prompted by a keen awareness that an increasing number of non-native students and researchers were having to write academic texts in English but could not find any tools that met their specific writing needs. At a higher intermediate/advanced proficiency level, students and researchers usually have a good mastery of grammar and an extensive vocabulary, including domain-specific vocabulary in the case of ESP students, but fail to produce texts that conform to the typical academic style. In particular, while such writers may know the meaning of a large number of words or phrases typically used in academic texts, whatever the discipline (such as *to conclude, however, research*), they are not always aware of their distinct academic "priming" (Hoey 2005) in terms of grammar (passive or active voice), position (initial, medial or final), style (formal or informal) and, most importantly, preferred lexico-grammatical patterning, i.e. collocations and lexical bundles. For example, learners regularly use the erroneous collocation *to make research* rather than the correct *to conduct, carry out, undertake research*. They also tend to use atypical extended sequences to convey important discourse functions such as concluding, comparing or illustrating. For example, they may conclude their essays with sequences such as *I would conclude by saying that* or *I want to conclude saying that* rather than the more typical *it can/may be concluded that* or *we may conclude that*. The LEAD provides collocations and lexical bundles for over 1,200 academic words. We made use of a corpus of expert writing (academic part of the *British National Corpus*) to extract typical patterning, and of a large learner corpus, the *International Corpus of Learner English*

<sup>1</sup> A beta version of the dictionary is available at <https://leaddico.uclouvain.be>

(Granger et al. 2009), to identify learners' difficulties. As a large number of the latter are L1-specific, users are asked to specify their native language on the home screen so as to ensure that they are only presented with warnings and error notes that reflect their own potential difficulties. The LEAD is also a learning tool: it contains exercises and these too are customisable according to the user's L1. In addition, L1 customisation allows users to search in the dictionary via the translation of the English academic word in their mother tongue<sup>2</sup>, a useful feature when users have difficulty recalling the English word. The dictionary also offers a second type of customisation: the examples of collocations and lexical bundles are automatically extracted from corpora representing the discipline that users have selected on the home page. The rationale for discipline customisation is that it is more useful and more motivating for learners to be presented with examples that match their writing situation as closely as possible. In addition, users have direct access to domain-specific corpora which they can use to search for any word, whether it features as an entry in the dictionary or not.

Error notes based on learners' authentic difficulties and tailored to their specific needs represent a significant advance on the intuition-based warnings that purportedly apply to all learners whatever their L1 background. However, they suffer from one major weakness that seriously diminishes their potential usefulness: learners will only look up words whose use they are unsure about, unaware that they may also need guidance on others. To solve this problem, a new functionality, called "highlighter", has been added to the LEAD. This dynamic interface not only hyperlinks all the academic words used in the user's text to their corresponding entries in the dictionary, but also, and more importantly, draws users' attention to potential (highlighted in orange) or real (highlighted in red) errors and displays explanations in the form of pop-ups. This feature of the dictionary is very much work in progress and is only implemented for French L1 learners at this stage.

Monolingual or bilingualised dictionaries are not the only types of dictionary that can benefit from learner corpus data. Corpus-based error notes or warnings are practically non-existent in bilingual dictionaries and yet this is arguably where they would be the most useful and easiest to implement as, unlike monolingual dictionaries, bilingual dictionaries are by their nature non-generic. However, including them would require a departure from the traditional bilingual dictionary model which caters for both L1 and L2 users of the two languages and serves both decoding and encoding purposes. A more flexible model consists in customising the dictionary to the L1 vs L2 status of the user and their encoding vs decoding purpose, as suggested by Thompson (1987: 284-5): "The dictionary should, like monolingual dictionaries, be aimed in one direction (e.g. for Chinese learners of English) rather than, like most bilingual dictionaries, in two directions (e.g. for Chinese learners of English and English learners of Chinese)." The idea of a "learner's bilingual dictionary" (Granger & Lefer 2016) is fully in line with this model. A unidirectional L1 → L2 encoding dictionary could include a host of useful error and usage notes reflecting learners' attested translation difficulties. These could, for example, include warnings against false friends and calques and, where appropriate, highlight the prevalence of translations by means of different grammatical categories and zero translations. The recently launched *Multilingual Student Translation* (MUST) collaborative initiative (Granger & Lefer 2018), which aims to collect translations produced by students in a wide range of languages, will be a particularly rich source for these notes.

## 4 Conclusion

Corpora have brought about a genuine watershed in lexicography, but the field has not yet reaped the full benefit of what they have to offer. In bilingual and pedagogical lexicography in particular, there is

<sup>2</sup> Currently Chinese, Dutch, French, German and Spanish



scope for more ambitious corpus-based and corpus-driven approaches. But there are nevertheless reasons to be optimistic about the future. New corpus collection initiatives will contribute to remedying the corpus shortage that the field is currently facing. For example, the *International Comparable Corpus* (Kirk & Čermáková 2017) will provide fully comparable corpora in nine languages. The recently released *Spoken BNC2014* (McEnery et al. 2017), which contains 11.5 million words of transcribed informal British English conversation, will provide much-needed up-to-date spoken data for English. There is unfortunately no collaborative initiative aimed at collecting large, representative translation corpora. This would, however, be entirely feasible as there are so many translated books available in electronic format. If several publishers were willing to lead this initiative in collaboration with academic research teams under the aegis of reputed associations such as EURALEX, ASIALEX, AFRILEX and the DSNA, bilingual lexicography could make a quantum leap forward. As to learner corpora, the *Multilingual Student Translation* corpus (Granger & Lefer 2018) and its standardised annotation system will be a very rich source of information on students' translation difficulties in a wide range of languages. Even now, however, use of available corpus resources and techniques, even very simple ones such as n-gram extraction, can already lead to substantial improvements. In particular, learner corpora can play a significant role in the new trend towards customisation of dictionaries given their ability to help identify language difficulties that are typical of particular learner groups. These insights were difficult to integrate into paper dictionaries but can be fitted seamlessly into web-based tools.

On a final note, it is salutary to remember that corpora are just collections of language data – undoubtedly data with high potential value, but data nonetheless. Even pre-analysed by powerful software tools, their use in lexicography will always require human expertise. As Moon (2008: 334) wisely reminds us, “lexicographers still have to use intuition and judgement in selecting, interpreting, and setting out the evidence, rather than simply relaying it to the user as quasi-scientific truth”.

## References

- Adamska-Sałaciak, A. (2010). Examining equivalence. In *International Journal of Lexicography* 23(4): 387-409.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, Pearson Education.
- Bowker, L. (2010). The contribution of corpus linguistics to the development of specialized dictionaries for learners. In Fuertes-Olivera, P. (ed.) *Specialized Dictionaries for Learners*. Berlin & New York, pp. 155-168.
- Farina, A. (2009). Problèmes de traitement des “pragmatèmes” dans les dictionnaires bilingues. In: Heinz, M. (ed.). *Le dictionnaire maître de langue*, Berlin: Frank & Timme, pp. 245-264.
- García-Salido, M., García, M., Villayandre, M., Alonso-Ramos, M. (2018). A lexical tool for academic writing in Spanish based on expert and novice corpora. *LREC 2018: Eleventh International Conference on Language Resources and Evaluation*, May 7-12, 2018, Miyazaki, Japan, pp. 260-265.
- Gouws, R.H. (2010). The presentation and treatment of collocations as secondary guiding elements in dictionaries. *Lexikos* 25, pp. 170-190.
- Granger, S. (1994). The learner corpus: a revolution in applied linguistics. In *English Today*, 39(10/3), pp. 25-29.
- Granger, S. (in press). Phraséologie et lexicographie bilingue: Apports croisés des corpus monolingues et parallèles. In Hanote, S. & Raluca, N. (eds.) *Autour de l'énonciation, de la lexicologie, de la morphophonologie et de la contrastivité : langues, discours, textes et corpus*. Presses universitaires de Rennes.
- Granger, S., Dagneaux, E., Meunier, F., Paquot, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Granger, S., Lefer, M.-A. (2012). Towards more and better phrasal entries in bilingual dictionaries. In *Proceedings of EURALEX 2012, August 2012*. Oslo, Norway, pp.682-692.
- Granger, S., Lefer, M.-A. (2013). Enriching the phraseological coverage of high-frequency adverbs in English-French bilingual dictionaries. In Aijmer, K., B. Altenberg, B. (eds.) *Advances in corpus-based contrastive linguistics*, Amsterdam & Philadelphia: Benjamins, pp. 157-176.

- Granger, S., Lefer, M.-A. (2016). From general to learners' bilingual dictionaries: Towards a more effective fulfilment of advanced learners' phraseological needs. In *International Journal of Lexicography* 29(3), pp. 279-295.
- Granger, S., Lefer, M.-A. (2018). MUST: A collaborative corpus collection initiative for translation teaching and research. Paper to be presented at the Using Corpora in Contrastive and Translation Studies (UCCTS2018) Conference, 12-14 September, 2018, Louvain-la-Neuve: University of Louvain.
- Granger, S., Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. In *Lexicographica* 31(1), pp. 118-141.
- Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Jackson, H. (ed.) (2013). *The Bloomsbury Companion to Lexicography*. London & New York: Bloomsbury.
- Kirk, J., Čermáková, A. (2017). From ICE to ICC: The new International Comparable Corpus. In: Bański, P., Kupietz, M., Lüngen, H., Rayson, P., Biber, H., Breiteneder, E., Clematide, S., Mariani, J. Stevenson, M., Sick, T. (eds.) *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing*. Birmingham, 24 July 2017. Mannheim: Institut für Deutsche Sprache, pp. 7-12.
- Maingay, S., Rundell, M. (1987). Anticipating learners' errors – implications for dictionary writers. In A.P. Cowie (ed.) *The Dictionary and the Language Learner*. Tübingen: Niemeyer, pp. 128-35.
- Matuška, O. (2018). Bilingual functionalities of Sketch Engine. In Granger, S., Lefer, M.-A. (eds.) *Book of Abstracts of the Fifth Using Corpora in Contrastive and Translation Studies Conference*, Louvain-la-Neuve, 12-14 September 2018. Louvain-la-Neuve: CECL Research Papers 1.
- McEnery, T., Love, R., Brezina, V. (2017). Compiling and analysing the Spoken British National Corpus 2014. Special issue of *International Journal of Corpus Linguistics* 22(3).
- Mogorrón Huerta, P. (2011). Compétences phraséologiques et traitement des expressions figées dans les dictionnaires. In: Van Campenhout, M., Lino, T., Costa, R. (eds) *Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face au défi de la diversité*. Paris, Éditions des archives contemporaines, pp. 517-535.
- Moon, R. (2008). Dictionaries and collocation. In Granger, S., Meunier, F. (eds.) *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: Benjamins, pp. 313-336.
- Nesi, H. (2014) Dictionary use by English language learners. *Language Teaching*, 47(1), pp. 38-55.
- Palmer, H. (1917). *The Scientific Study and Teaching of Languages*. University College, London. New York, World Book Company.
- Rundell, M. (1998). Recent trends in pedagogical lexicography. In *International Journal of Lexicography* 11(4), pp. 315-342.
- Rundell, M. (2007). The dictionary of the future. In Granger, S. (ed.) *Optimizing the Role of Language in Technology-Enhanced Learning*. Proceedings of the expert workshop organised by the Integrated Digital Language Learning seed grant project, Louvain-la-Neuve, 4-5 October 2007, pp. 49-51.
- Rundell, M. (2008). The corpus revolution revisited. In *English Today* 24(1), pp. 23-27.
- Rundell, M., Granger, S. (2007). From corpora to confidence. *English Teaching Professional* 50, pp. 15-18.
- Rundell, M., Stock, P. (1992). The corpus revolution (Part 1). In *English Today* 8(2), pp. 9-14.
- Rundell, M., Stock, P. (1992). The corpus revolution (Part 2). In *English Today* 8(3), pp. 21-32.
- Rundell, M., Stock, P. (1992). The corpus revolution (Part 3). In *English Today* 8(4), pp. 45-51.
- Salkie R. (2008). How can lexicographers use a translation corpus? In R. Xiao, L. He & M. Yue (eds) *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*. Zhejiang University, Hangzhou, 25-27 September 2008.
- Sinclair, J. 2010 [2007]. Defining the Definiendum. In de Schryver, G.-M. (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Kampala: Menha Publishers, 37-47.
- Tarp, S. (2009). Reflections on data access in lexicographic works. In S. Nielsen, S. Tarp (eds.) *Lexicography in the 21st Century: In Honour of Henning Bergenholtz*. Amsterdam & Philadelphia, Benjamins, pp. 43-62.
- Thompson, G. (1987). Using bilingual dictionaries. *English Language Teaching Journal* 41(4), pp. 282-286.
- Verlinde, S., Peeters, G. (2012). Data access revisited: The Interactive Language Toolbox. In: Granger, S., Paquot, M. (eds) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 147-162.
- Xia, L. (2015). Corpora and collocations in Chinese-English dictionaries for Chinese users. *English Language Teaching* 8(10), pp.162-167.

# Lexicography between NLP and Linguistics: Aspects of Theory and Practice

**Lars Trap-Jensen**

*Society for Danish Language and Literature*

*E-mail: ltj@dsl.dk*

## Abstract

Over the last hundred years, lexicography has witnessed three major revolutions: a descriptive revolution at the turn of the 20<sup>th</sup> century, a corpus revolution in the second half of the 20<sup>th</sup> century, and the digital revolution which is happening right now. Finding ourselves in the middle of a radical change, most of us have difficulties orienting ourselves and knowing where all this is leading. I don't pretend to know the answers but one thing is clear: we cannot ignore it and carry on as normal. In this article, I will discuss how lexicography and natural language processing can mutually benefit from each other and how lexicography could meet some of the needs that NLP has. I suggest that lexicographers shift their focus from a single dictionary towards the lexical database behind it.

**Keywords:** e-lexicography, NLP, interoperability, data structure

## 1 Introduction

Let me start with a confession: In my career, I have never been much concerned with natural language processing from a theoretical point of view. The reason is simple: My main interest lies with how natural language works, and I find that formal models of language have so far been unable to come up with convincing explanations of the way language works. This applies both to formal linguistic theories such as generative grammar in the Chomskyan tradition and to computational linguistic models such as the ones used in NLP. In my experience, NLP scholars are not concerned with linguistic theory either, they are, and I apologize for the generalisation, more like engineers: interested in making things work. Coming from computer science, their theoretical interests primarily lie in the mathematical models, algorithms and methods. This is both understandable and quite legitimate. For the same reason, NLP people are excited about big data because it makes their models work. General linguists, on the other hand, are less concerned because big data does not tell us much about how language works. However, when it comes to practical applications, much of the outcome of natural language processing is highly valuable indeed to lexicographers and linguists alike, irrespective of their theoretical positions.

Between NLP and general linguistics is computer linguistics. Computer linguists resemble NLP people in their interest in big data as attested in language corpora and the patterns you can find in them. But some computer linguists are also like corpus linguists interested in the empirical facts about language, in actual language performance as opposed to the more basic underlying linguistic system.

If we relate this to lexicography, we may perhaps draw an analogy to the relation that exists between meta-lexicography and concrete lexicographical products. In the UK, there has always been a strong empiricist tradition, and perhaps this explains why "the British are poor at lexicographic theorizing, but make the best dictionaries" (H. Bergenholtz, personal communication). In my view, a preference for linguistically observational facts is a healthy point of departure, no matter how intuitively it came about in the first place. Another but related aspect is caring about the user's needs. My experience tells me that this is something that everyone has always claimed to take seriously, but there is a huge

difference between deducing user needs from a theoretical starting point and by observing user behaviour. If you take the former standpoint, the user's need is what your theory tells you, or what you can deduce from it. This is why some people from the first camp are so fond of quoting Henry Ford: "If we had asked people what they wanted, they would have said faster horses" (e.g. Tarp 2009: 28), because people do not know what is best for them (a motor car). If you take an empiricist position, what people think and, most importantly, do does matter. To take just one example: if there are words in your dictionary that are never looked up, maybe the time is better spent revising entries that are looked up all the time.

In this paper, I will discuss why it is important to provide for both NLP and human users' needs in our lexicographical practice and propose possible ways in which it could be done. In order to do so, I will first place the current situation in the historical development of lexicography over the last hundred years and characterize three major developments in the field. From this follows that the current conditions for making lexicographical products are fundamentally different from what they were only 25 years ago when I entered the field. Most importantly, it also means that we cannot make dictionaries the way we used to, and towards the end I will discuss how we can go about this.

## 2 Background: Three Revolutions in Lexicography

### 2.1 The First Revolution: the Descriptive Paradigm

Going back about a hundred years in time, it is not difficult to spot the differences in the dictionary production process. This is around the time when the large national monolingual dictionaries emerged, created with the help of boxes full of excerpts and carried out by an army of hard-working lexicographers who tirelessly produced volume after volume, such as we know it from *Oxford English Dictionary* (OED) in England, the Dutch *Woordenboek van der Nederlandsche Taal*, *Deutsches Wörterbuch* by the Grimm brothers in Germany and *Svenska Akademiens Ordbok* (SAOB) in Sweden. These dictionaries are today widely known and classic works whose acronyms are familiar far beyond the narrow circle of lexicographers. But even though they are old projects and the methods may occur simple from a present-day point of view, they do not differ very much in nature from what we know and use today: descriptions of language and language usage based on empirical analysis – corpus-based as we would call it today. In those days, the corpus consisted of excerpts of language samples stored in file boxes and analyzed for each word and meaning.

In my own country, a national dictionary, *Ordbog over det danske Sprog* (ODS, Dictionary of the Danish Language), was compiled in a similar way, and the method used constituted a real break, a paradigm shift in the classic, Kuhnian sense: ODS broke off from the 19<sup>th</sup> century tradition and insisted on the descriptive approach. We can call this the descriptive revolution. The tradition prior to the ODS was characterized by what was called 'the academy principle' with reference to the practice typical of the French Academy dictionary. In accordance with this principle, the Danish dictionaries before ODS were prescriptive, and the motivation for their compilation educational: a wish to educate the common people and teach them good and exemplary language through beautiful examples by admired authors and linguistic role models. One of the founding fathers of the *Dictionary of the Royal Danish Academy of Sciences and Letters* (1793-1905), Jacob Langebek, expressed the opinion that the dictionary should only accept words that were "good, pure, generally usable and unmistakably Danish" words (ODS 1918: the preface), whereas he saw no room for:

All coarse, plump and horny words and words that strive against decency ... for they do not need to be known to those who do not pay heed to them, and those who want to learn them will get to know them anyway" (ODS 1918: the preface).



Consequently, the dictionaries must be selective and include only the good words, whereas the ‘coarse, plump and horny’ ones have no place in the dictionary.

We find the same way of thinking with another 19<sup>th</sup>-century Danish lexicographer, Christian Molbech, the author of the most widespread monolingual dictionary of Danish in those days. He said:

Even the most frequent use of a newly formed word, especially in spoken language, does not yield it any authority, and proves nothing for its usefulness in the pure language and good style, or for its acceptance in a dictionary, if it offends an ear cultivated towards fine language. (Chr. Molbech, the preface from 2<sup>nd</sup> edition 1859, here quoted from Dahlerup 1907).

The dictionaries should contain only ‘good’ words, ‘the most beautiful flowers in the language’.

It should thus be an honour for a word to be included in the dictionary, as it is an honour for a piece of art to be admitted into a national arts collection (Dahlerup 1907: 68).

This is how Molbech’s principle is expressed by Verner Dahlerup, the founder of the ODS. And Molbech himself writes that his dictionary should be “an interpreter of the proper use of the pure, educated written language of our present time”.

A final example of the academy principle comes from the first Danish slang dictionary, *Dictionary of the Vulgar Tongue and so-called Daily Speech*. The author was V. Kristiansen, a pseudonym for Viggo Fausbøll, a professor of Indian and Eastern Philology at the University of Copenhagen. In the preface to the dictionary, he writes that the purpose of his dictionary is not so much explanatory as it is a warning against vulgar language:

In recent years, this vulgar tongue ... is threatening to force its way into the families ... having collected some of what belongs to it, I have, apart from a purely linguistic aim, in addition wanted to draw attention to the danger and tried to provoke resistance against the same and I assume that once people have opened their eyes to the indecent crossing of the line, all educated people will agree to ban the vulgar tongue from good society and leave it to the guttersnipes and the adherents of Grundtvig in whose taste it may fall (V. Kristiansen 1866).

As a consequence, the most vulgar words were typeset in Greek letters in the first edition so as to prevent common people from exposure to words they were better off not knowing.

Another dictionary that appeared at the same time as ODS began was Dahl and Hammer’s *Danish Dictionary for the People*. It is well known for its puristic approach motivated by a desire to educate common people and spare them from seeing unwanted words, whether foreign words, slang, dialect or other types of language considered vulgar or non-standard.

The merit of ODS is that it broke away from this prescriptive tradition. Its founder, Verner Dahlerup, wanted ODS to be a science-based practical tool for language understanding, and to him it did not make sense to exclude words because they were thought to be dubious or destructive:

First of all, I cannot ask, “should this or that word be used?”. I ask instead: “is it used or has it been in use?” If so, I will include the word in so far as it falls within the scope of the dictionary. (Dahlerup 1907: 71).

This quote is taken from an article in the journal *Danske Studier* (Danish Studies) in which Dahlerup explains his ideas about the new dictionary. Dahlerup was not the only one who believed that actual language usage was the key to semantic description. It was a topical belief at the time among the Neogrammarians who had turned to the study of modern languages in their attempt to demonstrate the universality of the phonetic laws. In the latter half of the 19<sup>th</sup> century, this was an advanced and controversial position in linguistics but one that would eventually pave the way for the new structuralist

paradigm that came to replace comparative linguistics. The principle of empirical analysis as the basis for linguistic description prevails even today, more than a hundred years later. But the basis for empirical analysis and description has changed dramatically.

## 2.2 The Second Revolution: The Corpus Revolution

The Corpus Revolution is the next major leap that completely changed the way we make dictionaries. In terms of theory and method, excerpted language material and filing boxes are not fundamentally different from corpus concordances and semantic annotation; in a way, it is just a difference between analogue and digital methods. However, with the technical development that took place from the 1960s onwards, a substantial part of the work could be automated with the help of computers, resulting in a substantial expansion of the descriptive basis while at the same time editorial work could be carried out more efficiently.

The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.) which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the word being described. (Fillmore 1995).

The potential of the use of a corpus is enormous if you take seriously what Fillmore says here, that exhaustive word description implies the investigation of all connections with other words in which a word engages to see if they condition the meaning of that word in one way or another. Fillmore has done so himself in the FrameNet project, and in lexicography we have also seen some projects along similar lines. We will come back to this later.

In lexicography, the corpus revolution took off in earnest in the 1980s with the COBUILD project. At first, the greatest advantage of a corpus was that the sample material multiplied in comparison with the traditional procedure involving language slips and filing boxes. Measured in size, corpora have roughly increased tenfold every 10-15 years. In the 1960s and 1970s, the first corpora had about 1 million words – this is the size of the Brown corpus and also of the Swedish corpus Press65. COBUILD's corpus from 1985 had about 18 million words (today it has developed into the Bank of English with about 650 million words). The British National Corpus reached 100 million words in the mid-1990s. Among the largest corpora today are the German COSMAS (or DeReKo: Deutsches Referenzkorpus) from IDS in Mannheim with more than 5 billion running words, Collins Corpus with over 4.5 billion words, not to mention Google's corpus of texts that have been scanned for the Google Books project comprising more than 500 million books from 1500 to today for a number of the major languages. Different estimates exist, but it is not really important whether it contains 155, 175 or 200 billion words. To most people, the number is incomprehensible anyway or as good as infinite in size.

In my own part of the world, the Nordic region, we cannot quite match the size of the large English and German corpora, but even so the Language Bank in Gothenburg contains more than 1 billion words, and the situation is similar in Norway and Denmark: the Norwegian newspaper corpus contains more than 1 billion words, and the corpus developed at my own institution, the Danish Society for Language and Literature, has also reached a billion words.

The development in the use of corpora for dictionary work is parallel to the development in corpus size. In the early years of the corpus era, the advantage was access to larger sample material than was possible with the help of index cards. For the lexicographer, the task was to browse through concordances and arrange the tokens according to homographs and senses, quite similar to working with

index cards. However, anyone who has worked with concordances will know that even in a moderate corpus, this task becomes overwhelming when you are analysing the common words of a language because the words belonging to the core vocabulary are also frequent and, as a result, give a huge number of concordance lines.

That is why it was a step forward when it became possible to have annotated corpora. Restricting a search to, for example, verbal instances of a homograph or to particular inflectional forms, the lexicographer could skip a lot of unwanted and irrelevant tokens and thereby speed up the work process.

Around the turn of the century, we saw the first syntactically marked-up corpora, which further helped to find distinctive syntactic patterns in the texts, for example in the form of Word Sketches (Kilgarriff et al. 2004) or other forms of lexical profiles. And while the corpora grew in volume, it became increasingly necessary to do something to handle the overwhelming amount of information that came with it. The solution tends to be more and more preprocessing of the material, using techniques that pre-analyze the texts according to different parameters that allow the lexicographers to find just the instances they need in the relevant phase of the editing process. Let us look at some of the opportunities that are explored.

Sorting out corpus instances by homographs and senses is a key element in the daily routine of any lexicographer, and even though we do not yet, to my knowledge, have an operative technique for automatically sorting out concordances semantically, it needs to be mentioned first as this is the ultimate goal. So far, sense annotation is something that editors must do more or less manually. In the current state of corpus linguistics, lexical profiles can help to uncover some meanings automatically, as there is obviously some relation between semantic meaning and, for example, valency patterns or subject domains, both of which can be detected by means of corpus linguistic methods.

Using a corpus as a tool for lemma selection is another obvious possibility. Corpus frequency is one of the parameters used to determine which words should be included in a dictionary.

Statistical methods such as Mutual Information, T-score, log-likelihood etc. are methods that are suitable for demonstrating how words attract each other. Corpus linguistic techniques are therefore likely to be useful in finding and analyzing patterns within the field of phraseology. Valency patterns and lexical profiles can be found in a similar way, although it is not the direct attraction between words that is measured, but the linguistic material of syntactic categories and phrases in a syntactically marked-up corpus.

A fairly new and therefore less well-known application is the use of a corpus to monitor diachronic language development. The general idea is to compare any subset, the so-called focus corpus, against the entire corpus, in this context known as the reference corpus (see Cook et al. 2013). Significant features of the focus corpus can be said to be characteristic of that particular subset. If the focus corpus consists of a single year and we find a number of words and phrases that occur only here and not in the reference corpus, a reasonable hypothesis would be that we have come across neologisms and, possibly, lemma candidates for the dictionary. This is probably the most obvious use of monitor corpora for lexicographical purposes, but in principle anything could be investigated in similar ways. The focus corpus could be a domain specific corpus, and if one finds that common language words suddenly appear in the domain-specific corpus, it could be an indication that something interesting is going on that deserves closer inspection. An example would be if we found words like *mouse*, *cloud*, *worm* and *virus* in a computer focus corpus with above-normal frequency. This could be an indication that these common words had undergone some linguistic change in this domain-specific context, an observation that we know, in hindsight, to be true.

The development can also go in the opposite direction: words from a specific domain turn up in the general-language texts. This is the case when we find words and expressions from the world of sports

in general-language texts as a reflection of sports as a productive source of new metaphors: *below the belt, slam dunk, saved by the bell, checkmate, bullseye, the ball's in your court*.

The technique can be used similarly as a tool to mark senses with usage and domain labels. If a comparison reveals that a word or expression is overrepresented in texts from a particular domain, it is likely to be a domain-specific meaning. The method can be applied to any type of variable relating to speaker (age, gender, regional distribution, occupation etc.) or text type (style (formal/informal), channel (spoken/written), public/private, genre etc.) as long as the corpus has been marked up with the relevant metadata.

Finding good language examples is another area where pre-processing has proved successful. This basically means that only the best-suited examples are shown as the result of a corpus query or, alternatively, they are shown at the top of the concordance list. Finding good examples manually is a time-consuming and therefore expensive task, so there is much to be gained if an automatic procedure can find all and only the best examples. What is considered a good example may vary from dictionary to dictionary, depending on the intended user group and the dictionary's distinctive style but they probably have a few characteristics in common: a good example consists of a whole sentence, neither too long nor too short (about 15-20 words), it should not contain proper nouns (because they require cultural knowledge, are short-lived and may be a breach of people's privacy), it should not contain deictic expressions or pronouns referring to something outside the sentence; and it should preferably contain a typical collocation or some other idiomatic pattern, whereas difficult or rare words should be avoided. Criteria like these can be determined in advance and the query result sorted in such a way that the lexicographer will be presented with examples that meet all criteria first and can find an example that is a good candidate for inclusion in the dictionary. Such a feature is integrated with the SketchEngine corpus query tool (GDEX, see Kilgariff et al. 2008), and a similar system has been developed at the Berlin-Brandenburg Academy of Wissenschaften for German (Didakowski et al. 2012).

To summarize, the corpus revolution can hardly be described as a paradigm shift in the sense of Kuhn. But it brought the descriptive tradition, starting from the beginning of the 20<sup>th</sup> century, a giant step forward and took lexicography much closer to the goal: the dictionary description as a mirror of language in all its diversity. In the corpus era, description shifted in focus from the exemplary language of classic writers towards capturing more and more language varieties in an attempt to embrace and reflect the entire language. In the 1990s, many believed that representativeness was at least as important as volume, but that idea has now more or less been abandoned, probably due to practical rather than theoretical reasons: balanced corpora are much more difficult and expensive to develop, and none of today's mega-corpora are particularly well-balanced. Instead, they are either dominated by journalistic texts or have been harvested from the web. The keyword is accessibility.

Viewed from the users' perspective, the corpus revolution has hardly been noticed by many outside the lexicographic world because it did not change dictionaries and lexicographical products radically. But for lexicographers, it had a great impact as it improved the descriptonal basis and enabled us to make better dictionaries.

The growing size of corpora also changed the working conditions for practitioners in the field. A hundred years ago, the editors would spend most of their time editing an entry, starting from scratch until there was a complete article ready for publication. The only pre-processed material used in the process was the box with excerpts, the rest was the editor's responsibility. This has changed dramatically. In the early days of corpus lexicography, the editor was still to a large degree in charge of the entire process: the corpus was a mere tool, while the editor's responsibility was to read through concordances, sorting and selecting from them. But when one is working with huge corpora, as we



do today, this task becomes increasingly insurmountable. The material is simply too vast. The solution is pre-processing: the editor is presented with semi-manufactured elements, suggestions which the computer has analyzed in advance: spelling, inflection, valency patterns, collocations, idioms, morphological and syntactic restrictions, perhaps quotes and linguistic labels, just to name the most obvious options. The editor's role has changed into one of choosing, checking and validating from the pre-processed material presented to them. Consequently, the skills needed to become a qualified editor have changed: certain skills are no longer required, while others are becoming more important.

### 2.3 The Digital Revolution

Finally, I am getting to the point with the third major development that can be identified in the last century. It is perhaps also the most difficult to describe as we are still in the middle of it. It is of course what has been called the electronic or digital revolution: the development that has changed the dictionary from an analogous paper product and turned it into something digital, a webpage, an app or an embedded feature somewhere in cyberspace. The development is gradual and has been going on for more than 30 years. The first digitization projects were launched in the early 1980s (SAOB in Sweden, OED in the UK). In the 1990s, CDs and PDAs became popular, offering better search facilities, while content-wise remaining basically the same. During the 2000s, any dictionary with self-respect has either migrated from paper to screen or has gone out of business. Especially in the last 10 years, things have developed rapidly, not least due to the spread of smartphones and tablet computers. Young people today grow up in a world where communication, reading and learning are dominated by computers, and especially on small mobile devices of some sort.

If the influence of the corpus revolution was mostly an internal affair that affected the lexicographic community, the digital revolution has not only been evident to dictionary users but has in effect changed the daily life for every one of us. But it has of course also changed the way we make dictionaries.

In the time of paper dictionaries, there was a close relationship between the contents written and edited by the lexicographers and the finished product. The printed work was the main product, the output created by the lexicographer was an intermediate stage in the process, no matter if it was written on a sheet of paper, in a word processing program or in a database.

Today, we are much more aware of the fact that the database is the central element of the work, and that the database structure must be well organized and sufficiently flexible so as to publish in different media and for different platforms.

The digital revolution has changed dictionaries and dictionary-making in several ways: the business model, improved search possibilities and assistance to language learners and insecure spellers, user involvement and crowd-sourcing, to name but a few of the present challenges. In this context, we are principally concerned with certain aspects: the possibility to access, link and share data with others.

## 3 The Digital Era

In recent years, lexicography and the NLP community have been brought closer together, in particular for two reasons: the Internet and the use of hyperlinks to connect data and websites have made it possible, and the digital development has made it necessary: seeking information online has taught the users to be impatient. They want their questions answered quickly and they want the answers for free. Dictionaries are no longer the golden eggs they used to be for publishing companies, so they are forced to change their business models if they want to stay in the business while newcomers have entered the scene hoping to get their share.

Where does that leave lexicography in the current situation? On the one hand, existing dictionary providers have migrated from print to screen, having improved data structure and search facilities, perhaps added new information in the form of audio and video clips, in the hope that users will use and appreciate the new products. On the other hand, their efforts seem to have had limited success. Some people are quite pessimistic in their assessment:

the biggest problem of lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people, in fact, do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead (Simonsen, 2017: 409).

Undoubtedly, this is a generalization that ignores a lot of variation in both behaviour and experience that people have across languages and resources. But it is probably true that young people today are less willing to use dictionaries than the generation before them, and therefore dictionaries are not used as much as they could (or should) be. The answer why this is so, is complex and several factors are involved but the explanation made by Simonsen certainly plays a role:

why do not people use online or mobile dictionaries? Obviously, there are a number of reasons, but I would argue that the most important reason is that most lexicographic resources are not tool-integrated and not specifically related to the user's job tasks (Simonsen, 2017: 409f.).

When you are in the middle of a transition, standing at the crossroads, it is difficult to predict in which direction the future is pointing. Even so, some tendencies can be traced that we need to take seriously.

Traditional lexicography has been challenged by newcomers from natural language processing offering computationally developed lexicographical resources, either meant for computers, such as lexicons like WordNet, FrameNet and VerbNet and various lexical knowledge bases and ontologies used in the Semantic Web, or resources for human users obtained by combining already existing web resources in new ways, such as BabelNet, TheFreeDictionary.com and others. Also, new resources have been created through collaborative efforts by dedicated volunteers, with Wikipedia and Wiktionary as the best-known examples.

While most people agree that traditional lexicographical resources provide high-quality semantic descriptions of languages, it does not follow from this that traditional resources are the ones that are used in computational lexicography and NLP. In fact, they are not, and the explanation is simple: most traditional dictionaries were developed as self-contained entities, encoded in data structures known alone to the publishing company or the institution responsible and kept by them as a secret.

The world of the digital era, with Semantic Web and linked open data as the current buzz-words, is altogether different. Here the keywords are accessibility and interoperability. Even if we accept the fact that users are becoming increasingly reluctant to look up words in a dictionary, there is no reason to believe that their needs for language assistance have decreased. But we may have to meet their needs in new ways, most obviously by embedding language tools directly in the applications and other computer software.

In general, few users are interested in learning the totality of meanings of a particular word when they are looking up. More likely, they are interested in the meaning of the word in a specific context – the one that caused them to look up. So, if embedding is important, it is equally important to develop techniques to pick just the right lexical unit in a given context.

It has been well known for a long time that word sense ambiguation is a major challenge for natural language processing and computational lexicography. For a computer, it is very difficult to determine whether two words, or rather lexical units, are similar. Consider the following definitions of the adjective *kind*:

- (1) behaving in a way that shows you care about other people and want to help them (Macmillan)
- (2) generous, helpful, and thinking about other people's feelings (Cambridge English Dictionary)
- (3) caring about others; gentle, friendly and generous (OALD)
- (4) saying or doing things that show that you care about other people and want to help them or make them happy (LDOCE)

Taken individually, these definitions are quite acceptable ways of explaining what it means to be kind, and we, as lexicographers, make it a point of honour not to copy a definition from others but phrase it in a style that is characteristic of our dictionary. For a computer, however, they are too different for it to work out that they are different ways of explaining the same word meaning.

For comparison, consider the following definitions:

- (1) saying kind things to someone who has problems and behaving in a way that shows you care about them (LDOCE, *sympathetic*)
- (2) kind, helpful, and sympathetic towards other people (Macmillan, *caring*)
- (3) behaving in a pleasant, kind way towards someone (Cambridge, *friendly*)
- (4) (of a person) kind, friendly and sympathetic (OALD, *warm-hearted*)

Likewise, it is difficult or even impossible for a computer to tell that these explanations differ from the previous ones and are in fact explanations of four different but semantically related words: *sympathetic*, *caring*, *friendly* and *warm-hearted*, respectively.

If a computer should identify and link words like these in, say, a new and automatically generated resource, it needs a helping hand. That it is why resources like WordNet and FrameNet have become popular in natural language processing, because they do just this: label the semantic relations between words in an explicit way. In WordNet, *kind*, *sympathetic*, *caring*, *friendly* and *warm-hearted* can easily be identified as synonyms or near-synonyms if they belong to the same or neighbouring synsets.

Seen from the computer's point of view, the solution would be to give all words that are labelled synonyms identical definitions. In a computer lexicon, this would make sense: if meanings are synonymous, they should have the same denotation and so their definitions should reflect this fact.

Luckily (for the human user at least), this is to my knowledge not carried into effect in any existing dictionary. Few words, if any, are totally congruent in meaning anyway, and in a dictionary for human users, the most important thing is that the definition is well-phrased and is functioning in itself.

A similar case which is frequently mentioned as a problem for NLP exploitation of lexicographical data is Apresjan's notion of regular polysemy: the fact that the same meanings can regularly be identified for whole groups of words. For instance, words like *hospital*, *school*, *office*, *supermarket* etc. can all have the senses 1) a building ('she went into the office'), 2) the people working there ('the hospital decided to close the clinic') and 3) an institution or business ('the highest-ranking schools in the country'). Or a food container can refer to either the physical object ('a glass', 'a bottle') or its content ('they had two glasses and left'). In a computer lexicon, this would imply that these senses would have to be present for all hyponyms, or subordinate terms, of *school*, including *pre-school*, *high school*, *secondary school* and *summer school*. In real life – which in this connection means in concrete dictionaries – they are typically not, either accidentally if the lexicographer estimates that they need not be elaborated (*secondary school*: a school for children between the ages of 11 and 16 or 18, Macmillan), or deliberately so if some of the meanings happen to be too infrequent in the underlying corpus for them to be described in the dictionary.

The solution is, as I see it, in neither case to bring the dictionary data in harmony with the demands of the computer. If a definition works fine in itself and is helpful to its user, there is no reason to

standardize it just to make it compliant with the definition of a synonym. And there is no need to describe an infrequent sense just because it can be inferred from a superordinate term. Instead, if lexicographers want to make their data applicable for further exploitation and enrichment by the NLP community, there are several other ways of proceeding, and probably even more than a non-expert such as myself can imagine.

However, over the last 25 years I have seen what (my colleagues') foresight can bring, and if others can learn from that, I am happy to pass it on. Let us look at a few examples.

In the 1990s, during the preparation of The Danish Dictionary, we decided to create a separate element in the data structure devoted to the genus proximum of a word. Here the lexicographer would enter the nearest hyperonym (or superordinate term) of the lexical unit in question. At the time, we only had vague ideas about the future use of such an element, and it was of course not visible in the final printed dictionary.

Similarly, a systematic domain element was created where the lexicographer would assign a domain label to a lexical unit if at all possible. This element should not be confused with the traditional domain label. Whereas the domain label is visible in the dictionary and serves the function of marking a sense as a technical term, the systematic domain element was never intended for publication and simply makes explicit what part of the vocabulary a particular sense belongs to, whether used in common language or as a technical term.

These two elements turned out to be highly useful when years later we used the data from The Danish Dictionary to create the Danish WordNet, DanNet, in collaboration with Copenhagen University (Pedersen et al. 2009). It was also crucial for the decision to create the WordNet from scratch rather than translating from Princeton WordNet. Later, we used the structure of the WordNet to help us organize the meanings of The Danish Dictionary conceptually when we edited a Danish thesaurus (Nimb et al. 2014). And most recently, the data from the Danish Thesaurus have been used to develop a Frame Lexicon for Danish in another collaborative project with Copenhagen University (Nimb 2018).

## 4 Lexicographical Solutions

It is no coincidence that we are witnessing an increasing need for high-quality lexicographical data. Language technology and artificial intelligence are moving into a phase where the word lists and morphological lexicons developed inside the NLP environment itself are insufficient to meet the demands for developing smarter and more sophisticated products. Automatic content summaries, domain classification and virtual assistants are but a few examples of applications that require 'knowledge' or some way of handling the semantics of human language. By far the best existing semantic descriptions of language are dictionaries, and for that reason, it is obvious that existing dictionaries are interesting for developers of such applications. But it is also obvious that the structure and nature of data used for computer dictionaries are different than data for human dictionaries. The justification of a project such as ELEXIS (for example Pedersen et al. 2018) is exactly to bring these two communities together and explore how existing lexicographic descriptions in human dictionaries can be accessed and converted in such a way that they can serve as input to computers.

My point in this connection is to stress that convergence should take place in both directions and that there are several things that traditional lexicographers can do to meet the needs of language technology. First and foremost, it is necessary that lexicographers shift their focus away from the concrete end product and towards a lexical database that can serve both worlds at the same time. My expertise



does not suffice to make exhaustive suggestions what such a database should ideally look like, but my hope is that experts from both sides will join forces and agree on the minimum requirements and propose useful ideas. The recommendations could include the following – and probably many more:

#### 4.1 Accessibility

For data to be accessible, it is not enough that data owners are willing to share them with colleagues or customers on certain conditions (that can be specified in a license), the data itself must also be recognizable when exchanged. This is why a lot of discussions seem inevitable in any infrastructure project: we need to agree on certain data formats and standards if we want to profit from the benefits of exchanging data with others. If the data do not follow any of the current standards, at least they should be capable of conversion for export and exchange purposes.

#### 4.2 Unique Identification

Trivial as it may sound, it is important that the central units of the database can be uniquely identified. Traditionally, headwords have been the central units of dictionaries, but headwords are far from always enough to identify all occurrences of the words in a dictionary since many words can be either homographic or polysemous or both. When we say that *pale* and *light* are synonyms, this is, strictly speaking, not a relation between these two words themselves, but rather between the two words in one of their respective meanings. And when we talk about a *pale imitation*, we mean not ‘pale’ and ‘imitation’ as lemmas, but in a specific meaning of the two words. So, if we want to be able to identify word occurrences uniquely (and for computer purposes this is necessary), we must be able to point to the combination of lemma form and meaning. This combination is also known as a lexical unit, and undoubtedly it is wise to use a unique ID number for each lexical unit in the database. In this way, it would be possible to mark up all the words used in a dictionary (including definitions and examples) with a unique ID number, something which would be of help to a computer when given data as input for further processing.

#### 4.3 Data Structure

In the same way as hidden elements for genus proximum and systematic domain assignment were used in The Danish Dictionary, it is possible to add further elements or attributes to the database in the form of ID numbers or links to other relevant resources. This could be links to a synset in WordNet or to the proper WordNet super sense, it could be to a particular frame in FrameNet, etc. In this way, the resource would become gradually integrated with other resources, and this is indeed the whole idea behind linked open data and other data-exchanging communities.

#### 4.4 Consistency

Computers require more consistent data than humans, and for this reason the database should contain consistent descriptions compliant with systematic polysemy, rules of inheritance throughout the ontological hierarchies, etc. However, none of this needs to be visible in an extract of the database used to publish a traditional dictionary for humans as long as the elements in the database are clearly marked and consistently used for computer purposes alone.

## 5 Conclusions

By no means an NLP expert myself, I am sure that others will have both more and better suggestions on what it takes to improve language technology and how lexicography can contribute. There

is, however, little doubt in my mind that much is to be gained if lexicographers are willing to accede to the demands made by computational processing. These could involve, but do not exhaust, the following steps:

- (a) Use a lexical database for your data that is independent of the particular product that you are working on. But make sure to organize the database in such a way that it permits extraction of exactly the kind of information you need for a concrete dictionary to a given target audience.
- (b) Use a standard format (such as TEI) to make your data easy to export and modify when exchanged with others.
- (c) Use ID numbers to uniquely identify the central elements of the database; most often these will be the lexical units: a headword in one of its senses. Often this is also useful for internal purposes because ID markup of all the words in the dictionary, including definitions, synonyms, collocations and fixed phrases, is required for unique links between the words.
- (d) Use elements (or attributes) in the database that could be useful for NLP purposes: genus proximum, systematic domain assignment, ontological type and super senses have been mentioned above, but the only limitation is your imagination, and the NLP community is invited to make further suggestions to their needs.
- (e) Use attributes (or elements) that make explicit the exact position or relation of a lexical unit to one or more existing external NLP resources such as WordNet, FrameNet or VerbNet.

What has been said above is by no means in itself decisive, what matters is that NLP specialists and lexicographers realize that they are dependent on each other and must work together to find new ways of making lexical resources for the next generation. One such initiative is ELEXIS, and hopefully this project will pave the way for further co-operation and mutual understanding between the two fields. It is not necessarily, or at least not only, a question of giving NLP free and open access to data developed by lexicographers. It is as much a question of lexicographers realizing that their data are not exclusively made for a specific end-product, a traditional dictionary. The data are equally important as input for NLP and must be structured optimally to be suited for that purpose. In fact, we as lexicographers may soon come to realize that traditional dictionaries are no longer in demand. The users of tomorrow's computers may want their linguistic problems solved by simply talking to their device: "What does X means in this sentence?" or "give me another word for *kind*". In order for us to provide the answer, it is likely that a new division of labour is needed where lexicographers maintain the underlying database while other experts are responsible for the access paths and interfaces that connect the database content with the user. Lexicography and NLP must get together and work on joint solutions to meet the challenges of the digital era. This is crucial for the continued prosperity for lovers of language, words and reference works.

## References

- Apresjan, J. D. (1974). Regular Polysemy. In *Linguistics*, Volume 12, Issue 142, pp. 5–32. Cambridge English Dictionary. Accessed at: <https://dictionary.cambridge.org> [30/04/2018].
- Cook, P., Lau, J.H., Rundell, M., McCarthy, D & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Electronic lexicography in the 21<sup>st</sup> century: thinking outside the paper*. Proceedings of the eLex 2013 conference: Tallinn, Estonia. pp. 49–65.
- Dahl, B.T. & Hammer, H. (1907–14): *Dansk ordbog for Folket I–II* (Danish Dictionary for the People). Copenhagen and Kristiania: Gyldendalske Boghandel, Nordisk Forlag.
- Dahlerup, V. (1907). *Principer for ordbogsarbejde* (Principles of Lexicographical Work). In *Danske Studier* 1907, 65–78. DanNet. Accessed at: <http://wordnet.dk> [30/04/2018].
- Dansk Ordbog udgiven under Videnskabernes Selskabs Bestyrelse 1–8 (Danish Dictionary published under the Guidance of the Royal Danish Academy of Sciences) (1793–1905). Copenhagen.

- Den Danske Ordbog, DDO (The Danish Dictionary). Accessed at: <https://ordnet.dk/ddo> [30/04/2018].
- Didakowski, J., Lemnitzer, L., Geyken, A. 2012. Automatic example sentence extraction for a contemporary German dictionary. In Fjeld, R.V., Torjusen, J.M. (eds.) *Proceedings of the 15<sup>th</sup> EURALEX International Congress*. Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 343–349.
- Fillmore, C.J. (1995). The Hard Road From Verbs To Nouns. In M. Chen & O. Tzeng (eds.) *In honor of William S-Y. Wang*. Taipei, Taiwan: Pyramid press, 105–129.
- Grimm, J. & Grimm, W. *Deutsches Wörterbuch*. Accessed at: [http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui\\_py?sigle=DWB](http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui_py?sigle=DWB) [30/04/2018].
- Kilgariff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In Williams, G. & Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France, pp. 105–115.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, M. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 425–433.
- Kristiansen, V. (1866). *Bidrag til en Ordbog over Gadesproget og saakaldt Daglig Tale* (Contributions to a Dictionary of the Common Language and So-Called Vulgar Tongue). Copenhagen: Boghandler H. Hagerups Forlag.
- Longman Dictionary of Contemporary English, LDOCE. Accessed at: <https://www.ldoceonline.com> [30/04/2018].
- Macmillan English Dictionary Online. Accessed at: <https://www.macmillandictionary.com> [30/04/2018].
- Molbech, C. (1859): *Molbechs ordbog 1–2*. Copenhagen: Gyldendalske Boghandlings Forlag.
- Nimb, S., (2018). *Fra begrebsordbog til FrameNet* (From Thesaurus to FrameNet). In Nielsen, J.G., Petersen, K.S. (eds.) *DSL's Årsberetning 2017–2018* (Annual Report of DSL 2017–2018). Copenhagen: Society for Danish Language and Literature, pp. 80–86.
- Nimb, S., Trap-Jensen, L. & Lorentzen, H. (2014). The Danish Thesaurus: Problems and Perspectives. In Abel, A., Vettori, C. & Ralli, N. (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15–19 July 2014. Bolzano/Bozen: EURAC Research, pp. 191–199.
- Ordbog over det danske Sprog, ODS (Dictionary of the Danish Language) (1918–1956). Copenhagen: Society for Danish Language and Literature and Gyldendal Publishers. Accessed at: <https://ordnet.dk/ods> [30/04/2018].
- Ordbok över svenska språket utgiven av Svenska Akademien (Dictionary of the Swedish Language published by the Swedish Academy), SAOB (1898–). Accessed at: <https://www.saob.se> [30/4/2018].
- Oxford English Dictionary, OED. Accessed at: [www.oed.com](http://www.oed.com) [30/04/2018].
- Oxford Advanced Learner's Dictionary, OALD. Accessed at: <https://www.oxfordlearnersdictionaries.com/definition/english> [30/04/2018].
- Pedersen, B. S., McCrae, J., Tiberius, C., & Krek, S. (2018). ELEXIS – a European infrastructure fostering co-operation and information exchange among lexicographical research communities. In *Proceedings of Global WordNet Conference 2018*. Singapore.
- Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen N.H., Trap-Jensen, L. & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. In *Language Resources and Evaluation*, Volume 43, Number 3, Springer Netherlands, pp. 269–299.
- Simonsen, H.K. (2017). Lexicography: What is the Business Model? In Kosem, I., Tiberius, C., Jakubíček, M., Kallas J., Krek, S. & Baisa, V. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 conference*. Brno: Lexical Computing CZ s.r.o., pp. 395–415.
- Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In Bergenholtz, H., Nielsen, S. & Tarp, S. (eds.) *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang AG, International Academic Publishers, pp. 17–32.
- Trap-Jensen, L. (2014.) *Leksikografisk tradition og fornyelse: tre revolutioner på 100 år?* (Lexicographic tradition and innovation: three revolutions in 100 years). In Fjeld, R.V. & Hovdenak, M. (eds.) *Nordiske Studier i leksikografi* 12. NFL-skrift nr. 13. Oslo: Novus Forlag, pp. 42–68.
- Woordenboek der Nederlandsche taal. Accessed at: <http://gtb.inl.nl/search> [30/04/2018]]



# PAPERS



# **Research into Dictionary Use**





# Investigating the Dictionary Use Strategies of Greek-speaking Pupils

**Elina Chadjipapa**

*Democritus University of Thrace*

*E-mail: elinaxp@hotmail.com*

## Abstract

The purpose of this large-scale study was to determine the profile of Greek pupils as dictionary users. In particular, the study investigates the dictionary use strategies that Greek pupils adopt, and records those that they prefer in total and by category while using a dictionary. A total of 745 pupils attending the last three years of primary school and the first three years of junior high school participated in a survey that was carried out in 2014. The data was collected by using the S.I.D.U., a self-report questionnaire. The results revealed that Greek pupils cannot be characterized as strategic dictionary users, as the mean scores of all categories of the dictionary use strategies were below 3.4, which is considered to reflect medium usage. Furthermore, the participants stated that they prefer to employ the look-up and selection strategies more than the lemmatization and the awareness strategies. The medium scores of strategic dictionary use indicate that Greek pupils need training in order to become strategic users.

**Keywords:** dictionary use strategies, selection strategies, awareness strategies, lemmatization strategies, look-up strategies, strategic dictionary users

## 1 Introduction

A dictionary is a useful learning tool which provides linguistic information to users. It is not just a classroom tool, but can constantly accompany users in different circumstances (Kent 2011). However, pupils might choose to adopt various other strategies when they encounter a difficulty in the learning process, such as “guessing” or “asking the teacher” instead of using a dictionary, and this could be due to lack of education about using a dictionary properly. As such, students have to be taught how to use a dictionary effectively, and “the training should start early in a course of learning a language and should form an integral part of pupils’ learning” (Hurman & Tall 1998).

This approach is not applied in the Greek learning system, since very few pedagogical dictionaries have been published; dictionary use training is limited to the second grade of primary school (Gavriilidou, Sfyroera, & Beze 2006), and there is complete lack of it at the university level (Anastasiadis-Symeonidis 1999), and this may people do not know how to use a dictionary efficiently in the classroom. This is a vicious circle that should be addressed by educating users how to use a dictionary strategically.

Training users requires first of all investigating which dictionary use strategies they adopt with the intention to understand their needs, preferences, habits and so on. The present study attempts to profile Greek users in order to describe the situation in Greek elementary and secondary schools. The theoretical part of the study which examined dictionary use with regard to learning strategies and the effective dictionary use contributes to the interpretation of the results of this work.

## 2 Literature review

### 2.1 Dictionary use as a learning strategy

Dictionary use is considered as a cognitive strategy (O'Malley & Chamot 1990; Oxford 1990). According to O'Malley and Chamot (1990:44) cognitive strategies “operate directly on incoming information, manipulating it in ways that enhance learning” and according to Oxford (1990:37) they “enable learners to understand and produce new language by many different means”. On the other hand, Cohen and Weaver (2006) offer a more practical point of view by classifying learning strategies by language skills, which includes receptive skills such as listening and reading and productive skills such as speaking and writing. They also mention skill-related strategies, such as vocabulary learning and translation. The skills-based inventory of language strategy use (Cohen, Oxford and Chi 2001) categorizes dictionary use in terms of reading strategies, writing strategies and vocabulary learning strategies. Nation (2001) states that dictionary use affects the three basic learning processes: a) reception, b) production and c) vocabulary learning.

Researchers have investigated dictionary use as a reading strategy (Hosenfeld 1977; Barnett 1988; Carrell 1989; Sheorey & Mokhtari 2001; Cohen, Oxford & Chi 2002; Cohen & Oxford 2002) and showed that it has positive effects on reading comprehension. Moreover, some researchers deal with the effects of receptive dictionary use on L2 or FL reading (Bensoussan et al. 1984; Tono 1989; Nesi & Meara, 1991; Luppescu & Day 1993; Knight 1994; Nesi & Haill 2002), although just one (Miller 1995) focuses on L1. The majority of related studies (Tono 1989; Luppescu & Day 1993; Knight 1994; Nesi & Haill 2002) have found that the receptive dictionary use helps users perform better in reading comprehension.

Research which focuses on productive dictionary use is quite limited (Ard 1982; Christianson 1997; Harvey & Yuill 1997; Chun 2004; Santos 2006; Elola, Rodríguez-García & Winfrey 2008; Wolfer et al. 2016). It has to be reported that students prefer to use dictionaries more in receptive circumstances than in productive ones (Tomaszczyk 1979; Béjoint 1981; Scholfield 1982; Γαβριηλίδου 2002). Nesi (1987) agrees with this, and raises the reasonable question of whether this preference is due to users' incomplete knowledge of dictionary use, or to the incomplete information that dictionaries' microstructures provide related to written production. However, research has shown that productive dictionary use positively affects the writing process (Harvey & Yuill 1997). Cohen, Oxford and Chi (2002) and Cohen and Oxford (2002) include dictionary use in the lists of writing strategies, as it contributes to the cultivation and development of writing skills required for written production.

A lot of studies investigate the effects of dictionary use on vocabulary learning, and many categorize dictionary use among the various vocabulary learning strategies. Gu and Johnson (1996) classify vocabulary learning strategies into six categories, one of which is “dictionary strategies”. Schmitt (1997) categorizes dictionary use as one of the “determination strategies” which users adopt in order to discover the meaning of a new word, without asking for help, while Nation (2001) classifies dictionary use in the general class of strategies “sources: finding information about words”. Dictionary use is a vocabulary learning strategy that is often adopted by dictionary users (Fan 2003; Gu 2003; Asgari & Mustapha, 2011), and one that facilitates the process of learning new words (Summers 1988; Luppescu and Day 1993; Hulstijn, Hollander & Greidanu 1996) and helps users maintain new vocabulary in their long-term memory (Knight, 1994; Chen, 2011).

### 2.2 Effective dictionary use

Often dictionary users wonder which type of dictionary is the most appropriate to meet their needs during the learning process, and what are the steps to follow in order to get the required information.

Researchers have been examining these issues for decades, and thus explored user behaviours, opinions, ideas, difficulties, dictionary misuse, reference needs and skills, as well as the various dictionary use strategies. The identification of all these parameters contributes to profiling the users and finally to effective training in dictionary use.

By investigating the reasons why people choose to use a dictionary researchers can better understand the needs of the users. Hartmann and James (1998: 116) define reference needs as “the circumstances that drive individuals to seek information in reference works such as dictionaries”. Nation (2001) notes that a dictionary can be used for receptive and productive reasons, as well as for learning vocabulary. In order to fulfill those needs particular skills should be developed, and reference skills are “the abilities required on the part of the dictionary user to find the information being sought” (Hartmann and James 1998: 117). It is obvious that there are direct relationships between reference needs and skills, as the last are intended to meet the needs arising while using the dictionary. Users who develop these skills are able to conduct effective searches in the dictionary and enhance their language learning.

Theoretical studies describe the reference skills that users should develop to use a dictionary effectively (Roberts, 1997; Hartmann 1999; Thornbury 2002) and also classify them (Scholfield 1982; 1999; Nesi 1999; Lew & Galas 2008). Nesi (1999) suggests an exhaustive list of skills that a student might need to use dictionaries effectively, which are chronologically listed in six stages: (a) before study, (b) before dictionary consultation, (c) locating entry information, (d) interpreting entry information, (e) recording entry information, and (f) understanding lexicographical issues. All these are time-based except for the sixth, which includes general lexicography skills. Lew and Galas (2008) propose the categorization of skills in terms of: a) reference (alphabetical ordering, the ability to use a dictionary for equivalents, definitions, spelling, pronunciation and obtaining grammatical information; locating words using initial letters), b) inference (the ability to establish and interpret parts of speech; the ability to correctly interpret meanings; finding and handling meanings; grammatical awareness), c) understanding of vocabulary conventions (the awareness of dictionary features and layout knowledge of phonetic symbols; knowledge of parts of speech; word formation; derivatives; past forms; countable and uncountable nouns; awareness of idiomatic expressions; awareness of phrasal verbs; pronouns), and (d) acquiring extra information (the ability to obtain socio-cultural information). Users who have not developed sufficient reference skills will make errors while using a dictionary (Nesi & Meara 1994; Christianson 1997; Nesi 1999; Nesi & Haill 2002). These errors have been classified according to the learning context (Meara & English 1987; Maingay & Rundell 1987; Nuccorini 1994) and the time when they take place (Neubach & Cohen 1988). However, some researchers (Tickoo 1989; Hulstijn & Atkins, 1998) attribute the difficulties that users face during dictionary use to the content and the structure of the dictionaries.

A new approach to effective dictionary use is proposed by Gavriilidou (2012; 2013), who refers for the first time to Dictionary Use Strategies (DUSs) and not to reference skills. This earlier study focuses on the user’s conscious effort while searching in a dictionary, and defines DUSs as “techniques used by the effective dictionary user in order to make a successful search in the dictionary”. Based on the above characteristics, we could add more information and define DUSs as “techniques which the dictionary user adopts consciously, firstly for efficient dictionary consultation and secondly for self-regulation and autonomy while using the dictionary” (Χατζηπαπά 2018)

As the construct of DUSs is a relatively recent one, very few studies have been done so far to describe, record and classify these (Gavriilidou 2011; Gavriilidou 2012, 2013; Chadjipapa & Papadopoulou 2016; Chadjipapa & Papadopoulou to appear).

Gavriilidou (2014) classifies DUSs into four categories:

1. The awareness strategies, which lead to a decision to use a dictionary in order to resolve a problem encountered inside or outside the class
2. Selection strategies, which enable users to select an appropriate dictionary type depending on the problem to be solved and guarantee familiarity with one's own dictionary
3. Lemmatization strategies, which are strategies which help finding the citation form of inflected forms found in the text. Users should be able to use the morphological indices (stems, prefixes, suffixes, inflectional morphemes) of the unknown word that has been met in the text in order to make hypotheses about the look-up form of that word, or should be acquainted with alphabetical sequencing, otherwise lemmatization is not possible
4. Look-up strategies, which control and facilitate the localization of the correct part of the entry where different meanings of the same word form are included.

Effective dictionary use is influenced from either reference skills or the dictionary use strategies the user adopts, and from the information dictionary provides. Atkins and Varantola (1998) state there is a need to both improve dictionaries and train users. However, teaching users how to carry out reference skills or DUSs is perhaps the most important task, as the literature shows that dictionaries are improving day by day, becoming better structured and more user-friendly.

### 3 Research Method

#### 3.1 Purpose of the study

The purpose of this study was to assess if Greek pupils employ DUSs and to investigate which they prefer to adopt when they consult a dictionary. In other words the study attempts to describe the Greek users' profiles in order to ascertain if they are able to realize successful and effective look-ups in a dictionary. The investigation of these particular DUSs can help us to identify the related reference needs that the participants want to satisfy, which of them are due to incomplete reference skills, and which are due to the poor structures of their dictionaries. In addition, the results of the present study lead to conclusions about the necessity to teach DUSs in Greek schools at in all education levels, and this is also the first work that explores strategic dictionary use among pupils of particular ages.

#### 3.2 Participants

The research was conducted between February and May 2014. The participants were pupils that were attending the fourth, fifth and sixth grades of Greek primary school, and the first, second and third grades of Greek junior high schools. In Greek education, the distribution of the grades is based on the age of the pupils<sup>1</sup>, in contrast to other countries educational systems, where the distribution of the grades is based on a combination of age and learning level. The pupils that participated in the research were studying in various Greek public schools, and the total of 745 pupils were collected using convenience sampling. In terms of gender, the pupils were almost equally distributed (see Table 1).

Table 1: Sample distribution by gender.

Gender	Pupils	%
Male	360	51,7
Female	385	48,3
Total	745	100,0

<sup>1</sup> Pupils in the Greek educational system start primary school at the age of 5.5-6.5 years, so at the age of 8.5-9.5 years old are studying in grade 4 of primary school, pupils aged 9.5 -10.5 years old in grade 5 and 10.5-11.5 years old in the 6th grade. As for the junior high school, pupils aged from 11.5-12.5 years are attending the 1st grade, from 12.5-13.5 the 2nd grade and from 13.5-14 the 3rd grade.

With regard to the age/grade, 13.3% of the pupils were attending the fourth grade, 16.5% the fifth grade and the 14.4% the sixth grade of primary school, while 18.3% were attending the first grade, 18% the second grade and 19.6% the third grade of junior high school (see Table 2).

Table 2: Sample distribution by age/class.

Grade	n	%
4 <sup>th</sup> grade	99	13,3
5 <sup>th</sup> grade	123	16,5
6 <sup>th</sup> grade	107	14,4
1 <sup>st</sup> grade	136	18,3
2 <sup>nd</sup> grade	134	18,0
3 <sup>rd</sup> grade	146	19,6
Total	745	100,0

The majority of the participants (94.5%) speak Greek as native language and less than 3% speak as native languages Albanian, Turkish, Russian and so on. The schools that participated in the research are located in Greek cities (Kavala, Komotini, and Chalkida) and on Greek islands (Lesvos and Chios). The primary schools that participated in the study were the Primary School of Polichnitos in Lesvos, the First Primary School of Lesvos, the Fifth Primary School of Komotini, the First Primary School of Chalkida, and the Primary School of Paleochori in Kavala. The high schools that participated in the study were the Second High School of Mitilini, the High School of Antissa and Petra in Lesvos, the Music High School of Komotini and the First High School of Chios.

### 3.3 Operationalization

In order to conduct the research, the Strategy Inventory for Dictionary Use - S.I.D.U. (Gavriilidou 2012; 2013) was used. The S.I.D.U. questionnaire is a newly-developed self-report instrument that is both internationally standardized and reliable. It has been used in Greek for evaluating strategic dictionary use, mainly focused on printed dictionaries<sup>2</sup>. It consists of 36 items measured on a five-point Likert-scale (never or almost never true of me, generally not true of me, somewhat true of me, generally true of me, always true of me)<sup>3</sup>, and the questions are divided into four categories as Gavriilidou (2013) organizes them: (a) awareness strategies (14 questions, 1-14), (b) selection strategies (seven questions, 15-21), (c) lemmatization (seven questions, 22-28) and (d) look-up strategies (eight questions, 29-36). S.I.D.U. gives the present study the ability to evaluate strategic dictionary use from a quite large sample in manner that is easy, fast and economical, while helping to profile individuals or groups of users.

The data collection started after the principals' and teachers' agreement, and of course the agreement of the pupils' parents at all school units. Participants completed the Greek version of S.I.D.U. and followed the instructions that the researcher gave them in most cases, or the teachers' instructions when the researcher could not be present. When the researcher was not present, the teachers received clear instructions how to present the purpose of the study to the pupils and how to complete the questionnaire. The participants had 45 minutes to fill in the questionnaire, but none of them needed more than 30 minutes, even those pupils attending primary school.

2 S.I.E.D.U. (Strategy Inventory Electronic Dictionary Use) assesses users' skills and strategies in online electronic dictionary searches (Gavriilidou & Mavrommatidou 2016).

3 The S.I.D.U. is standardized in English (Gavriilidou 2014).



It is important to note that this particular instrument describes the DUSs that users declare they adopt. That means that the answers that pupils provide could reveal either the real actions that they use when using a dictionary or the actions that they wish to employ while consulting a dictionary.

## 4 Results

IBM SPSS Statistics version 21.0 was used for the data analysis. The questions were encoded by assigning specific numbers to each category of the five-point Likert scale, for example, “never or almost never true of me” was coded 1, “generally not true of me” 2, “somewhat true of me” 3, “generally true of me” 4, and “always true of me” 5.

Cronbach’s coefficient alpha analysis was performed to check the reliability of each category with regard to internal consistency.. In addition, the correlations between each category (sub-scale) and total score were computed, and scores higher than 0.3 were considered satisfactory. Synthetic variables (one for each strategy group) were then constructed for further analyses. Descriptive statistics (frequencies, averages, and standard deviations) were used to investigate the frequency of dictionary strategy use by the respondents.

The findings indicate excellent internal consistency ( $\alpha = 0.93$ ) with regard to the total scale, with results very close to those of Gavriilidou (2013). Likewise, all four sub-scales indicate acceptable to good reliability, ranging from 0.784 to 0.881. In addition, the correlations of each sub-scale within the overall scale for all strategies groups were found to be from 0.350 to 0.645, and higher than 0.30 (see Table 3).

Table 3: Results of reliability testing.

Scale	Items	Correlation	Cronbach’s $\alpha$
Awareness strategies	14	0.497 – 0.592	0.881
Selection	7	0.350 – 0.574	0.792
Lemmatization	7	0.431 – 0.541	0.784
Look-up	8	0.437 – 0.645	0.828
Total	36	0.350 – 0.645	0.930

The results of the descriptive analysis of the participants’ responses are presented in Tables 4 to 8.

Initially, it was found that the strategic dictionary use was in the range of the medium scores (mean 3.05, see Table 4). Among the four categories of DUSs, the participants stated that they are more likely to employ look-up strategies, with an average of 3.39, and dictionary selection strategies with an average of 3.31, while they stated that they less often adopt the awareness strategies (average = 2.55) and the lemmatization strategies (mean = 2.89) (see Table 4).

Table 4: MS and SD of the four categories of the dictionary use strategies

Dictionary use strategies	Mean Score	Standard Deviation
Awareness strategies	2.55	0.80
Selection strategies	3.31	0.99
Lemmatization strategies	2.89	0.94
Look-up strategies	3.39	0.91
Total	3.05	0.74

In relation to dictionary awareness strategies (see Table 5), the pupils stated that they use a dictionary mainly at home (M.S. = 3.29) and usually to help themselves with translation (3.02). They use a dictionary to find the meaning of a word (2.97), while they rarely use a dictionary when they read a text (2.14) or to find antonyms (2.22).

Table 5: Frequencies of dictionary awareness strategies' use.

Dictionary awareness strategies	Mean Score	Standard Deviation
1. I use a dictionary to find the meaning of a word	2.97	1.25
2. I use a dictionary to find the spelling of a word	2.50	1.33
3. I use a dictionary to find synonyms	2.49	1.31
4. I use a dictionary to find antonyms	2.22	1.25
5. I use a dictionary to check how a word is used	2.37	1.36
6. I use a dictionary to find the origin of a word	2.44	1.36
7. I use a dictionary to help myself in translation	3.02	1.48
8. I use a dictionary to find the syntax of a word	2.25	1.25
9. I use a dictionary to find the derivatives of a word	2.47	1.30
10. I use a dictionary to find word families	2.61	1.42
11. I use a dictionary to find the meaning of an expression	2.71	1.36
12. I use a dictionary at home	3.29	1.38
13. I use a dictionary when I read a text	2.14	1.26
14. I use a dictionary when I write a text	2.37	1.34
Total	2.55	0.80

While selecting a dictionary users reported (see Table 6) that they know why they need it (3.90) and they know what bilingual dictionary is and what it is used for (3.73). They also stated that they know what an etymological (3.11), general (3.37) and dictionary of technical terms are (2.79) and what they are used for.

Table 6: Frequencies of dictionary selection strategies' use.

Dictionary selection strategies	Mean Score	Standard Deviation
15. Before I buy a dictionary, I know the reason why I need it	3.90	1.46
16. Before I buy a dictionary at the bookshop, I glance through it to see what information it provides	3.12	1.53
17. I choose a dictionary because it has a lot of entries and a lot of information in each entry	3.22	1.42
18. I know what an etymological dictionary is and what it is used for	3.11	1.51
19. I know what a general dictionary is and what it is used for	3.37	1.49
20. I know what a bilingual dictionary is and what it is used for	3.73	1.52
21. I know what a dictionary of technical terms is and what it is used for	2.79	1.52
Total	3.31	0.99

Regarding the frequencies of use of lemmatization strategies (see Table 7), the participants reported that when they come across an unknown word in a text they often try to think in what form they should look for it in the dictionary (3.37), when they hear a word that they do not know, they consider various spelling possibilities and look it up in the dictionary accordingly (3.11). They also claimed that when they do not find a word where they believed it would be they begin a new search with other criteria until they

find it (3.09). In addition, the students reported that they are not used to carefully studying the abbreviations of a new dictionary (2.65) or to carefully reading the introduction (2.67), and they sometimes use the usage labels provided in the entry to see how a word is used in spoken language (2.54).

Table 7: Frequencies of lemmatization strategies' use.

Lemmatization strategies	Mean Score	Standard Deviation
22. Before I use my new dictionary, I carefully read the introduction	2.67	1.55
23. Before I use my new dictionary, I carefully study the list of abbreviations	2.65	1.45
24. When I come across an unknown word in a text, I try to think in what form I should look it up in the dictionary	3.37	1.44
25. When I can't locate a proverb or a set phrase in the entry where I thought I would find it, I begin a new search	2.90	1.44
26. When I hear a word I don't know, I consider various spelling possibilities and look it up accordingly	3.11	1.43
27. When I can't find a word where I thought I would find it, I begin a new search until I find it	3.09	1.41
28. To see how a word is used in spoken language, I use the usage labels provided in the entry	2.54	1.29
Total	2.89	0.94

When they look up for a word (see Table 13), the pupils stated that they consider its initial letter and then look up where they think this letter is in the dictionary (3.97). They also reported that they always have the word in mind during the search (3.83). Finally, they stated that before they use a word they found in the dictionary when writing a text, they read all the information on the grammar of that word (conjugation, syntax) to be sure of the correct usage (3.01).

Table 8: Frequencies of look-up strategies' use.

Look-up strategies	Mean Score	Standard Deviation
29. When I look up a word beginning with E, I search in the first quarter pages as E is one of the first letters of the alphabet	3.50	1.50
30. When I look up a word beginning with L, I open my dictionary in the middle	3.52	1.39
31. When I look up a word, I bear in mind its initial letter and then search where I believe this initial letter is in the dictionary	3.97	1.34
32. When I look up a word, I simply open the dictionary and see if I am near the specific initial letter	3.30	1.43
33. When I look up a word, I constantly bear it in my mind during the search	3.83	1.39
34. When I realize that the word I am looking for has various different meanings, I go through them all one by one, assisted by the example sentences	3.33	1.40
35. When I find the word that I was searching for, I return to the text to confirm that the word matches the context	3.45	1.41
36. Before I use a word I found in the dictionary when writing a text, I read all the information on the grammar of that word (conjugation, syntax) to be sure of the correct usage.	3.01	1.49
Total	3.39	0.91



To sum up, we have to say that the Greek pupils examined in this work had medium range scores for the dictionary use strategies in the four categories, and that they adopt look-up strategies and dictionary selection strategies more frequently, while they seldom adopt the awareness and the lemmatization strategies.

## 5 Discussion

The present study has investigated the dictionary use strategies that pupils of the last three grades (fourth to sixth) of primary school and the first three grades (first to third) of junior high school adopt by using the S.I.D.U. The results indicate that the pupils' DUSs are in the medium range, and thus that Greek pupils might not be very competent dictionary users, and so their efforts in this regard might not be very effective. The reasons for the non-strategic dictionary use found in this work may be as follows:

- The participants of the survey do not consciously adopt the DUSs.
- The Greek school system does not systematically train users in strategic dictionary use, and teachers do not create opportunities for effective dictionary use.
- There are few pedagogical dictionaries with regard to Greek lexicography.

We can also assume that Greek users do not have sufficient knowledge of effective dictionary use, as a dictionary is not a reference or educational tool in Greek schools. One of the reasons that dictionary use is neglected in the Greek classroom could be the beliefs of the Greek teachers. Chadjipapa and Papadopoulou (to appear) show that Greek teachers in both primary and secondary schools, while claiming that they are "efficient" dictionary users, rarely use a dictionary in the classroom. Thus, if we take into consideration the issue of lifelong learning in the educational process we can conclude that the moderate use of DUSs by Greek pupils could be due to the incomplete use of a dictionary by their teachers. Undoubtedly, non-strategic dictionary use occurs because there is no systematic incorporation of dictionary use strategies into the Greek educational process. Except for very few attempts to integrate DUSs in Greek schools (Gavriilidou, Sfyroera, & Beze 2006) and to inform pupils of the basic functions of a dictionary (Αγγελάκος, Κατσαρού, Μαγγανά 2006; Γαβριηλίδου, Εμμανουηλίδης, Πετρίδου-Εμμανουηλίδου 2006), there has been little effort to improve this situation at any educational level. As such, the pupils' training in DUSs depends on their teachers' intentions.

Further conclusions can be drawn by examining each category of dictionary use strategy. The results showed that pupils stated they adopt the look-up and the selection strategies more frequently, and the lemmatization and the awareness strategies less frequently. The results of the present study are consistent with those of Chadjipapa and Papadopoulou (2016) and Chadjipapa and Papadopoulou (to appear) that explored the DUSs adopted by students and teachers, respectively. Both studies have shown that users encounter difficulties in using the lemmatization and awareness strategies.

This also indicates that pupils know how to use a dictionary effectively but do not know what kind of information they can search for and find in a dictionary. As regards the awareness strategies, the respondents stated that they use a dictionary in order to help themselves with translation issues. That means that Greek pupils use the dictionary mainly for learning FL or L2, and less for learning L1.

Another statement that it is worth mentioning is that the pupils stated that they prefer to use a dictionary to find the meaning of a word or an expression, and much less often use a dictionary to find the spelling, etymology, or syntax of a word, synonyms, opposites and derivatives of a word (awareness strategies). That is very interesting, because the students also stated that they use a dictionary more for productive purposes and less for comprehension reasons. This contradiction shows what the

pupils would like to use a dictionary for, and maybe that they do not know that spelling, syntax and so on are related to written production, a demanding process which requires specific skills and strategies. The statistical analyses revealed that look-up strategies are at the top of the users' preferences. In spite of pupils' medium range scores in strategic dictionary use we speculate that they have developed certain metacognitive abilities, which indicates that after a short time of training in dictionary use they would be able to choose DUSs consciously and make successful look-ups. The second choice is the dictionary selection strategies. The pupils claimed that they know all types of dictionaries, except from the type with technical terms. This is probably because they are not familiar with this particular type and because there are not many such dictionaries on the Greek market. Finally, with regard to the lemmatization strategies, the pupils seemed to more often use those that are related to the macrostructure of the dictionary (a. when I come across an unknown word in a text, I try to think in what form I should look it up in the dictionary; b. when I hear a word I don't know, I consider various spelling possibilities and look it up accordingly; c. when I can't find a word where I thought I would find it, I begin a new search until I find it) than the strategies that are related to the megastructure and the microstructure (a. to see how a word is used in spoken language, I use the usage labels provided in the entry; b. before I use my new dictionary, I carefully study the list of abbreviations; c. before I use my new dictionary, I carefully read the introduction).

## 6 Limitations of the study

The results of this research are based on the S.I.D.U., a self-report instrument that describes the dictionary use strategies that users declare they adopt. Consequently, their statements may differ from what they actually do in the learning process, and may instead express their willingness to be strategic user or to give a false impression about their skills. The verification of the results with other self-referencing tools, such as oral protocols or interviews, but also with non-self-referencing methods such as observation or their combination (Dörnyei 2007), would thus provide greater reliability and validity to the research. However, triangulation is a time consuming process that is quite difficult to carry out when working with numerous subjects.

Finally, it should be noted that convenience sampling was used for data collection. Certainly random sampling enhances the external validity of any study, and requires more time and cost; however, the particular group examined in this work was collected from the Greek mainland and islands, as well as from urban and non-urban areas, which contributes to the reliability of the results.

## 7 Pedagogical implementations and further investigations

The medium scores of DUSs' adoption found in this work indicate that there is a need to teach such strategies to Greek students and incorporate dictionary use in the educational process. The integration of DUSs into the educational process would ensure more effective dictionary use and successful look-ups. Several researchers focus on the need to teach how to use a dictionary effectively, and suggest training by enhancing the users' reference skills or DUSs (Ard 1982; Béjoint & Moulin 1987; Herbst and Stein 1987; Gavriilidou 2017). Some other studies have investigated how intervention programs contribute to the development of reference skills or DUSs, and thus more effective dictionary use (Kipfer 1987; Bishop 2001; Głowacka 2001; Carduner 2003; Chi 2003; Lew & Galas 2008, Gavriilidou 2017). In the Greek literature, there are studies that propose intervention exercises or programs in the educational process, but without any control groups to show their effectiveness (Αναστασιάδη-Συμεωνίδη 1997; Γαβριηλίδου 2000; Gavriilidou, Sfyroera, & Beze 2006;

Νικηφοράκη 2003; Ευθυμίου & Μητσιακή 2007; Γαβριηλίδου, Γιούλη και Λαμπροπούλου 2008; Ευθυμίου 2013). Finally, we have to note that training in effective dictionary use should include entertaining and interesting activities for users (Alhaysony 2011), such as role-plays (Wright 1998), and that the teaching of DUSs should be integrated into the language teaching process by informed instructors (Gavriilidou 2017).

Teachers often suggest dictionary use during the educational process and encourage look-ups in class, but barely know their pupils' reference needs and if they are effective dictionary users (Neubach & Cohen 1988). Chadjipapa and Papadopoulou (to appear) has also shows that a very small percentage of teachers, teaching in Greek schools at both educational levels (primary and secondary), use a dictionary in the classroom. This may be due to the teachers' "non-strategic" dictionary use, as well as to their belief that dictionary use does not offer particular benefits during the learning process. The researchers thus believe that Greek teachers need further training, as their beliefs about dictionary use may influence their students' opinions about strategic dictionary use.

As such, the investigation of how interventional programs impact strategic dictionary use and how Greek teachers' beliefs influence their students' beliefs will contribute to the completion of the Greek dictionary users' profiles carried out in this work. It would also be interesting to investigate if there are any correlations between the users' effective dictionary use and their performance in school. Finally, it is important to understand the relations between the use of DUSs and cognitive and metacognitive skills, and this is another direction for future research.

## 8 Conclusion

Dictionary use contributes to the language learning process by enhancing language skills. Effective dictionary use implies the adoption of dictionary use strategies, which are part of a student's language learning strategies.

This study attempted to investigate the dictionary use strategies that Greek pupils studying in the three last grades of primary school and in the first three grades of junior high school adopt, and to capture the profiles of the specific users, based on the S.I.D.U. self-report instrument. The moderate use of DUSs, as users of all grades claim, demonstrates the incomplete integration of dictionaries as a reference and educational tool in Greek schools. However, the users' preferences for the look-up and selection strategies indicates high levels of metacognitive ability.

The results of this work help reveal dictionary use in the Greek educational system. The profile of the "non-strategic" user found in this work calls for more dictionary use training and its integration in the educational process, and also the improvement of the Greek pedagogical dictionaries. It should also be noted that improved strategic dictionary use is also an issue for teachers, as a well-trained teacher who is considered an effective user can appreciate the contribution of dictionary use to the learning process, and thus teach how to use such tools effectively.

The present study hopes to reinforce the integration of dictionary use in the Greek educational process and be the basis for further studies on dictionary use strategies.

## References

- Alhaysony, M. (2011). Dictionary-use Difficulties Encountered by Saudi Female English-Major Students. *European Journal of Scientific Research*, 65 (1), pp. 109-120.

- Anastasiadis-Simeonidis, A. (1999). The Dictionary Scene in Greece. In R. R. Hartmann (Ed.), *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the Thematic Network Project in the Area of Languages, Sub-Project 9: Dictionaries*, (pp. 22-23).
- Ard, J. (1982). The use of bilingual dictionaries by ESL students while writing. *ITL Review of Applied Linguistics*, 58, pp. 1–27.
- Asgari, A., & Mustapha, A. (2011). The Type of Vocabulary Learning Strategies Used by ESL Students in University Putra Malaysia. *English Language Teaching*, 4.2, pp. 84-90.
- Atkins, B. T., & Varantola, K. (1998a). Language learners using dictionaries: The final report on the EURALEX/AILA Research Project on Dictionary Use. In B. Atkins (Ed.), *Using dictionaries. Studies of dictionary use by language learners and translators. (Lexicographica Series Maior 88.)* (pp. 21-81). Tübingen: M. Niemeyer.
- Barnett, M. (1988). Reading through context: how real and perceived strategy use affects L2 comprehension. *The Modern Language Journal*, 72, pp. 150-162.
- Béjoint, H. (1981). The foreign student's use of monolingual English dictionaries: a study of language needs and reference skills. *Applied Linguistics*, 2 (3), pp. 207–222.
- Béjoint, H., & Moulin, A. (1987). The place of the dictionary in an EFL programme. In A. Cowie (Ed.), *The dictionary and the language learner: Papers from the EURALEX seminar at the University of Leeds 1-3 April 1985* (pp. 97-114). Tübingen: Max Niemeyer Verlag.
- Bensoussan, M., Sim, D., & Weiss, R. (1984). The Effect of Dictionary Usage on EFL Test Performance Compared With Student and Teacher Attitudes and Expectations. *Reading in a Foreign Language*, 2, pp. 262-76.
- Bishop, G. (2001). Using quality and accuracy ratings to quantify the value added of a dictionary skills training course. *Language Learning Journal*, 24: 1, pp. 62 - 69.
- Carduner, J. (2003). Productive Dictionary Skills Training: What Do Language Learners Find Useful? *Language Learning Journal*, 28, pp. 70-76.
- Carrell, P. (1989). Metacognitive awareness and second language reading. *Modern Language Journal*, 73, pp. 121-134.
- Chadjipapa, E., & Papadopoulou, E. (2016). Investigating Greek student's dictionary use strategies. *Selected Papers of the 21st International Symposium on Theoretical and Applied Linguistics (ISTAL 21)*, (pp. 516-526). School of English, Aristotle University of Thessaloniki.
- Chadjipapa, E., & Papadopoulou, L. (to appear). Profiling Greek teachers: Dictionary use. London.
- Chen, Y. (2011). Dictionary Use and vocabulary learning in the context of reading. *International Journal of Lexicography*, pp. 1-32.
- Chi, M. L. (2003). *An Empirical Study of the Efficacy of Integrating the Teaching of Dictionary Use into a Tertiary English Curriculum in Hong Kong*. Hong Kong: Language Centre, Hong Kong University of Science and Technology.
- Christianson, K. (1997). Dictionary use by EFL writers: What really happens? *Journal of Second Language Writing*, 6 (1), pp. 23– 43.
- Chun, Y. V. (2004). EFL Learners' Use of Print and Online Dictionaries in L1 and L2 Writing Processes. *Multimedia-Assisted Language Learning*, 7(1), pp. 9–35.
- Cohen, A. D., & Oxford, R. L. (2002). Young Learners' Language Strategy Use Survey. (pp. 75-78). Minneapolis, MN: University of Minnesota Center for Advanced Research on Language Acquisition. Available at: <https://fdde40f2-a-551982af-s-sites.googlegroups.com/a/umn.edu/andrewdcohen/documents/2002-Cohen%26OxfordYongLearners%27LgStratUseSrvy.pdf?attachauth=ANoY7cpB8CcmbeV> [27/2/2016].
- Cohen, A. D., & Weaver, S. J. (2006). *Styles and strategies-based instruction: A teachers' guide*. Minneapolis, MN: Center for Advanced Research on Language Acquisition University of Minnesota.
- Cohen, A. D., Oxford, R. L., & Chi, J. C. (2002). Language Strategy Use Survey. (pp. 68-74). Minneapolis, MN: Center for Advanced Research on Language Acquisition, University of Minnesota. Available at: <https://fdde40f2-a-551982af-s-sites.googlegroups.com/a/umn.edu/andrewdcohen/documents/2002-Cohen%26Oxford%26ChiLanguageStrategyUseSurvey.pdf?attachauth=ANoY7crg> [27/2/2016].
- Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: OUP.
- Elola, I., Rodríguez-García, V., & Winfrey, K. (2008). Dictionary use and vocabulary choices in L2 Writing. *Estudios de lingüística Inglesa aplicada*, 8, pp. 63-89.
- Fan, M. (2003). Frequency of Use, Perceived Usefulness and Actual Usefulness of Second Language Vocabulary Strategies: A study of Hong Kong learners. *The Modern Language Journal*, 87(2), pp. 222-241.
- Gavrilidou, Z. (2011). Profiling Greek Adult Dictionary Users. *Studies in Greek Linguistics*, 31, pp. 166-172.



- Gavriilidou, Z. (2012). Construction, validity and reliability of the Strategy Inventory for Dictionary Use. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki, & P. Kambaki-Vougioukli (Ed.), *Selected papers of the 10th International Conference of Greek Linguistics*. Komotini: Democritus University.
- Gavriilidou, Z. (2013). Development and validation of the strategy Inventory for Dictionary Use (S.I.D.U). *International Journal of Lexicography*, 22(2), pp. 135-154.
- Gavriilidou, Z. (2014a). User's abilities and performance in dictionary look-up. In N. Lavidas, T. Alexiou, & A. Sougari (Ed.), *In Major Trends on Theoretical and Applied Linguistics, Vol. 2*, pp. 41-52.
- Gavriilidou, Z. (2017). The effect of an intervention programme on improving electronic dictionary reference skills of students attending secondary schools in Greece. *Oral presentation in Elex 2017, Electronic Lexicography from Scratch*. Leiden 19-21 September 2017.
- Gavriilidou, Z., & Mavrommatidou, S. (2016). Construction of tool for the identification of electronic dictionary users skills. In T. Margalitadze, & G. Meladze (Ed.), *Proceedings of the XVII Euralex International Congress, Lexicography and Linguistic Diversity*, (pp. 168-178). Tbilisi State University.
- Gavriilidou, Z., Sfyroera, M., & Beze, L. (2006). *A trip into language's world [IN GREEK], School book for Grade B Elementary*. Organization of Publication of School Books.
- Głowacka, W. (2001). *Difficulties with understanding dictionary labels experienced by Polish learners of English using bilingual dictionaries*. Adam Mickiewicz University: M.A. Dissertation.
- Gu, P. (2003). Vocabulary Learning in a Second Language: Person, Task, context and Strategies. *Teaching English as a Second or Foreign Language, TESL-EJ*, 7(2).
- Gu, P. Y., & K., J. R. (1996). Vocabulary learning strategies and language learning outcomes. *Language Learning*, 46(4), pp. 643-79.
- Hartmann, R. R. (1999). Case Study: The Exeter University Survey of Dictionary Use. In R. R. Hartmann (Ed.), *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the Thematic Network Project in the Area of Languages Sub-Project 9: Dictionaries* (pp. 36-52). Berlin: Freie Universität Berlin.
- Hartmann, R. R., & James, G. (1998). *Dictionary of Lexicography*. London. New York: Routledge.
- Harvey, K., & Yuill, D. (1997). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics*, 18(3), pp. 253- 278.
- Herbst, T., & Stein, G. (1987). Dictionary-using skills: A plea for new orientation in language teaching. *The Dictionary and the Language Learner*. Tübingen: Max Niemeyer Verlag.
- Hosenfeld, C. (1977). A preliminary investigation of the reading strategies of successful and unsuccessful second language learners. *System*, 5, pp. 11-123.
- Hulstijn, J. H., & Atkins, B. T. (1998). Empirical research on dictionary use in foreign language learning: Survey and discussion. In B. T. Atkins (Ed.), *Using L2 dictionaries: Studies of dictionary use by language learners and translators*, (pp. 7-19).
- Hulstijn, J., Hollander, M., & Greidanu, S. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, 80(3), pp. 327-339.
- Hurman, J., & Tall, G. (1998). *The Use of Dictionaries in GCSE Modern Foreign Languages Written Examinations*. Birmingham: School of Education, University of Birmingham.
- Kent, D. B. (2001). A Survey of Korean University Freshmen Dictionary Use and Perceptions. *Korea TESOL Journal*, 4(1), pp. 73-92.
- Kipfer, B. A. (1987). Dictionaries and the Intermediate Student: Communicative Needs and the Development of User Reference Skills. In A. P. Cowie (Ed.), *The Dictionary and the Language Learner, Lexicographica Series Maior 17*, (pp. 44-54). Tübingen: Niemeyer.
- Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, 78(3), pp. 285-299.
- Lew, R., & Galas, K. (2008). Can dictionary use be taught? The effectiveness of lexicographic training for primary school level Polish learners of English. In E. a. Bernal (Ed.), *Proceedings of the XIII EURALEX International Congress* (pp. 1273-1285). Barcelona: Universitat Pompeu Fabra.
- Lupescu, S., & Day, R. R. (1993). Reading Dictionaries and vocabulary learning. *Language Learning*, 43(2), pp. 263-287.
- Maingay, S., & Rundell, M. (1987). Anticipating learners' errors: Implications for dictionary writers. In A. Cowie (Ed.), *The dictionary and the language learner: Papers from the EURALEX seminar at the University of Leeds* (pp. 128-135). Tübingen: Max Niemeyer Verlag.

- Meara, P., & English, F. (1987). Lexical errors and learners' dictionaries. *Technical Report*. Birkbeck College, London.
- Miller, S. M. (1995). Vocabulary development and context usage, Available at: <http://files.eric.ed.gov/fulltext/ED379642.pdf> [20/4/2017].
- Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge: CUP.
- Nesi, H. (1987). Do dictionaries help students write? In T. Bloor, & J. Norrish (Ed.), *Written Language*. London: CILT.
- Nesi, H. (1999). The Specification of Dictionary Reference Skills in Higher Education. In R. Hartmann (Ed.), *European Language Council. Thematic Network Project in the area of Languages. Sub-project 9: Dictionaries*, (pp. 53-67. Available at <http://www.fu-berlin.de/elc/tnp1/SP9dossier.doc> [21/4/2011].
- Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography*, 15(4), pp. 277-306.
- Nesi, H., & Meara, P. (1991). How Using Dictionaries Affects Performance in Multiple-choice EFL Tests. *Reading in a Foreign Language*, 8(1), pp. 631-643.
- Nesi, H., & Meara, P. (1994). Patterns of misinterpretation in the productive use of EFL dictionary definitions. *System*, 22(1), pp. 1-15.
- Neubach, A., & Cohen, A. D. (1988). Processing strategies and problems encountered in the use of dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 10, pp. 1 - 19.
- Nuccorini, S. (1994). On Dictionary Misuse. In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenberg, & P. Vossen (Ed.), *EURALEX 1994 Proceedings* (pp. 586-597). Amsterdam: New University.
- O' Malley, J., & Chamot, A. (1990). *Learning strategies in second language acquisition*. Cambridge, England: Cambridge University Press.
- Oxford, R. (1990). *Language Learning Strategies: What Every Teacher Should Know*. New York: Newbury House.
- Roberts, R. (1997). Using dictionaries efficiently. In *38th Annual conference of the American Translators Association*. Washington: American Translators Association.
- Santos, S. (2006). Dictionary use in L2 writing. *Memorias del III Foro Nacional de Estudios en Lenguas (FONAEL 2006)*, (pp. 1 -12. Available at: <https://www.yumpu.com/en/document/view/30706700/246-dictionary-use-in-l2-writing-1-dictionary-use-in-l2-written/12> [9/5/2011].
- Schmitt, N. (1997). Vocabulary Learning Strategies. In M. McCarthy, & N. Schmitt (Ed.), *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.
- Scholfield, P. (1982). Using the English Dictionary for Comprehension. *TESOL Quarterly*, 16, pp. 185-94.
- Scholfield, P. (1999). Dictionary use in reception. *International Journal of Lexicography*, 12(1), pp. 13-34.
- Sheorey, R., & Mokhtari, K. (2001). Differences in the Metacognitive Awareness of Reading Strategies among Native and Non-Native Readers. *System*, 29(4), pp. 431-449.
- Summers, D. (1988). The role of dictionaries in language learning. In R. Carter, & M. McCarthy (Ed.), *Vocabulary and language teaching* (pp. 111-125). London: Longman.
- Thornbury, S. (2002). *How to Teach Vocabulary*. Harlow: Pearson.
- Tickoo, M. (1989). Which dictionaries and why? Explaining some options. In M. L. Tickoo (Ed.), *Learners' Dictionaries. State of the Art* (pp. 184-203). Singapore: SEAMEO.
- Tomaszczyk, J. (1979). Dictionaries: users and uses. *Glottodidactica*, 12, pp. 103 - 119.
- Tono, Y. (1989). Can a dictionary help one read better? In G. James (Ed.), *Lexicographers and their works. (Exeter Linguistic Studies Volume 14)* (pp. 192-200). University of Exeter.
- Wolfer, S., Bartz, T., Weber, T., Abel, A., Meyer, C., Müller-Spitzer, C. and Storrer, A. (2016). The Effectiveness of Lexicographic Tools for Optimising Written L1-Texts. *International Journal of Lexicography*, 31(1). Available at: <https://academic.oup.com/ijl/article/31/1/1/2549205> [8/5/2018]
- Wright, J. (1998). *Dictionaries: Resource books for teachers*. Oxford, UK: Oxford University Press.
- Αγγελάκος, Κ., Κατσαρού, Ε., & Μαγγανά, Α. (2006). *Νεοελληνική Γλώσσα Α' Γυμνασίου Σχολικό εγχειρίδιο Βιβλίο Μαθητή*. Αθήνα: ΟΕΔΒ. Παιδ. Ινστιτούτο.
- Αναστασιάδη - Συμεωνίδη, Α. (1997). Η λεξικογραφία στην εκπαίδευση. *Πρακτικά του 2ου Πανελλήνιου Συνεδρίου για τη Διδασκαλία της Ελληνικής Γλώσσας* (pp. 149-176). Θεσσαλονίκη: Κώδικας.
- Γαβριηλίδου, Ζ. (Ed.). (2000). *Παιδική Λεξικογραφία και Χρήση Λεξικού στην Προσχολική και Πρώτη Σχολική Ηλικία* (p. 72). Ξάνθη: Υπηρεσία Δημοσιευμάτων Δ.Π.Θ.
- Γαβριηλίδου, Ζ. (2002). Η διερεύνηση των λόγων χρήσης λεξικού ως προϋπόθεση για τη διδασκαλία στρατηγικής χρήσης του λεξικού στην τάξη. In P. Kambaki (Ed.), *Η διδασκαλία της νέας ελληνικής ως μητρικής γλώσσας (Teaching Greek as L1)*, (pp. 45-60). Komotini.

- Γαβριηλίδου, Μ., Εμμανουηλίδης, Π., & Πετρίδου-Εμμανουηλίδου, Έ. (2006)). *Νεοελληνική Γλώσσα Β' Γυμνασίου*. Οργανισμός Εκδόσεως Διδακτικών Βιβλίων.
- Γαβριηλίδου, Μ., Λαμπροπούλου, Π., & Γιούλη, Β. (2008). Το ερμηνευτικό λεξικό για το γυμνάσιο. *Πρακτικά του 8ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας*, (pp. 684-698). Πανεπιστήμιο Ιωαννίνων, Τομέας Γλωσσολογίας, Τμήμα Φιλολογίας.
- Ευθυμίου, Α. (2013). *Η διδασκαλία του λεξιλογίου στο δημοτικό σχολείο: θεωρία και εφαρμογές*. Θεσσαλονίκη: Επίκεντρο.
- Ευθυμίου, Α., & Μητσιάκη, Μ. (2007). Το πρώτο μου λεξικό ως εργαλείο διδασκαλίας της ελληνικής σε αλλόγλωσσους. In Κ. Ντίνας, & Α. Χατζηπαναγιωτίδη (Ed.), *Πρακτικά διεθνούς συνεδρίου: Η Ελληνική Γλώσσα ως δεύτερη/ξένη. Έρευνα, Διδασκαλία, Εκμάθηση*. (pp. 237-260). Θεσσαλονίκη: University Studio Press.
- Νικηφοράκη, Μ. (2003). Η αξιοποίηση του Λεξικού στη Διδασκαλία της Μητρικής Γλώσσας. Εφαρμογή στη σημασιολογία. *Πρακτικά 23ης Ετήσιας Συνάντησης με θέμα: Μελέτες για την Ελληνική Γλώσσα*. Θεσσαλονίκη.
- Χατζηπαπά, Ε. (2018). *Διερεύνηση στρατηγικών χρήσης λεξικού μαθητών Δημοτικού και Γυμνασίου*. PhD Thesis, Κομοτηνή.

## Acknowledgements

I would like to express my special thanks to all the pupils who participated in this research and to their teachers who gladly let me into their classrooms and gave some of their time to this project.





# Everything You Always Wanted to Know about Dictionaries (But Were Afraid to Ask): A Massive Open Online Course

**Sharon Creese<sup>1</sup>, Barbara McGillivray<sup>2</sup>, Hilary Nesi<sup>1</sup>, Michael Rundell<sup>3,4</sup>, Katalin Sule<sup>4</sup>**

<sup>1</sup>Coventry University, <sup>2</sup>The Alan Turing Institute/University of Cambridge, <sup>3</sup>Lexical Computing, <sup>4</sup>Macmillan

E-mail: [creeses@uni.coventry.ac.uk](mailto:creeses@uni.coventry.ac.uk), [bm517@cam.ac.uk](mailto:bm517@cam.ac.uk), [h.nesi@coventry.ac.uk](mailto:h.nesi@coventry.ac.uk), [michael.rundell@lexmasterclass.com](mailto:michael.rundell@lexmasterclass.com), [k.sule@macmillaneducation.com](mailto:k.sule@macmillaneducation.com)

## Abstract

We have created a Massive Open Online Course (MOOC) about dictionaries and dictionary-making, to be hosted by FutureLearn. This paper discusses the design and development of this course, which is pitched at high school and undergraduate level participants as well as language enthusiasts around the world. The MOOC will answer questions such as: how dictionaries are made and how this process has changed over time; what goes into a dictionary and who decides; and what kinds of language evidence underpin the information which dictionaries provide. Participants will be encouraged to compare the quantity and quality of information in different types of dictionary, and will investigate corpus-based and computational lexicographic methods. It will also consider dictionary users' attitudes and common misconceptions, taking into account the requirements and habits of English language learners as well as fluent speakers. By the end of the course, participants will know about some of the latest trends in lexicographic research, the roles of language technology, corpora and crowdsourcing in the dictionary compilation process, the range of possible dictionary entry components, lexicographical choices and computational methods surrounding the selection and ordering of word meanings, and the content and wording of definitions.

**Keywords:** MOOC, massive online open course, dictionary skills, lexicography, history of dictionaries, dictionary-making, corpus linguistics, neologisms, dictionary inclusion criteria, dictionary typologies, lexicographic evidence, crowdsourcing, meaning and definition, corpus-based lexicography

## 1 Introduction

This paper describes the design and development of a Massive Open Online Course (MOOC) on dictionaries, provided by FutureLearn and produced by Coventry University in collaboration with the Alan Turing Institute and Macmillan Dictionaries. The MOOC aims to provide an introduction to the world of dictionaries to a broad, non-technical audience, which includes language teachers and students but also laypeople with an interest in dictionaries and language. It is due to run for the first time in the autumn of 2018, and to the best of our knowledge it is the first MOOC in the field of lexicography.

FutureLearn is a private company owned by The Open University which provides a platform for a variety of courses created at other (mostly UK) universities and institutions. FutureLearn courses generally run from between two to eight weeks and are repeated at regular intervals. Participants enroll for free, although in most cases they can pay a small fee to retain access to the course materials after the course has ended. Institutions hope to gain a reputational advantage from their investment in FutureLearn course development, for example by establishing a positive relationship with users and attracting them onto related fee-paying programs offering certification.

FutureLearn takes a constructionist approach to knowledge acquisition, as described by Sharples (2018) and Sharples et al. (2012), and is based on the view that all human learning involves interaction. FutureLearn courses therefore restrict teacher input while maximizing the opportunities for peer-to-peer conversation. Source materials are generally expected to be provocative and short, and are always followed by at least one task requiring learners to post their own responses and comment on contributions from other learners. Video and audio recordings typically last no longer than four minutes, and more often are less than two minutes long. Written ‘articles’ are generally only a few paragraphs in length, providing a summary of the key ideas under discussion; links can be made to more extensive open-access material, but learners are not expected to have the skills to negotiate longer texts unaided: further reading activities are usually heavily guided.

MOOC tasks generally fall into two categories; feedback is either provided by peers (a model suited to the sharing of experiences and opinions), or automatically in response to quiz questions (a model more suited to the acquisition of factual information). Participants can pay to receive a certificate if they have completed the course, but they are not graded in any way, and do not produce assignments that require evaluation by a subject expert. This approach suits FutureLearn MOOC environments, where the number of online learners vastly exceeds the number of available moderators; it works best when the focal learning outcome is either an alteration in learner attitudes or the acquisition of factual knowledge, neither of which require scaffolding from experienced practitioners over an extended period of time. However FutureLearn MOOCs can be embedded within more conventional educational programs, such as face-to-face degree modules, and can thus be used to supplement and expand the learning experiences of students working towards the acquisition of more complex skills. Our dictionary MOOC might be incorporated not only into linguistics or lexicography courses, but also into courses in many other disciplines where an understanding of dictionary types and contents would help learners to communicate and study more effectively.

FutureLearn courses are divided into ‘Weeks’, described as “personally meaningful study periods” (Sharples, 2017); the MOOC we describe here consists of six Weeks, each providing about four hours of study time. The content of each Week is generally characterized by a ‘big question’, designed to stimulate participants’ interest, and it is broken up into several topics or ‘Activities’, each with a defined goal. Within each Activity there are a series of learning tasks or ‘Steps’, each lasting about 20 minutes, which lead the way to the designated learning outcome. Activities are often introduced in a provocative way, and the first Step in an Activity is typically an invitation to discuss the topic. Reading or listening to input comes after this initial collaborative task, and is always followed by discussion and reflection. Later Steps in the Activity may include problem-solving, search engine browsing, investigating materials or authentic situations and co-creating or sharing artefacts. Learners on our MOOC will be asked to explore dictionary sites, for example, and create and share their own lexicographical material. Introductory textbooks such as Mugglestone (2011) cover some of the same ground, but the interactivity and interconnectivity of the MOOC makes for a rather different learning experience, adaptable to the learners’ own dictionary-using habits and contexts.

Participants work at their own pace and can access any part of the course at any point, so the MOOC learning environment changes constantly as more and more contributions are submitted. One or more moderators see to the day-to-day management of these contributions, although on courses with high participation rates they cannot be expected to read all the comments that users post. The Week’s activities are generally concluded with a round-up Step, asking participants to discuss what they think they have learnt so far. The course educators may also contribute summary postings at key points in the course.

Our MOOC is emphatically not a training course in dictionary-making or in the theoretical and computational research which underpins contemporary lexicography. Rather, it is intended to provide a

lively introduction to a topic about which non-experts often have strong opinions but little real knowledge. The course aims to bring users up to speed with recent developments in this rapidly-changing field. It will also challenge common misconceptions about dictionaries, such as the notion that it is a dictionary's role to pronounce on "correct" usage, or that a word's inclusion in "The Dictionary" confers some kind of official approbation. As well as explaining the kinds of information that dictionaries contain (much of which is often ignored by dictionary users), we focus especially on the evidence base of dictionaries, on issues around meaning and how it is encoded and understood, and on the technologies which are transforming dictionary-making and dictionary-publishing.

## 2 Week 1: Why Use Dictionaries?

We anticipate that the MOOC will attract a wide variety of participants from different educational backgrounds, and with different levels of linguistic and lexicographical expertise. The first Week of the course therefore starts by asking "Why use dictionaries when you can use search engines?", a question designed to provoke more conventional, and possibly older, dictionary-using participants, and also to acknowledge the reservations about dictionaries that other, possibly younger, participants might feel. Questionnaire findings suggest that learners do not always choose to use highly-regarded monolingual dictionaries, even if they consider these to be the best sources of information about words. We hope our 'big question' will indicate to participants that the MOOC is a non-judgmental environment, and will encourage honest reflection on possibly contradictory behavior and attitudes. The question may also expose differences of opinion regarding what actually constitutes a 'dictionary' and what constitutes a 'search engine'. The distinction might be puzzling to users if their search engine queries about word meaning or translation lead directly to online dictionary entries, without any need to specify a particular type of dictionary or to include mention of dictionaries in the query. Further confusion might be caused by the fact that, in some regions of the world, bilingual electronic dictionaries are not considered as 'dictionaries' at all, but are described simply as 'translators'.

The first Steps serve to introduce learners to the course educators and moderators and to each other. Following on from this, a series of Steps encourage consideration of possible reasons for dictionary use, as suggested in interviews with a representative selection of users, and in summaries of findings from various user surveys, from Barnhart (1962) and Quirk (1975), through to the recent large scale European survey on dictionary use. A central aim of the MOOC is to alter the participants' perceptions of the role of dictionaries in society. A poll will be taken in Week 1 to ascertain attitudes at the start of the course, and again in Week 6, so that participants can see if their own and others' attitudes have changed. The authority of dictionaries, differences between 'landmark' dictionaries such as the *Oxford English Dictionary* and less conventional products such the *Urban Dictionary*, and the rise of collaborative and crowdsourced dictionaries are all topics introduced in Week 1, to prepare the way for more in-depth investigations of dictionary content in later Weeks. We anticipate that as participants develop their ideas about the authority of different types of dictionary they will become more critically aware, and better able to distinguish between good and bad lexicographical practices.

## 3 Week 2: What's in a Dictionary Entry?

Week 2 examines the dictionary entry, and considers what kinds of information dictionaries need to provide for different types of word, and for different types of users engaging in different types of task. It begins provocatively, by suggesting that information encoded within an entry might be difficult to understand, and might often be ignored. Once again, this critical attitude towards well-respected

dictionaries is likely to surprise certain more traditional participants, but is also likely to reassure those who have themselves had difficulty interpreting dictionary entries; it should remind them that all kinds of opinion, including criticism of established practices, will add value to the discussion.

The tasks for Week 2 introduce participants to the basic components of the dictionary entry – the information that enables users to understand a word’s meaning and use it appropriately. Participants will look at standard defining practices and their alternatives, which might involve illustrations, sounds, and perhaps even smells!. They will also be asked to consider the challenges of creating entries for words whose usage is restricted in some way, for example in terms of register, region, prosody or phraseology. Having broadened their understanding of the range of information that might be included, participants will examine specific dictionary entries, and consider the appropriacy of the information provided, and its accessibility to the user.

Of course dictionary users’ information needs depend on the tasks they are engaged in at the time of dictionary consultation, so Week 2 includes consideration of the contexts of dictionary use, bearing in mind the portability of mobile phones and recent technological advances in speech production and voice recognition. Participants will be invited to share their experiences of dictionary use in both conventional and less-conventional settings (for example when speaking to shop assistants, or asking for directions on the street). A discussion of the value of different types of dictionary entry for different types of activity should also activate opinions concerning monolingual versus bilingual dictionaries, print versus e-dictionaries, and learner dictionaries versus those intended for fluent users of the language.

Thus, as well as introducing participants to some basic concepts, to be explored in greater depth in Week 5, the work in Week 2 helps to develop participants’ critical faculties, and may challenge some of their long-held views about the authority of dictionaries.

#### **4 Week 3: Evidence and Method: Where Does the Information in Dictionaries Come From?**

Week 3 is dedicated to illustrating the evidence and the methods used by lexicographers. This is intended as an introduction to the dictionary-making process for non-specialists, as we believe that learning about evidence sources for lexicography can raise the participants’ proficiency as dictionary users too. This Week builds on the previous ones, particularly the introduction to the content of dictionary entries in Week 2, and prepares the ground for the following weeks, particularly the dictionary inclusion criteria described in Week 4 and the definition-writing process explained in Week 5.

We encourage participants to reflect on the sources that could be relied on to create dictionary entries, before introducing them to an overview of evidence sources in lexicography. We then provide a historical overview of the use of evidence sources in lexicography, starting from pre-computational approaches based on paper slips, with reference to the *Oxford English Dictionary*. Following on from this we explain the “corpus revolution” and show how it brought fundamental changes to the dictionary-making process. Participants’ engagement is facilitated by a series of tasks aimed at raising their awareness of the role played by corpora in lexicography, and a deeper understanding of the nature of dictionary content and its relation to linguistic evidence. The tasks include a comparison between a dictionary entry and concordances of the headword in a corpus, with the aim of discovering gaps in the dictionary entry, as well as exercises about using corpus data to create a dictionary entry and identifying the corpus evidence relative to different components of dictionary entries. All this is aimed at triggering a critical discussion of the relationship between dictionaries and language data, and appreciating some of the subtle points regarding admissible and suitable evidence in lexicography. This



way, the course participants will reach a fuller understanding of the nature of dictionary content and its relation to language.

We also cover the various stages in corpus-based lexicography, from corpus design and text collection, to corpus annotation and ‘word sketches’, so that participants see how linguistic analysis levels, such as morphological or syntactic and so on, can be surfaced in the lexicographic process, become familiar with such a critical aspect of dictionary-making practice, and recognize the challenges of dictionary-making. We finish by presenting the role played by corpora and quantitative information in understanding language and therefore describing it in dictionaries, and a comparison of how corpus evidence is reflected in different dictionaries. This will relate to the participants’ personal experience with dictionaries and encourage them to be more aware users of dictionaries.

## 5 Week 4: What Goes into a Dictionary - Who Decides, and How?

Week 4 of the MOOC invites participants to explore what goes into a dictionary, who decides what is included, and how those decisions are made. This is done through a focus on the selection processes followed by dictionaries of various types (both ‘traditional’ dictionaries such as bilingual, monolingual and learner, and ‘less-conventional’ ones such as *Wiktionary* and the *Urban Dictionary*). Although inclusion criteria are more relaxed for less-conventional dictionaries than for their traditional counterparts, participants will be led to realize that attestation of real-world usage is crucial when differentiating genuine neologisms from ‘buzzwords’ (defined by Neuman, Nave and Dolev as ‘fashion words that enter the language and rapidly acquire great popularity’ (2010: 58, 67), yet regularly fade into obscurity).

We begin by asking learners to compare their estimates of the number of new words entering ‘traditional’ dictionaries each year with the figures given in the publicity material produced by renowned publishers of English dictionaries. From here participants are asked to consider why some words are accepted into dictionaries, while others are rejected. This will encourage discussion of the way some of the dictionary-making processes introduced in previous Weeks are put into practice, and will also introduce participants to some of the key issues facing lexicographers and collaborative dictionary contributors.

Participants are then asked to decide whether or not some newly-created words should be included in a dictionary of their choice, and to produce and discuss their own inclusion criteria, bringing their own practical knowledge and experience to bear. To support this discussion, we provide dictionary inclusion criteria from published sources, and input from professional lexicographers regarding the factors which actually influence the acceptance and rejection of new words or new meanings in a dictionary. The activity takes into account the different needs of different types of user: the *Oxford English Dictionary*, for example, focuses on those words deemed most likely to survive long term, rather than those which are most current (Algeo 1993: 283), while entries into *Wiktionary* tend to be more fluid. Of course, neologisms are not the only words added to dictionaries, so participants will also be introduced to other types of candidate, including borrowings, and words which have undergone a change of meaning or word class.

Finally, the Activities in Week 4 show participants how the lines between ‘traditional’ and ‘less-conventional’ dictionaries are becoming blurred, and how the use of crowdsourcing and user-generated content can add value to the former by incorporating elements of the latter. The concept of true ‘crowdsourcing’ and its application to dictionary-making is explored, along with the distinction between crowdsourced and strictly ‘collaborative’ dictionaries, and participants are asked to consider what collaborative and/or crowdsourced dictionaries are available in their own home countries, which ones they use, and why.

## 6 Week 5: Meanings and Definitions

Week 5 is perhaps the most intellectually-challenging part of the course. In this section, we tackle the subject of meaning: how meanings are created, why some words have more than one meaning, how lexicographers identify meanings, and how meanings are explained in a dictionary.

One of the big lessons of corpus linguistics is that the neatly-divided numbered “senses” in dictionaries imply a level of certainty around meaning which is not always supported by the evidence of language in use. So a major objective in this part of the course is to encourage participants to confront the reality that meanings are often less stable and less discrete than dictionaries suggest. As in much else of this MOOC, the message we want participants to absorb is that these things are not fixed “from above”, but often require difficult judgement calls on the part of lexicographers.

We start by asking the question “How do we know what words mean?”, and this is developed through a number of tasks, short articles, and videos. In one exercise, participants are presented with a series of (corpus-derived) sentences illustrating several uses of a polysemous word. Their task is to assign each sentence to a specific numbered sense in a dictionary entry, and it quickly becomes clear that such mappings are not always straightforward. In another Step, participants are introduced to the concept of polysemy, and they are asked to explain why this does not (as might be expected) lead to ambiguity and confusion among speakers and listeners (or writers and readers). We demonstrate that, in normal communicative interactions, fluent users of a language reliably identify the “right” intended meaning (when several possibilities exist), and we ask how they do this. Participants analyze corpus data to match a given instance with a specific meaning, and are then asked to find the contextual clues (syntactic, collocational, phraseological, and so on) which led them to associate each occurrence of a word with its dictionary meaning. This leads participants to the discovery that context almost always resolves any potential ambiguity. Following this, participants build up an inventory of criteria for distinguishing the various meanings of a complex word, and thus gain an understanding of how lexicographers approach the task of word sense disambiguation. This part of Week 5 ends with an article summarizing what has been learned about the relationship between meanings in the real world and “senses” in a dictionary.

In the second part of Week 5 we move on to the topic of definition, and think about how dictionaries explain meaning. Participants are first encouraged to give their own definitions of familiar objects and concepts, and then to share and discuss their output. This gives them an insight into the challenges involved, promotes an understanding that there may be several quite different ways of achieving the same result, and leads them to think about what definitions are actually for.

We then look at a range of dictionary definitions for the same “simple” word (such as the name of a familiar animal): why aren’t these definitions all the same, and what factors account for the differences among them? This leads on to a discussion of the needs, prior knowledge, language proficiency and so on of different types of dictionary user, and how this affects the structure and content of definitions. This point is consolidated through a task where participants evaluate different definitions for the same word, and give their views on which is best (and what “best” means in this context).

The issue of what makes a good definition is further developed through a number of Steps (articles, videos, exercises). Looking at a range of definitions in well-known dictionaries, we focus first on the informational content required for a definition which will resolve a user’s communicative problems: how much is enough, and is there such a thing as too much information? We introduce participants to the idea of the definition as a “typification”, rather than an attempt to account for every conceivable use of a word in text. This leads to the recognition that the word “definition” itself is problematic: the Latin root implies a level of certainty and precision which is unattainable for most items of everyday vocabulary.



Moving on to the language and structure of definitions, participants are introduced to a range of defining styles, from the most traditional (“the act of x-ing, etc.”), through folk-defining techniques, to more contemporary approaches such as full-sentence definitions. The relative merits of these different styles are evaluated. We also address the issue of dictionaries’ supposed objectivity – and its limits. The *Urban Dictionary* provides clear cases where a definition conveys opinions rather than facts, but although “serious” lexicographers strive to avoid subjective judgements when writing definitions, there are certain types of word where culture-specific norms are difficult to avoid – words such as *disabled*, *marriage*, *civilization*, or *god*.

In a final exercise, participants consolidate what they have learned by writing – and then discussing – their own definitions for selected words. A video in which the educators reflect on the main issues covered during the Week concludes this part of the course.

## 7 Week 6: What Does the Future Hold?

The final Week of the course is devoted to new research trends in computational lexicography (including automatic neologism detection and new sense detection) and dictionary use, as well as a summary of the course and a look at the future of dictionaries. The content is highly interactive, encouraging participants to reflect on their learning and how the course has changed their view of dictionaries. Participants will also be invited to discuss what the role of dictionaries in today’s and tomorrow’s world should and will be.

## 8 Conclusion

We have described what we believe to be the first ever MOOC devoted to dictionaries. We see this as a significant development, which will familiarize non-experts with the challenges faced by lexicographers and introduce them to the wide range of activity within our field – as well as, hopefully, dispelling some common misconceptions about dictionaries and how they are made.

The course will be offered free of charge, and is open to everyone, so some participants will probably choose to study it independently and on a voluntary basis. These might include language teachers and teacher trainees, and laypeople with an interest in language issues. However we believe that a significant number of participants will take the course as part of a credit-bearing program within their school or university, and we anticipate that some institutions will blend our online course content with input from their own teachers, to create modules tailored to the specific needs of their students. Clearly there is scope for learners to progress beyond the content of this MOOC to more advanced lexicography and dictionary user research. Once the MOOC is running, we will be able to monitor its effects on participants, and perhaps consider developing a second stage.

Although the course’s primary purpose is pedagogical, the inclusion of tasks requiring participants to upload information about their own dictionaries, and (later) their own dictionary-using habits, could provide a rich source of research data, available not only to us, the course developers, but also to any participant who enrolls. If the MOOC is embedded within a formal university program, for example, this data could be used in student assignments, mini research projects, or even dissertations or theses, and the Terms and Conditions set by FutureLearn allow the anonymous data collected from the course to be used in this way.

In addition, we hope to gain valuable insights from participants' feedback and discussions throughout the course. This could well have value in informing dictionary content, developing teaching materials in relation to dictionary use, and as a data source for researchers at all levels.

## References

- Algeo, J. (1993) 'Desuetude among New English Words'. *International Journal of Lexicography* 6(4), 281-293.
- Barnhart, C. (1962) Problems in Editing Commercial Monolingual Dictionaries. - In: F. Householder and S. Saporta (eds.) *Problems in Lexicography*. Publication twenty-one of the Indiana University Research Centre in Anthropology, Folklore and Linguistics, 161 - 181. Bloomington: Indiana University.
- Kosem, I., Lew, R., Müller-Spitzer, C., Wolfer, S. (2017). The European survey of dictionary use. In *Electronic lexicography in the 21st century: lexicography from scratch, the eLex 2017 conference*, Leiden, the Netherlands (book of abstracts). Leiden: Dutch Language Institute; Brno: Lexical Computing; Ljubljana: Trojina Institute for Applied Slovene Studies, page 53.
- Mugglestone, L.(2011) *Dictionaries: A Very Short Introduction*. Oxford: Oxford University Press
- Neuman, Y., Nave, O. and Dolev, E. (2010) 'Buzzwords on Their Way to a Tipping Point: a View from the Blogosphere'. *Complexity*, 16(4), 58-68
- Quirk, R. (1975): The social impact of dictionaries in the UK. In: R. McDavid and A. Duckert (eds.): *Lexicography in English*, 76 - 88. New York: Annals of the New York Academy of Sciences 211.
- Sharples, M. (2018) *The Pedagogy of FutureLearn: How our learners learn*. Available at <https://about.futurelearn.com/research-insights/pedagogy-futurelearn-learners-learn>
- Sharples, M. (2017) Pedagogy of FutureLearn: How our learners learn. Powerpoint presentation available at <https://www.slideshare.net/sharplem/pedagogy-of-futurelearn>
- Sharples, M., McAndrew, P., Weller, M., Ferguson, R., FitzGerald, E., Hirst, T., Mor, Y., Gaved, M. and Whitelock, D. (2012). Theory and practice of teaching, learning and assessment. Innovating Pedagogy 2012: Open University Innovation Report 1. Milton Keynes: The Open University.

## Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

# Researching Dictionary Needs of Language Users Through Social Media: A Semi-Automatic Approach

**Jaka Čibej<sup>1,3</sup>, Špela Arhar Holdt<sup>2,3</sup>**

<sup>1</sup>Jožef Stefan Institute, <sup>2</sup>Faculty of Computer and Information Science, University of Ljubljana

<sup>3</sup>Centre for Language Resources and Technologies, University of Ljubljana

E-mail: jaka.cibej@ijs.si, spela.arharholdt@fri.uni-lj.si

## Abstract

With the rise of digital media in the last decades, many language-related discussions have found home on various fora and social media such as Facebook, where users can participate in a shared-interest group to discuss language use, problems and resources. The posts in these groups are formulated by language users as a genuine response to a specific disruption in language use and offer an empirical starting point for studying language problems. We propose an automatic approach to extracting questions from language-related Facebook groups and describe the procedure in consecutive steps. We also address the issues of copyright, privacy and ethical constraints, and propose ways to overcome them. We present the extraction method on a case of two Slovene language-related Facebook groups: *Za vsaj približno pravilno rabo slovenščine* and *Društvo ljubiteljskih pravopisarjev in slovničarjev*. Both groups allow users to discuss language-related problems and find answers to their questions within the community. Our first extraction from these groups yielded approximately 1,900 posts (written by approximately 500 users) and 13,000 comments (posted by more than 900 users), providing ample material that can be analyzed to reveal the users' most frequent language problems.

**Keywords:** lexicographical user research, language problems, social media, automatic extraction, Facebook, Slovene

## 1 Introduction

As in many other fields, the development of the digital medium has brought an array of new possibilities to the field of dictionary user research. New procedures and methods used in this field, e.g. surveys, tests, evaluations, log file analyses (Welker 2013a, 2013b), have become less cumbersome or, in some cases, possible for the first time. This has enabled researchers to harness the change in interpersonal communication caused by the online environment. With the rise of digital media and computer-mediated communication in the last decades, many language-related discussions have found home on various fora and social media such as Facebook, where users can participate in a shared-interest group to discuss language use, problems and resources. The posts in these groups are formulated by language users as a genuine response to a specific disruption in language use. This data is especially valuable when taking into account the difference between what users believe their needs are (either in general, in relation to a specific language resource that is being evaluated, or when presented with hypothetical scenarios, which do not necessarily reflect their actual language dilemmas) and the actual language problems they encounter (i.e. what users really need when faced with an authentic language problem). From this perspective, observing user-reported language problems offers a more objective perspective on language problems compared to methods based on users reporting their problems post festum (e.g. interviews and questionnaires). Another aspect of this method that is of particular importance for user research is the broad scope of participants: while the population of Facebook cannot be considered as truly representative of all language users, the posts nevertheless

reveal the problems, needs and opinions of a large and diverse number of language users, regardless of which language resources – if any – they use.

As we point out in the following sections, the method of collecting, classifying and conducting both a quantitative and qualitative analysis of self-reported language problems has already been tested. However, manual data extraction remains time-consuming and less than trivial. In this paper, we further develop this method by presenting a number of (semi-)automatic improvements. We first present an overview of related work done in this field and continue with a step-by-step description of the method to automatically extract data from Facebook groups in order to obtain a large quantity of Facebook posts representing authentic language problems, which can be analysed in order to obtain an overview of the most typical user needs. The main purpose of this approach is to facilitate the acquisition of empirical data on language users' authentic communication dilemmas, which the dictionary as a tool (alongside other language resources) should be designed to resolve. While we focus predominantly on Slovene data, the methodology is language-independent, as similar language-related discussion groups can be found for Slovene (*Za vsaj približno pravilno rabo slovenščine* 'For an at Least Approximately Correct Use of Slovene', *Društvo ljubiteljskih pravopisarjev in slovničarjev* 'Association of Amateur Orthographers and Grammarians'), Swedish (*Sverige mot särskrivning*, Sweden against Writing Separately; *Sprakpolisar*, 'Language Police'), Danish (*Sprog for sjov – og i alvor*, 'Language for fun – and for real'), Italian (*Gli amanti della lingua italiana* 'Fans of the Italian Language'), and German (*Deutsch verbindet - Deutsch lernen* 'German Unites – Learning German'), to name just a few.

## 2 Related Work

In recent decades, lexicography has demonstrated an increasing interest in the needs, preferences and habits of dictionary users, with initiatives in dictionary-user research dating back as far as the 1960s (e.g. Barnhart 1962, Householder 1967, Tomaszczyk 1979) and gaining momentum in the 1980s (e.g. Hartman 1987, Wiegand 1987) and 1990s (e.g. Atkins 1998, Nesi 2000, Tono 2001). The emergence of the digital medium in the 2000s, however, allowed for new methodologies in dictionary-user research (Bergenholtz & Johnsen 2013, Müller-Spitzer 2014, Lew & De Schryver 2014). Different approaches – such as questionnaires, interviews, experiments, and research of actual dictionary use through think-aloud protocols, eye-tracking, log-file analysis, or user feedback collected through the dictionary interface – provide answers to which language resources dictionary users know and use, how they estimate their needs and habits in terms of the dictionaries they use, etc. This information is an invaluable foundation for dictionary development and has been increasingly frequently implemented in modern lexicographical projects.

However, existing approaches to dictionary users provide very little insight into why the user actually decided to consult the dictionary in question. Mentrup (1984: 160) proposed that the interest of the field '[. . .] should not start with the intangible dictionary usage situations but – as it were one level below – with language-related disruptions in language use situations'. This is later echoed by Tarp (2009), who suggests several possible approaches to address this gap, e.g. tests and interviews to investigate the readers' comprehension level and reception problems; analysis of text revisions; or simply the extension of existing methods (log files, eye-tracking, protocols) from dictionary use situations to extra-lexicographical situations, while already acknowledging that these approaches are mostly qualitative as well as time-consuming and expensive (ibid.: 293).

An alternative approach to identifying user needs, namely through user-generated content in digital media was proposed in Arhar Holdt et al. (2017) and Čibej et al. (2016). These two studies have

confirmed that language-related discussions in social media groups provide a wide range of implicit and explicit information that can be useful when designing user-friendly and user-oriented language resources. However, both of them were based on a limited number of (at the time most recent) posts that were extracted manually, forming a small sample that was not representative of the entire group. The evaluation of the method highlighted that the procedure would benefit greatly from automatization.

### 3 Automatic Extraction of User Posts

Automatic harvesting of information from social media is already commonplace in natural language processing (e.g. for sentiment analysis, opinion mining, and author profiling). We extend the use of this method to dictionary user studies by presenting an automatic approach to extracting questions from language-related Facebook groups through a Python script that makes use of the official Facebook Graph API. In this section and the following subsections, we describe the procedure in consecutive steps from identifying relevant Facebook groups and creating an app in the Facebook API, to extracting posts and comments. The Facebook Graph API allows for the extraction of all the posts (and comments posted as replies to those posts) from a Facebook group, along with a number of relevant metadata (e.g. user, time of publication, number of comments, likes and other reactions, links to resources and pages provided by the users) according to which a more representative and/or relevant sample can be made.

More detailed instructions on the use of the Python script are available on GitHub. In this paper, we only provide the basic steps.

#### 3.1 Facebook Graph API

The Facebook Graph API is the primary way for apps to read (and write) to the Facebook social graph. In order to use the Facebook Graph API, a Facebook account is required. After logging in, the user must create an app with which to access Facebook data. The app must then be reviewed and approved by the Facebook staff, which usually takes several days.

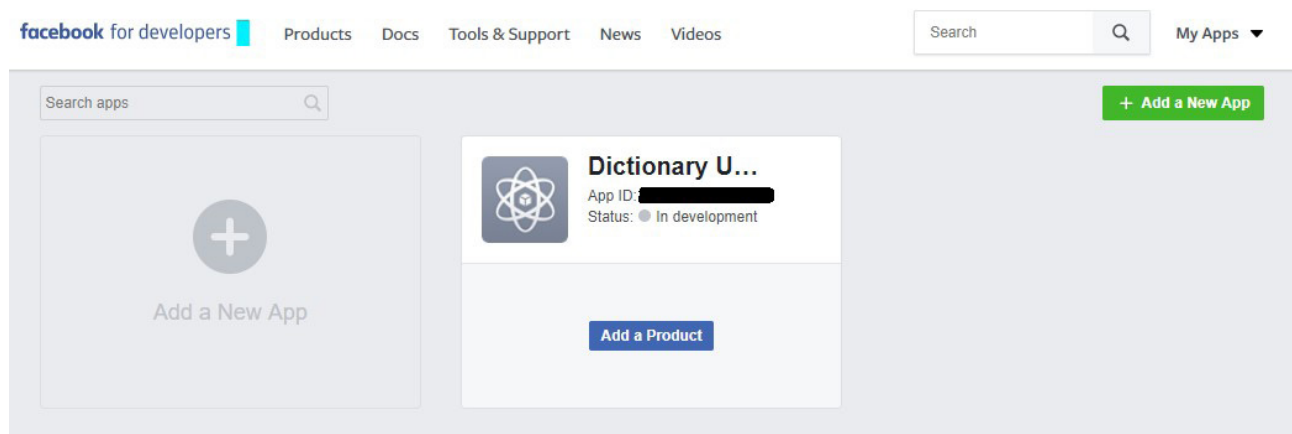


Figure 1: A screenshot of the Facebook Graph API interface. The censored part contains the app ID, which should be kept secret.

The user must then obtain the *app ID* (a unique, 16-digit number identifying the app), the *app token*, and the *access token* (an automatically generated sequence of alphanumeric characters that acts as



an access code; a short-term access token expires in an hour, while an extended access token last for several months), all of which must be incorporated into the script.

### 3.2 Facebook group IDs and Access Permissions

Extracting data from a Facebook group also requires the group's ID number. As of April 2018, the ID of a group can be obtained by inspecting the source code of the group page and finding the 'entity\_id' attribute (see Figure 2), which is a 15-digit number.

```
<script>require("TimeSlice").guard(function() {(require("ServerJSDefine")).handleDefines([]);new (require("ServerJS"))().handle({"require":
[["ScriptPath","set",[["\groups\profile.php:feed","dc67ef5a","imp_id":"ae7152d0","entity_id":"398216690214010"]]]]}), "ServerJS define",
{"root":true})();</script><title id="pageTitle">Za vsaj približno pravilno rabo slovenščine.</title><link rel="search"
type="application/opensearchdescription+xml" href="/osd.xml" title="Facebook" /><meta property="al:android:app_name" content="Facebook" /><meta
property="al:android:package" content="com.facebook.katana" /><meta property="al:android:url" content="fb://group/398216690214010" /><meta
property="al:ios:app_name" content="Facebook" /><meta property="al:ios:app_store_id" content="284882215" /><meta property="al:ios:url"
content="fb://group/?id=398216690214010" /><link rel="shortcut icon" href="https://static.xx.fbcdn.net/rsrc.php/yo/r/iRmz9lCMBD2.ico" />
```

Figure 2: A screenshot of the source code for the page of the public Slovene Facebook group *Za vsaj približno pravilno rabo slovenščine*. The group ID attribute is highlighted.

At this point, however, it should be noted that not all Facebook groups are equally accessible, as there are currently three types of groups: public groups (which new users can join freely once they have been confirmed by an administrator or another group member), closed groups (which are visible to the public, but can only be accessed by group members), and secret groups (which are invisible to the public and cannot be accessed by non-members). In terms of our post extraction method, there is an important difference between public and non-public groups. Data from public groups can be extracted by anyone with a Facebook account (and a Facebook Graph API app), regardless of whether they are a group member. With private (and secret) groups, however, this can only be done by the group administrator(s). The easiest way to bypass this restriction is to contact the group administrator(s) directly and explain the scope and goal of the research at hand. We discuss this in further detail in Section 5.

### 3.3 Data Extraction

The script we use<sup>1</sup> requires Python 3 and, in its current version (1.0), consists of two parts: the first part extracts the group posts, while the second extracts the comments beneath the posts. The output files can later be joined to form threads as seen in Facebook groups, or analysed separately – while user posts provide insight into the most frequent questions (or types of questions), user comments provide replies and, perhaps more importantly, the types of resources used to find solutions to the questions.

The script uses the Facebook Graph API syntax to request the following data: post ID, message (the text of the post or comment), username or pseudonym (e.g. User001), link (e.g. a link to a website if included in the post by the user), type of post (regular text-based post, link to video, etc.), time of publication, image (if included in the post by the user; the image is downloaded separately and is not included in the output file, but can be collocated with the correct post or comment through its ID), and finally, the number of comments, shares and different reactions (currently, Facebook allows users to mark posts with the following reactions: *like*, *love*, *wow*, *sad*, and *angry*). The data returned by the Graph API is first loaded in JSON format and then written to a CSV output file, which can be opened and analysed in most statistical analysis software (Excel, R, etc.). An example is shown in Figure 3.

<sup>1</sup> Our script is based on a script made by GitHub user Max Woolf (<https://github.com/minimaxir/facebook-page-post-scraper>). Our version is also available on GitHub: [https://github.com/jakacibej/dictionary\\_user\\_needs](https://github.com/jakacibej/dictionary_user_needs)



status_id	status_message	status_author	link_name	status_type	status_link	status_published	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
398216690214010_1310622185640118	Živjo, živim v Sloveniji nekaj časa in mislim, da knjižni jezik je dost drugačni od tega, kar ljudje vsakodnevno govorijo - slišim samo "kr neki" in "dej no"... tuki je en video, kako mi, tuji v Sloveniji, vidimo vaš jezik :) Upam, da Vam bo všeč :)	User1	20 phrases Slovenes use the most	video	https://www.youtube.com/watch?v=IDJrUms9B	12.04.2017 19:39	25	3	2	23	1	0	1	0	0
398216690214010_1306040452764958	Danes v oddaji za otroke Ringa raja na1.programu radia : pri kolesu moraš pregledati BREMZE. Lepo, da že male otroke učimo lepe slovenščine.	User2		status		8.04.2017 08:31	2	3	0	1	0	0	1	0	0
398216690214010_1304850312883972	Pozdravljeni. Zanima me, kako je s presledki pri računalniškem navajanju datumov, ali je presledek med pikami ali ne? Hvala.	User3		status		7.04.2017 10:49	0	2	0	0	0	0	0	0	0
398216690214010_1301547756547561	Me lahko kdo razsvetli glede rabe "občnega zbora" pri raznih društvih in organizacijah - je ta izraz pravilen ali ne? Ali bi moralo biti "obči zbor"?	User4		status		4.04.2017 09:04	0	3	0	0	0	0	0	0	0
398216690214010_1252373568131647	Z nami v oddaji po desetih? Pravilno ali ne? Meni se vsekakor bolje sliši po deseti (uri), ampak sem prvo različico že večkrat slišala na radiu (val202).	User5		status		7.02.2017 08:42	1	4	0	1	0	0	0	0	0
398216690214010_1291992627503074	Eno hitro pomoč bi potreboval Štajerc ali Štajerec? namešč word prvo besedo podčrta kot nepravilno, drugo pa ne kar pomeni da naj bi bila pravilna. Sam menim osebno je prva boljša.	User6		status		24.03.2017 20:58	1	4	0	1	0	0	0	0	0
398216690214010_1291906407511696	Na Valu202 so se pa danes GUŽVALI.	User3		status		24.03.2017 19:00	0	2	0	0	0	0	0	0	0

Figure 3: Example of an output CSV file imported into Excel.

By default, the script anonymizes all the usernames, replacing them with generic codes (*User1*, *User2*, etc.), which remove problems with privacy protection while still enabling posts to be grouped by user. The script does allow the automatic anonymization to be turned off, but in this case, researchers treat the extracted data as carefully as possible and take every precaution to protect user privacy (for instance, unanonymized data is not suitable for publication in publicly available corpora).

## 4 The Case of Slovene Language-Related Facebook Groups

We present the results of our automatic extraction method on a case of two Slovene language-related Facebook groups: *Za vsaj približno pravilno rabo slovenščine* (For an at Least Approximately Correct Use of Slovene) and *Društvo ljubiteljskih pravopisarjev in slovničarjev* (The Association of Amateur Orthographers and Grammarians). Both groups allow users to discuss language-related problems and find answers to their questions within the community.

### 4.1 Quantitative Overview

As of April 2018, the groups consist of more than 2,500 and 1,800 members, respectively, and have been active since 2011 and 2012, respectively. Our first extraction (see Table 1) from these groups yielded approximately 1,700 posts (written by approximately 500 users, some of which are members of both groups) and 13,000 comments (posted by more than 900 users). The data is shown in Table 1 below. As can be seen, the method provides ample material that can be analyzed to reveal the users' most frequent language problems and identify the areas in which existing language resources could be improved in order to better fulfil the needs of language users. We describe this in more detail in the following subsection (4.2).

Table 1: Number of users, posts and comments extracted from the groups.

Group	Users	Posts	Comments
Za vsaj približno pravilno rabo slovenščine	562	604	4.315
Društvo ljubiteljskih pravopisarjev in slovničarjev	273	1.135	8.548

An overview of the number of posts and comments per user shows that while the majority of users (approximately 90 %) posted only a handful of posts and comments (between 1 and 10), there are

nevertheless several very productive users (with up to 105 posts and 822 comments). On average, users posted approximately 9 posts and 15 comments.

The fact that users post unevenly was one of the problems encountered with manual extraction of Facebook group posts. The sample collected in this way was very prone to skewing, as there is a higher chance to include only very active users while neglecting the ones that may have posted only a handful of questions, especially if they have not been very productive at the time of data collection. Automatic extraction allows for stratified sampling by user to ensure that all users that posted in the group are included in the sample.

## 4.2 Qualitative Overview

The posts are a valuable source of information to be implemented in the design of digital lexicographic resources, as already confirmed by the results of Arhar Holdt et al. (2017) and Čibej et al. (2016): according to their typology, the questions found in the posts can be divided into 17 categories, which cover diverse scenarios such as *Which of these options is better?*, *Is this word correct or not?*, *What does this word mean?*, and so on. The examples below are English translations<sup>2</sup> of posts extracted from the Facebook group *Za vsaj približno pravilno rabo slovenščine*. The questions cover a variety of different topics, including orthography (examples 4 and 5) and variation (examples 1 and 6), semantics, word form (example 2), word origin, translation (example 3), and metalinguistic or other external data.

- (1) Hello. One question – šola astme (school of asthma), šola astma ali astma šola? And why. (I'm for 'šola astme' analogous to the expressions 'šola hujšanja' (school of weight loss), 'šola zdravega načina življenja' (school of healthy lifestyle), but I have no other arguments for it). Thanks. And have a nice Wednesday.
- (2) How do we call the inhabitants of Sicily? (And I don't mean Italians ;))
- (3) Does anyone know how to translate "zero anaphora"?
- (4) UV light or UV-light?
- (5) hi, I'd like to know how to correctly write the expression ad-hoc/ad hoc – in italics? (when speaking of an ad-hoc decision, an ad-hoc work group). thank you for the help&advice.
- (6) when speaking of the Jedi from Star Wars: "jedijski" or "jedijski" – the results in Gigafida are approximately equally frequent for both, with slightly greater frequency for the second. What do you think? Thanks for your replies.

The analysis and a thorough overview of the most common categories of user problems can provide a lexicographical project with several guidelines on how to prioritize dictionary content, how to structure the dictionary interface and what functionality it should offer. As pointed out by Arhar Holdt et al. (2017), for many of the needs revealed by the material extracted from language-related Facebook groups, a number of solutions are already available, for example query lemmatisation, the did-you-mean function, pronunciation sound clips, and interconnectivity with other resources (these are also mentioned in Lew and De Schryver (2014)). However, the analysis shows some user needs

<sup>2</sup> Slovene originals:

- (1) Dan. Eno vprašanje - šola astme, šola astma ali astma šola? In zakaj. (Jaz zagovarjam 'šola astme' po vzoru šola hujšanja, šola zdravega načina življenja, drugega argumenta pa nimam). Hvala. In lep preostanek srede.
- (2) Kako rečemo prebivalcem Sicilije? (In ne mislim Italijani ;))
- (3) Morda kdo ve, kako se prevede "zero anaphora"?
- (4) UV svetloba ali UV-svetloba?
- (5) živijo. zanima me, kako pravilno zapišemo izraz ad-hoc/ad hoc - v italic? (ko govorimo o ad-hoc odločitvi, ad-hoc delovni skupini). hvala za pomoč&nasvet.
- (6) v zvezi z jediji iz Vojn zvezd: "jedijski" ali "jedijski"- Gigafida daje oboje v približno enakem številu, rahla prednost drugega. Kaj mislite? Hvala za odzive

that are not as frequently discussed, even though they could probably be met with relatively simple steps. For example, users often wish to compare two or more language variants. The comparison of two (semantically similar) words was also one of the most typical and frequent scenarios identified by Čibej et al. (2016), who analysed the posts in the Slovene Facebook group *Prevajalci, na pomoč!* (Translators, help!) and demonstrated that a great number of users would benefit from a Slovene synonym dictionary, a lacuna that has since been filled by the Thesaurus of Modern Slovene (Krek et al. 2017). The Thesaurus of Modern Slovene was designed as a direct response to the identified user needs, and among other functions, it offers the possibility of comparing two synonyms in context by providing their most typical collocates (see Figure 4, showing the most typical collocates for *razvoj* ‘development’ and *napredek* ‘progress’) and examples of use. This is thus a good-practice example of how the analysis of language-related user-generated content can directly contribute to user-friendly dictionary design.

The screenshot shows the website 'cvt sopomenke .io' with a search bar containing 'razvoj'. The main content area displays a comparison of collocates for 'razvoj' and 'napredek'. On the left, there is a sidebar with filters: 'Relevantnost' (dropdown), 'Pogostost' (slider), and 'usmeritev' (dropdown). Below these, a list of categories is shown: 'napredek' (selected), 'usmeritev', 'potek', and 'sprememba'. The main area shows a grid of collocates for 'razvoj | napredek'.

razvoj   napredek			
tehnološki	v Sloveniji	prispevati k	spremljati
hiter	na področju	pripomoči k	doseči
gospodarski	v letu	skrbeti za	omogočiti
nadaljnji	v primerjavi	vplivati na	omogočati
trajnosten	v smeri	vlagati v	spodbujati

Figure 4: Collocations page of the Thesaurus of Modern Slovene, allowing a comparison between two synonyms.

## 5 Personal Data Protection and Ethical Restrictions

When dealing with Facebook data, a number of legal and ethical restrictions need to be taken into account. In this section, we describe these issues and propose solutions to overcome them.

The first issue concerns personal data protection, as data obtained from Facebook most often contains personal information. In our case, the most problematic are the users' usernames, which usually consist of their real-life names and surnames. It is crucial to take every precaution to ensure that the users' rights to privacy are not violated. Our script automatically anonymizes all usernames, but also allows this option to be turned off (if names, and e.g. gender, which can be deduced from them, are important to the goals of the research at hand). In this case, the researcher(s) dealing with the data should ensure that all personal data is used only for research purposes and never shared outside the research group unless informed consent has been acquired from the group members and the material properly anonymized.

The second issue concerns ethical restrictions. In the case of public groups, the data and posts were publicly accessible and, until the most recent version of the Facebook Graph API (v2.12, April 2018),

could be harvested even without explicit permissions from group members and/or administrators. Access has been restricted since then. In any case, it is advisable to establish contact with the group and explain the nature of the project, especially if the result (e.g. a language resource) will benefit the community. In the case of non-public groups, data has never been publicly accessible without the permission of the group administrator(s), so adequate contact with the community is obligatory. There are two ways a group administrator can grant access to the Facebook group data. The first way is by creating their own Facebook Graph API app and providing the researcher with a (temporary) access token that will allow the script to download group posts and comments. However, the access token either expires within an hour (which is usually too short a time to finish downloading all the data from the group) or within several months (which may raise suspicion among group members). In addition, this solution requires a lot of unnecessary work on the group administrator's part. The second way is to ask the group administrator(s) to accept the researcher's request to join the group and then temporarily (e.g. for a day or another fixed amount of time) promote them to group administrator. With administrator permissions, the researcher can then access the group's data through their own Facebook Graph API app. However, understandably, administrators will be reluctant to accept responsibility of allowing the data of the entire group to be accessed by a third party, which is why it is advisable to inform the community of the research taking place, the exact type and format of the data that will be collected, the purposes for which it will be used, and lastly, that data extraction will only take place at a pre-determined time, and anonymized. While it is often impossible or at least impractical to obtain consent from every single group user, a poll can be held within the group to vote on whether they are willing to allow access or not. The administrators can then determine a threshold, e.g. if more than 60 % of the votes are in favor of the data extraction, the researcher shall be granted access. It is also advisable for the researcher to draft an official statement signed by themselves and their institution, stating the conditions under which the data can be harvested (e.g. used only for scientific purposes).

Contacting a group is also important for dissemination purposes and community building. The researcher should keep in touch with the community even after data extraction to inform them about the progress of the project and perhaps post some interesting findings to allow the community to provide feedback. It is important not to treat the group simply as a source of information, but as a community that can contribute to dictionary design in a number of different stages of development.

## 6 Conclusion

In the paper, we have presented a method to automatically obtain large quantities of authentic language-related user questions (as well as their solutions) from Facebook groups on social media. The script used to extract posts and comments from Facebook groups is language-independent and is openly accessible on GitHub for the benefit of the research community. The extracted posts include invaluable implicit and explicit information that can be analysed in order to form guidelines for a more user-friendly and user-oriented approach to the design and compilation of new language resources. It is also worth noting that the method produces posts that include a number of relevant metadata that can be processed during the analysis to find or filter the most relevant posts, e.g. with the most comments or the most reactions. In addition, the method enables the creation of a sample that is more representative of the entire group, as it allows for stratified sampling by user.

However, the method does have several potential weak spots that need to be addressed. First, in light of recent controversial events with social media and discussions on data privacy, any restrictions to Facebook API policy, although unlikely to completely ban all automatic extraction, may prove problematic. Second, so far, it is impossible to extract metadata on the users themselves (e.g. their education level, age, etc.). While this method does sample a larger number of users compared to individual



interviews, the results should not be interpreted as representative of the entire population of language users. In certain situations, groups include people with a shared professional background (e.g. *Prevajalci, na pomoč!* for translators), which allows the researcher to more accurately deduce the type of users being researched. In other situations, it might be prudent to conduct a poll within the group to determine, at least approximately, the type(s) of users being researched.

There are several other possibilities and improvements to the method to be explored as future work. First, within dictionary-compilation projects, it is possible to encourage the growth of a separate community to collect feedback on various versions of the project and, in later stages, to evaluate the interface. This method is being pioneered by the Thesaurus of Modern Slovene, which has a dedicated Facebook group aimed specifically at collecting user feedback on the Thesaurus. Feedback can then be automatically extracted and sorted by metadata.

Second, we intend to extract posts from all relevant Facebook groups for Slovene and conduct another analysis along the lines of Arhar Holdt et al. (2017), with particular emphasis on improving their bottom-up typology of user-generated language-related questions. Their analysis has namely shown that the method would benefit from a multi-layer categorization, with more robust categories for each layer if possible. These would be more adequate for further automatic processing. We namely also plan to implement machine learning to check whether user posts can be classified automatically, e.g. by language of interest (Slovene, English), by linguistic field (semantics, orthography, morphology, lexis), by potentially helpful resources (thesaurus, monolingual dictionary, bilingual dictionary), and so on. This will further automatize the entire process of analysing user needs through social media, and, if successful, provide an instant general overview of the most frequent user needs.

## References

- Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (2017). Value of language-related questions and comments in digital media for lexicographical user research. *International journal of lexicography*, 30 (3), pp. 285-308.
- Atkins, B. T. S. (ed). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.
- Barnhart, C. L. (1962). Problems in Editing Commercial Monolingual Dictionaries. *International Journal of American Linguistics*, 28(2), pp. 161-181.
- Bergenholtz, H. & Johnsen, M. (2013). User Research in the Field of Electronic Dictionaries: Methods, First Results, Proposals. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 556-568.
- Čibej, J., Gorjanc, V. & Popič, D. (2016). XVII EURALEX International Congress, 6-10 September, 2016, Tbilisi. Analysing translators' language problems (and solutions) through user-generated content. In T. Margalitadze & G. Meladze (eds) *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 158-167.
- Householder, F. W. (1967). Summary Report. In F. W. Householder & S. Saporta (eds) *Problems in lexicography*. Bloomington: Indiana University Publications, pp. 279-282.
- Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch*, 19-21 September 2017. Leiden, Netherlands, pp. 93-109.
- Lew, R. & De Schryver, G. M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography*, 27(4), pp. 341-359.
- Mentrup, W. (1984). 'Wörterbuchbenutzungssituationen-Sprachbenutzungssituationen. Anmerkungen Zur Verwendung Einiger Termini Bei HE Wiegand.' In W. Besch, K. Hufeland, V. Schupp & P. Wiehl (eds) *Festschrift für Siegfried Grosse zum 60. Geburtstag*. Göttingen: Kümmerle Verlag, pp. 143-173.
- Müller-Spitzer, C. (ed). (2014). *Using Online Dictionaries*. Berlin, Boston: De Gruyter Mouton.

- Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries*. Tübingen: Max Niemeyer Verlag.
- Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexikos*, 19(1), pp. 275–296.
- Tomaszczyk, J. (1979). Dictionaries: Users and Uses. *Glottodidactica* 12, pp. 103–119.
- Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Berlin: Walter de Gruyter.
- Welker, H. A. (2013a). Methods in Research of Dictionary Use. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 540–547.
- Welker, H. A. (2013b). Empirical Research into Dictionary Use since 1990. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 531–540.

## Acknowledgements

The research presented in this paper has received funding from the Ministry of Culture of the Republic of Slovenia (Thesaurus of Modern Slovene: By the Community for the Community), the infrastructural program of the Centre for Language Resources and Technologies (funded by ARRS, the Slovenian Research Agency) and the ARRS-funded research program P6-0215 Slovene language – basic, contrastive, and applied studies.



# The DHmine Dictionary Work-flow: Creating a Knowledge-based Author's Dictionary

**Tamás Mészáros<sup>1</sup>, Margit Kiss<sup>2</sup>**

<sup>1</sup>Budapest University of Technology and Economics, <sup>2</sup>Institute for Literary Studies, Hungarian Academy of Sciences

E-mail: meszaros@mit.bme.hu, kiss.margit@btk.mta.hu

## Abstract

Digitalized author's dictionaries could play an important role in humanities research. Not only could they provide better ways to study an individual author's vocabulary, but they could also act as a knowledge source for other computer-based methods. We present the process of making an author's dictionary of headwords, writing variations, word forms and corpus citations extended with part-of-speech, linguistic, literary and semantic information. We also describe how this extended dictionary incorporates knowledge from linked open data sources and from critical annotations and builds an RDF knowledge base attached to the dictionary. The result is a vast knowledge source about an author's oeuvre that can be studied and used to enhance corpus analysis. We demonstrate our method on processing a large text corpora of 1.5 million words from the 18th century and on creating the digital author's dictionary of Kelemen Mikes.

**Keywords:** author's dictionaries, knowledge-based systems, corpus analysis, linked open data

## 1 Aim of the Research

The ongoing DHmine project at the Budapest University of Technology and Economics aims to create a software tools to support various digital humanities (DH) research tasks (Mészáros 2016). In cooperation with the Institute for Literary Studies of Hungarian Academy of Sciences, we processed the works of Kelemen Mikes, an 18th-century author often called the “Hungarian Goethe” (Kiss 2012).

Our main goal was to create an author's dictionary of Kelemen Mikes. This was a groundbreaking work since this era is rather underrepresented in computerized corpus building, and no complete digital author's dictionary had been created in Hungarian language before. Our aim was thus to establish a work-flow for creating such dictionaries and also to demonstrate the possible benefits of information technology in this field.

We concentrated our work on two main aspects: increasing the efficiency of the dictionary-making process by utilizing various software tools, and taking a step beyond data-centric digitalization and introducing knowledge-based methods in creating and using the dictionary. Our research aim was to develop methods for incorporating various kinds of knowledge in digital author's dictionaries and then utilize them in corpus analysis.

## 2 Previous Work

Computerized tools play an increasingly important role in humanities research. They provide efficient tools for storing, searching, retrieving and displaying digitalized texts, they also vastly improve the

efficiency of various research tasks, including making concordance lists and performing many different kinds analyses on the collected data.

Creating computerized concordance lists has been a focus of humanities researchers for many decades now. Roberto Busa (Busa 1980) was a pioneer in this field, creating a list of concordance called *Index Thomisticus* in 1951. Since the 1950s research and development in computerized lexicography has yielded significant results by processing the works of Kant, Shakespeare, Goethe, Dante and many other authors. In the Hungarian language the work of Ferenc Papp in the 1970s is also notable. He was processing Ady's oeuvre and emphasized the importance of computer-created concordance as a raw material of author's dictionaries (Mártonfi 2014).

Although a concordance list itself is a valuable data source for lexicography research, it can be augmented with many different kinds of information to form a dictionary of an author's oeuvre. Completing it with grammatical, semantical, stylistic, historical or cultural information yields a vast knowledge store for research (see Mattaush 1990: 1552-1553; Mártonfi 2014). The resulting author's dictionary is "a type of reference work which provides information on the vocabulary of a specific author." (Hartmann & James 2001: 10). Many author's dictionaries have been created in various languages (Goethe, Schiller, Thomas Mann, Bertolt Brecht, Dante, Ibsen, Shakespeare, etc.). In the Hungarian language the Petőfi dictionary is a notable example (J. Soltész et al. 1973-1987).

Paper-based dictionaries often face problems due to space limitations and the substantial manpower required to complete them. These dictionaries were typically created slowly, with a detailed entry structure and a sophisticated meaning description, stylistic qualifications, phraseological references, and so on, and computerized tools can help to overcome the related time and space limitations. They provide a virtually endless storage capacity and also speed up certain work phases of the dictionary-making process. Our aim was thus to provide such tools for processing Mikes' oeuvre, for the dictionary making process and for storing and accessing the related corpus and vocabulary.

Recently some pioneering research work has applied artificial intelligence (AI) techniques in the field of digital humanities. AI is a vast field of research that tries to grasp human knowledge and mimic our behavior in problem solving. Knowledge-based systems are especially successful in representing and using human knowledge in computerized systems. As digital humanities projects have already accumulated a lot of data in digital form, it is a natural step forward to transform these data into knowledge and utilize them to support human researchers. This can be done by, for example, connecting contextual knowledge to a corpus (Bartalesi et al. 2015), semantic corpus annotations, knowledge extraction from text using NLP techniques, or adding sentiment information to lexicons (Nugues et al. 2016). Our research focused on how knowledge can be incorporated, represented and used in digital authors' dictionaries.

### 3 The Dictionary-making Process

We present the process of making the digital Mikes dictionary based on the works Kelemen Mikes, a famous Hungarian author from the 18<sup>th</sup> century.

#### 3.1 The Mikes Oeuvre

Our basis for compiling the dictionary was the critical edition of Mikes' work created by Lajos Hopp (Hopp 1966-1988). This contains Mikes' letters written from exile on almost 6,000 pages, and it also contains the critical annotations and research notes of Lajos Hopp. The initial electronic form of the corpus was created in cooperation with the National Széchényi Library. They performed the OCR

process on the scanned documents. After manually correcting scanning and recognition errors we created an electronic version of the original text and the critical annotations. This was the initial corpus for our dictionary-making process.

### 3.2 Creating the List of Concordance with Full-text Citations

The process started with creating a full concordance list of words with attached full-text citations. This automatically generated concordance list was not only linked to the appropriate corpus locations, but entries were also extended with full-text citations. An entry in this list contains a word form (e.g. “*Constáncinapolyban*”) and all its occurrences in the corpus with full-text quotes. It is important to note that digital dictionaries do not have the space limitations that their paper-based versions exist with. Thus we generated all citations in their full-length form in the concordance list, helping researchers to analyze them and create the dictionary entries later on.

We chose the XML standard for storing the concordance list and also for authoring the dictionary. This is a common choice for DH researchers, as it is very flexible in storing various kinds of data in a self-describing format, and it is also easily processable by computerized tools. We used a subset of the XML TEI standard to encode the content: tags were marking headwords, writing variations, word forms, corpus citations, source references, and so on. We also developed a simplified syntax to support researchers in using their favorite non-XML text editor during the authoring process of the dictionary.

The following two examples show excerpts with citation data for the word form “*Constáncinapolyban*”.

Simplified XML syntax showing the original sentence, the word form marked and the location in the corpus (TL 250):

- (1) ö parancsollya. innét csak hamar <I>Constáncinapolyban</I> megyünk, azért hogy meg lássuk (TL 250)

The above sample encoded in XML TEI after an automated transformation is as follows:

- (2) <cit type=“example”><quote>ö parancsollya. innét csak hamar <I>Constáncinapolyban</I> megyünk, azért hogy meg lássuk </quote><bibl>TL 250</bibl></cit>

The result of the automated concordance-making process was a huge text document (roughly 60,000 pages, 150MB). This full concordance list contained roughly 260,000 different word forms with 1.6 million citations. This formed the base data set of making the dictionary.

### 3.3 Creating Headwords

This huge concordance list was then processed manually by researchers to identify the list of headwords and to attach the word forms to them. This time-consuming process required many man-months of expert work. During this, researchers corrected the errors of the automatically recognized word forms, identified the proper headwords and attached the word forms and citations to them. They also identified the headword’s modern form and common writing variations and extended the dictionary entries with this information.

For the above example the headword “*Constancinapoly*” was created. In simplified XML form (excerpt):

- (3) <U>Konstantinápoly</U>
- (4) <Q>Constáncinapoly</Q>
- (5) <B>Constáncinapolyban</B> – 1
- (6) ö parancsollya. innét csak hamar <I>Constáncinapolyban</I> megyünk, azért hogy meg lássuk (TL 250)

For this headword researchers identified 10 writing variations (Q tag), 22 word forms (B) in the corpus, they attached 99 citations (I) and also noted its “*Konstantinápoly*” (Constantinople) modern form in Hungarian (U tag).

After processing all word forms and creating their respective headwords, the core of the digital Mikes dictionary was created (Kiss 2012).

### 3.4 Extending the Core Mikes Dictionary

Our main goal was to build an extended dictionary that acts as a knowledge source for literary research and also supports corpus analysis. In order to achieve this we extended the dictionary entries in various ways.

#### 3.4.1 Incorporating Additional Lexicographic Knowledge

We analyzed dictionary data (word forms and citations) and extended them with lexical and other attributes. For example, many Turkish words were used in the texts written by Mikes (the author spent many years in exile in Turkey), and he also created his own words that could not be found in other dictionaries. These were marked in the dictionary.

A rather complex task was to extend the citations with part-of-speech (POS) information. In order to perform this we developed a tool to automatically identify the POS roles of word forms based on the common rules of the Hungarian language grammar of that time period. After automatically extending the headwords with this information the researchers corrected these manually by reviewing all word form uses in the attached full-text citations. When a word form had multiple POS roles in the citations they were also marked manually.

#### 3.4.2 Adding Semantic Knowledge to Headwords

The other type of knowledge we introduced in the dictionary was semantic information about headwords.

Word forms in Hungarian have changed significantly since the 18<sup>th</sup> century. Adding the modern word form to headwords already helps greatly in accessing more information about them.

We also marked named entities in the dictionary. Headwords of place and person names were enriched with semantic markings. In order to speed up this process we developed automated entity recognizers to detect proper names of persons and geographic locations in the corpus (Mészáros 2016). The proposed entities were then reviewed by human experts and this information was also added to the appropriate headwords. For example, the “*Konstantinápoly*” headword was marked as a place name.

It is important to note that although these extensions describe knowledge about headwords they do not alter the structure nor the usage of the dictionary, they merely store additional information about headwords.

### 3.5 Linking the Dictionary with Knowledge Sources

Our aim was to create a knowledge base that grows beyond the structure (and capabilities) of a traditional author’s dictionary. In order to achieve this we took a step forward in extending the dictionary to better utilize knowledge-based methods and tools. To implement this change we decided to link the dictionary with additional data sources and store this information as a knowledge base attached to the dictionary.

The dictionary is already linked to the corpus in many ways through word forms and sentences, but there are already other data sources available that can be referred to. Firstly, the set of critical annotations is a natural place to investigate. Secondly, the semantically enriched headwords can be linked to knowledge sources related to these concepts. For example, the recognized named entities (e.g. person and place names) can be connected to Linked Open Data (LOD) like DBpedia (Auer S. et al. 2007) sources that contain more information about them. We explored these two possibilities during our research and present the results in the following sections.

### 3.5.1 Attaching Research Notes

The first candidate to attach more knowledge to the dictionary was the set of critical annotations created by Lajos Hopp (Hopp 1966-1988). Critical annotations are a primary knowledge source when researching an oeuvre. They are traditionally text fragments attached to various parts of the corpus. In contrast to the author's texts we can observe that they usually have a more or less well-formed structure and they can be categorized based on their primary purpose: linguistic, historic, external reference, etc.

Lajos Hopp wrote more than 5,000 research notes about various parts of the Mikes oeuvre. He created his annotations in a more or less standardized form: a citation and a reference to the corpus, his researcher note, and references to external documents. The following example shows a note about the usage of “*Constantinapolyban*” (excerpt):

(7) [1.]

(8) 0 Constantinapolyban — Előfordul még Constantinápoly, Constancinapoly, Constancinápoly, Constáncinápoly [...]

We can observe the corpus reference in lines 7 and 8 of the above example. Line 7 selects the first letter of Mikes, and the very beginning of line 8 specifies that the word can be found at the beginning of that Mikes' letter. Line 8 in the example also shows the original text “*Constantinapolyban*” and the attached note (after a “—” sign) that lists other word forms and describes historical notes and Mikes' personal attachment to the city.

By observing this structure we were able to create a software tool that automatically transforms these annotations into a structured XML TEI form including citations, references and the annotations themselves. The following example shows the result of this automatic transformation.

(9) <note type=“critical annotation”>

(10) <text>Előfordul még Constantinápoly, Constancinapoly, Constancinápoly, Constáncinápoly [...]</text>

(11) <cit><quote>Constantinapolyban</quote><bibl unit=“line” from=“0”>TL.1</bibl></cit>

(12) </note>

We also analyzed what kinds of annotations were created by Lajos Hopp. We categorized them into 10 subtypes (like historical, social, geographical, grammatical, and so on), and then manually labeled the annotations based on their categories to determine what type of knowledge is stored in them. For example, the above note was categorized as geographical (G) and historical (H). To store these categories line 9 in the previous example was changed to

(13) <note type=“critical annotation” subtype=“G,H”> ...

From the XML version of the critical annotations we created a database and automatically attached its entries to the appropriate parts of the dictionary and to the corpus by identifying word forms in the citations and by extracting references from the XML tags and attributes. These annotations and their links to other data elements were the first type of knowledge attached to the dictionary.



### 3.5.2 External Knowledge Bases

In order to incorporate more knowledge about headwords we explored external knowledge sources. DBpedia (Auer et al. 2007) is a well-known LOD data source that contains a vast amount of knowledge. For example, it contains many kinds of information related to the headword “*Constantinople*”: it is a general concept that is a subject of many other entities, it is one of the names of Istanbul, a geographic location, it was also the capital city of many empires, etc.

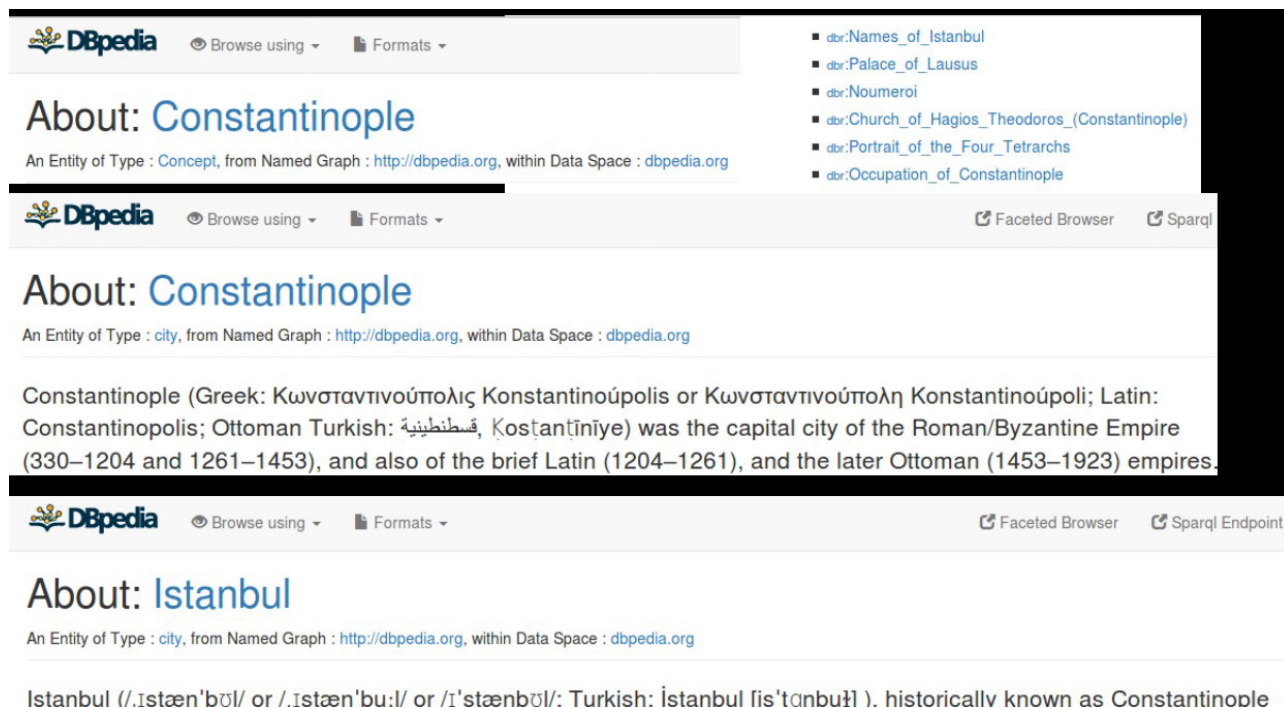


Figure 1: Excerpts of DBpedia entries related to Constantinople

Obviously, it is not suitable to incorporate all this knowledge in the dictionary. Since we focused on named entities we decided to narrow queries to these. Even this dataset was too large to include, so further selection had to be made based on the possible applications of this knowledge. We followed a restricted set of DBpedia links like geographical location, person, temporal and name alternatives to select the proper subset of information.

We developed an automated software tool to connect dictionary entries to these external sources and retrieve a restricted set of data from them. This tool builds a small-scale local mirror of selected knowledge pieces retrieved from the sources. As named entities were already marked in the dictionary, and their modern word forms were also included, it was fairly easy to perform DBpedia lookups to find the appropriate external records. The list of entries proposed by this tool was also manually reviewed by researchers before starting the knowledge transfer in order to solve disambiguation problems. The selected entries were transferred to a local graph database. The details of this will be discussed in the next chapter.

## 4 Implementing the Extended Author's Dictionary

The main goal of the dictionary is to support literary research by storing data and providing various functions to retrieve the stored information. We explored several use cases regarding authoring,



storing and accessing the information in order to develop appropriate data models, storage methods and system functions, and to select the appropriate software tools to implement them.

## 4.1 Data and Knowledge Storage

The extended dictionary contains many types of data, including dictionary entries, corpus text, critical annotations and semantic knowledge. There are also various types of connections among these entities: dictionary entries refer to the corpus through citations and their references, words in the corpus can be found in the dictionary as word forms, critical annotations are attached to the corpus but they are also linked to the relevant headwords, and pieces of knowledge from external sources are attached to headwords.

An SQL database system is a natural choice in order to store the electronic version of the dictionary. A NoSQL database was chosen for corpus texts and critical annotations, and a graph database for describing semantic and structural knowledge in a unified system.

### 4.1.1 Headword Database

As the headwords were authored using a predefined XML schema (a detailed and structured data model) importing them to a database was a straightforward process. Following the headword structure we designed a simple database model to store them. During database modeling we also took possible data queries into consideration and created separate data tables and indexes for those pieces of data. The system was implemented using the MySQL database engine.

### 4.1.2 Corpus Storage

As our goal was to maintain a complete list of text occurrences at each dictionary entry we stored the entire corpus in the system. Since our work concentrates on the dictionary and we store corpus only for related citations, we decided not to use complex corpus analysis tools but choose a simple solution to store the attached text fragments.

Corpus texts are encoded using XML TEI, but they are far less structured than dictionary entries. In contrast to the data-centric dictionary XML data, the corpus XML is a document-centric data set. In order to store it a schemaless (NoSQL) database system, MongoDB was chosen. NoSQL databases have the required flexibility to store and organize such documents, and they also provide scalability and performance with regard to accessing them. The corpus was stored at various levels of granularity in the database (e.g. chapters, sections, paragraphs, lines, etc.) to ensure that we can create different kinds of relations between the dictionary and corpus.

Critical annotations are traditionally text fragments attached to various parts of the corpus. Our system stores these text annotations using the same technique in the NoSQL database. It also uses the same referencing system: annotations refer to corpus entities at various granularity levels, and the word forms found in annotations can be looked up in the dictionary.

### 4.1.3 Knowledge Store

The incorporated external knowledge and internal references between dictionary entries, corpus parts and critical annotations can all be represented as graphs. Since external knowledge bases (LOD sources) use the Resource Description Framework (RDF) to store and exchange this kind of information, we decided to follow this practice to avoid transformation during the import process.

RDF is a very flexible system to represent knowledge. Its atomic data entity is a triplet consisting of a subject, a predicate and an object. Together they represent a fact. From these building blocks a

graph can be constructed to store a knowledge base. Due to this graph storage it can be extended or modified without altering previously inserted knowledge. Its flexibility is very useful if we do not know in advance what kind of knowledge will be stored or how it will be extended later. In order to implement the knowledge store we have selected the RDF4J open source software.

In addition to the retrieved DBpedia RDF dataset, we also assigned resource identifiers for corpus and dictionary entries and stored these and their relations in the RDF database. This way we stored the knowledge about their relations and also the relevant semantic information in a unified system.

RDF also provides a very powerful graph query language, SPARQL. Inserting knowledge as RDF triplets is a straightforward process but performing a query on a graph database is a more complex task, since we may want to formulate complex conditions for graph matches. SPARQL allows such constructs. We designed many kinds of query functions in SPARQL in order to support corpus and dictionary information retrieval, as detailed in the next section.

## 4.2 Query Functions

The main use of the system is to access dictionary data. The input is typically a keyword (headword, writing variation, word form) and its attributes (e.g. part-of-speech role), and the result is a dictionary entry. It is also common to restrict the query by selecting a part of the corpus in which the keyword has to be found. These queries were implemented in SQL using the dictionary database.

For completeness, we also developed corpus search methods attached to the dictionary. In this case the input is again a keyword and its attributes, but the result is a corpus entry. This is similar to dictionary search, but it also supports query expansion to replace keywords with other data (e.g. with other word forms), while keyword and corpus normalization are used to provide better precision and recall during the search process.

Finally, knowledge retrieval is the third kind of search function that allows queries regarding the semantic information of the stored data. In this case the input is a complex query that specifies entities (e.g. places or persons, dictionary and corpus entries), their attributes and relations to other entities. These functions were implemented in SPARQL that queries the RDF knowledge store of semantic and reference information.

One interesting aspect of these query functions is the query expansion backed by the RDF knowledge store (Varga et al. 2003). The search engine is capable of recognizing concept and entity names in the user's query and it is able to replace them with related keywords when needed. For example, it is able to answer queries like

(XX) Retrieve headwords related to cities

By using knowledge from external data sources like DBpedia the following complex query can be also answered

(XX) Retrieve citations related to geographic locations within 100 km of Constancinapoly.

The system identifies that *Constancinapoly* is a writing variation of Constantinople using the dictionary. It also finds out from the attached RDF store that it is a city with a known geographic location. To answer the query it searches for other geographic locations in the knowledge base that are within the given range. After finding them it retrieves the relevant citations from their dictionary entries and presents the results back to the user.

### 4.3 Web Interface for the Extended Dictionary

We created a web-based software tool to import, store, query and display the corpus, the dictionary and critical annotations. This tool provides an administrative interface for importing XML data (the dictionary, corpus and annotations) and it also has a user interface for the previously outlined query functions and for displaying their results. The web interface was implemented using the open source ProcessWire PHP framework with additional modules developed for XML import and displaying dictionary entries and corpus annotations.



Figure 2: The Mikes dictionary web interface showing a dictionary entry (excerpt) and an original letter from Kelemen Mikes.

## 5 Summary

Creating an author's dictionary is a labor-intensive task, and information technology can greatly help researchers in this process. It provides efficient tools for making concordance lists, editors for supporting the authoring process, various methods to store available data, and a user interface for accessing the dictionary and attached corpus. We have demonstrated how the digital Mikes dictionary was created using such IT methods and tools.

Our aim was then to take a step further by introducing knowledge-based methods in creating and using the dictionary. In order to achieve this we extended the dictionary with additional knowledge and also connected it to already available knowledge sources like critical annotations and linked open data sets. We created an RDF knowledge base by retrieving information from these sources and linking them to the appropriate dictionary entries and corpus locations. This base contains semantic information imported from linked open data sources, and it also incorporates knowledge from critical annotations by linking them to dictionary and corpus entries. We also developed advanced dictionary and corpus search functions that utilize this knowledge using query expansion to acquire higher quality search results.

There are many possible benefits of developing an extended author's dictionary. For example, it enables access to much more information about an author's vocabulary than before, and provides better awareness of the related linguistic, historical and semantic information. This vast knowledge base can also be a basis for further research. We can conduct new kinds of analyses on the integrated data, text and knowledge base. Finally, the added knowledge could also improve traditional corpus analysis and search methods by supporting corpus normalization and query expansion techniques.

There are many possible ways to take this research even further. Extending and analyzing this knowledge store opens up many possibilities. We are investigating how to extend the semantic knowledge beyond the scope of named entities, and how bibliographical data found in critical annotations can also be incorporated. Another interesting research topic is how to incorporate other kinds knowledge acquisition methods into the system. For example, we are experimenting with controlled natural language interfaces to allow researchers to introduce new knowledge to the RDF store using a natural language online interface.

Most of the software tools developed in the framework of the DHmine project is open source, and they are available in the following GitHub repository: <https://github.com/mtwebit/dhmine>.

## References

- Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z. (2007) DBpedia: A Nucleus for a Web of Open Data. In: *Aberer K. et al. (eds) The Semantic Web. Lecture Notes in Computer Science*, vol 4825. Springer, Berlin, Heidelberg
- Bartalesi, V. et al. (2015). Towards a Semantic Network of Dante's Works and Their Contextual Knowledge. In *Digital Scholarship in the Humanities*, pp. 28-35.
- Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. In *Computers and the Humanities*. 14(2), pp. 83-90.
- Hartmann, R.R.K., James, G. (2001). *Dictionary of Lexicography*. London and New York: Routledge.
- Hopp, L. (1966-1988). *Mikes Kelemen összes művei*. L. Hopp (eds.) Budapest: Akadémiai.
- Kiss, M. (2012). The Digital Mikes-Dictionary. In G. Tüskés et al. (eds.) *Literaturtransfer und Interkulturalität im Exil [...]*. Bern: Peter Lang Verlag, pp. 288-297.
- Mártonfi, A. (2014). Számítógép és írói szótár – különös tekintettel a készülő József Attila szótárra. In *Magyar Nyelv*, 110(1), pp. 30-46.
- Mattaush, J. (1990). Das Autoren-Bedeutungswörterbuch. In Hausmann, Franz Josef et al. (Hrsg.) *Wörterbücher–Dictionaries–Dictionnaires [...], An international encyclopedia of lexicography [...]* 2. Berlin–New York: Walter de Gruyter, , pp. 1549-1562.
- Mészáros, T. (2016). Agent-supported knowledge acquisition for digital humanities research. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3936-3941.
- Nugues, P. et al. (2016). From Digitization to Knowledge: Resources and Methods for Semantic Processing of Digital Works/Texts. Workshop proceedings. In *Digital Humanities 2016*, July 11, 2016, Krakow, Poland.
- J. Soltész, K. et al (1973-1987). *Petőfi-szótár*. Budapest: Akadémiai Kiadó.
- Varga P., Mészáros T., Dezsényi C., Dobrowiecki T.P. (2003) An Ontology-Based Information Retrieval System. In: *Chung P.W.H., Hinde C., Ali M. (eds) Developments in Applied Artificial Intelligence. IEA/AIE 2003. Lecture Notes in Computer Science*, vol 2718. Springer, Berlin, Heidelberg

## Acknowledgements

This research has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00013). Sándor Jan Dobi developed software for the corpus and knowledge store, and he also implemented a programming interface to these components. The initial concordance list and the first version of the web-based dictionary (not shown in this paper) were created by Attila Mártonfi.



# Analyzing User Behavior with Matomo<sup>1</sup> in the Online Information System *Grammis*

**Saskia Ripp, Stefan Falke**

*Institut für Deutsche Sprache, Mannheim*

*E-mail: ripp@ids-mannheim.de, falke@ids-mannheim.de*

## Abstract

The grammatical information system *grammis* combines descriptive texts on German grammar with dictionaries of specific word classes and grammatical terminology. In this paper, we describe the first attempts at analyzing user behavior for an online grammar of the German language and the implementation of an analysis and data extraction tool based on Matomo, a web analytics tool. We focus on the analysis of the keywords the users search for, either within *grammis* or via an external search platform like Google, and the analysis of the interaction between the text components within *grammis* and the integrated dictionaries. The overall results show that about 50% of the searches are for grammatical terms, and that the users shift from texts to dictionaries, mainly by using the integrated links to the dictionary of terminology within the texts. Based on these findings, we aim to improve *grammis* by extending its integrated dictionaries.

**Keywords:** user behavior, online information systems, automated tracking, Matomo, online grammars, online dictionaries, keyword analysis

## 1 Introduction

While much is known about the use of online dictionaries, only little is known about the use of complete online grammatical information systems such as *grammis*. *Grammis* is a grammatical online information system, hosted by the Institute for the German Language in Mannheim (Institut für Deutsche Sprache, IDS), that combines descriptive texts on German grammar with dictionaries on grammatical terminology and selected word classes (*grammis* 2018). It was created in the early 1990's as a research project that dealt with the complexity of writing grammars<sup>2</sup> and the challenges of transferring the linear structures of grammar books into hypertext formats. The CD-ROM-based version was changed to an online version in 2004 (Schneider & Schwinn 2014), and had its last redesign in 2017/18. The goal was to update the system technically in order to adopt the latest standards in web development, like the three-tier architecture including a MVC-PHP-Framework and mobile friendly design (Krasner & Pope 1988; Olanrewaju et. al. 2015). The restructuring and updating of the contents which started with the terminology on German grammar (Suchowolec et al. 2017) is still ongoing. The redesigned version of *grammis* went live on January 23<sup>rd</sup> 2018, while the old system will run in parallel until the final server shutdown in April 2018<sup>3</sup>.

Today, *grammis* is structured into the three main parts: “Forschung” (Research); “Grundwissen” (Basic Knowledge); and “Ressourcen” (Resources), as shown in Figure 1.

1 The web tracking system Piwik was renamed Matomo on January 9<sup>th</sup>, 2018.

2 In this paper we use the term “grammar” in the sense of reference books of the grammar of a certain language, in this case German.

3 In the course of this paper we will refer to the old version of *grammis* as the “old *grammis*” and the new version as the “new *grammis*”.

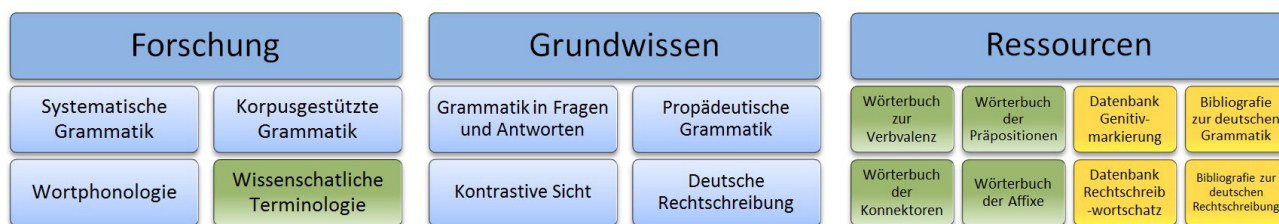


Figure 1: Main structure of *grammis* from the *grammis* homepage.

These three main modules contain the components shown in Figure 2, where the light blue components are full text<sup>4</sup> components containing descriptive text passages on grammatical topics, the green ones are the dictionaries, and the yellow ones are research tools or bibliographies.

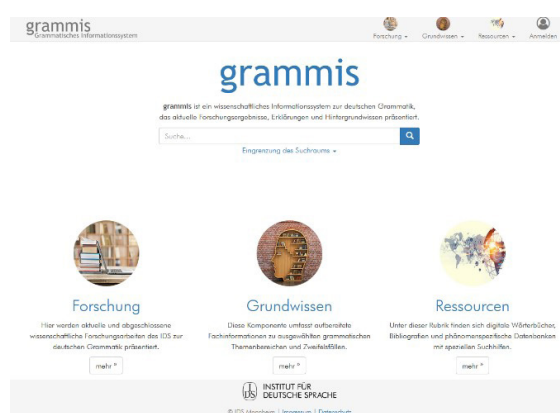


Figure 2: Components of the new *grammis*.

With the release of the new *grammis*, a new project focus lies on the evaluation of the system from the users' perspective. In order to make *grammis* more user-friendly, one aspect of restructuring and updating its contents is to take the actual users into account.

By doing so, we want to fulfil the demand for research on grammar use which has not yet been met, despite calls by Helbig (1992), Klein (2004) or Hennig (2010), in contrast to research on dictionary use, which has become a canonical research field in (online) lexicography (c.f. Müller-Spitzer 2014b; Tarp 2009). With regards to dictionaries, Lew (2015) states that the web has and will continue to bring out a great number of online dictionaries, which we believe is also true for online grammars. Besides *grammis* for the German language, other grammars like *canoo.net* (2018), or the recently provided “Variantengrammatik des Standarddeutschen” (Variation Grammar of Standard German, Dürscheid et al. 2018) are also available on the internet.

To begin with, we focus on the interaction between full texts and dictionaries within our online information system *grammis*, on the question of what exactly the users are searching for, and also look at the search results lists of the full text search feature. For that purpose we take a closer look at the four main dictionaries in *grammis*: the dictionary of the “Wissenschaftliche Terminologie” (Scientific Terminology), the “Wörterbuch der Präpositionen” (Dictionary of Prepositions), the “Wörterbuch der Konnektoren” (Dictionary of Connectors<sup>5</sup>) and the “Wörterbuch der Affixe” (Dictionary of Affixes).

<sup>4</sup> We distinguish full texts from dictionary entries, whereas full texts mean descriptive text passages on German grammar.

<sup>5</sup> In *grammis*, the class of connectors unites expressions that organize specific semantic relations between sentences. Traditionally, these include conjunctions, some adverbs and particles (*grammis* Konnektoren 2018).



Our interest lies in their interaction with the full text components describing the grammar of the German language. The three dictionaries on the word classes *prepositions*, *affixes* and *connectors* provide mainly grammatical information on the respective lemmas, like the position in a sentence, phrase structure, government, function and meaning. Access to the dictionaries is provided either through the separate dictionary components or via hyperlinks within the full texts on grammar. The latter are implemented as modal windows which open when the user clicks on a linked term. The terms within the dictionary components themselves can be accessed through either an alphabetical list or by typing a word into the search field.

For a quantitative overview of the content of *grammis*, Tables 1 and 2 show how many texts and entries per dictionary *grammis* has.

Table 1: Number of texts per component.

Component	Number of texts
Systematische Grammatik (Systematic Grammar)	928
Korpusgrammatik (Corpus Grammar)	136
Wortphonologie (Word Phonology)	11
Grammatik in Fragen und Antworten (Grammar in Questions and Answers)	223
Propädeutische Grammatik (Propaedeutic Grammar)	205
Kontrastive Grammatik (Contrastive Grammar)	867
Deutsche Rechtschreibung (German Orthography)	80
<b>Total</b>	<b>2,450</b>

Table 2: Number of entries per dictionary.

Component	Number of entries
Verbvalenz (Verbal valency)	677
Präpositionen (Prepositions)	132
Konnektoren (Connectors)	369
Affixe (Affixes)	285
Terminologie (Terminology)	388
<b>Total</b>	<b>1,851</b>

Since a lot of dictionary user research has been done (as stated above), we will take the results of previous studies within online dictionary user research into account, especially concerning the research methods.

## 2 Previous Research on (Dictionary) User Behavior

Research on grammar user behavior is still a desideratum in grammar writing. While the call for research on the user's perspective in grammar use came up in 1992 (c.f. Helbig 1992), the first attempts at investigating user needs and behavior concerning grammars of the German language were only recently made by Hennig and Lotzow (2016). Besides a questionnaire-based study (Hennig 2010) on the use of German grammars in general (What do you do when you have a question on grammar? Which grammar do you use and for what purpose? What do you expect of a grammar? etc.), Hennig (2010), Hennig and Löber (2010) and Hennig and Lotzow (2016) mainly investigated user behavior for the so-called *Duden-Grammatik* (Duden Grammar, Duden 2005; 2009), which is considered to be the most frequently used German grammar book in Germany (Hennig 2010: 20). These studies

were based on problem solving tasks and questionnaires with only a few subjects ( $n = 42$  for Hennig & Löber (2010) and  $n = 6$  for Hennig & Lotzow (2016)). As stated above, online lexicography has been taking the users' perspective into account for quite a long time (see short summary in Bergenholtz & Johnsen 2005: 118f.), which is why we want to draw on this research tradition. Bergenholtz and Johnsen (2005; 2007), for example, state that surveys based on questionnaires are problematic because they do not reflect the real user situation and are based on the users' memories of a rather artificially-constructed situation, and are sometimes even based on predictions of what the users think they might do in the future. Instead, they showed how log file analysis for dictionary research can be useful for gaining information about user behavior:

With a log file, you can track every single use of the dictionary, depending, of course, on the search possibilities. If it is only possible to search for the lemma, only data for the first access step in the dictionary will be available. Which lemmas have been looked up how often? Which lemmas have never been looked up at all? And which words have been used in the search field without result, i.e. how many and which lemma lacunas does the dictionary use indicate? (Bergenholtz & Johnsen 2005: 121)

Although this method has some limitations (e.g. the search possibilities mentioned above), it has the advantage that a huge amount of user data can be analyzed at once, unlike in user studies. Furthermore, Müller-Spitzer (2014a) states that using log files for the research into dictionary use is a "promising method" as it captures the usage in a real and authentic user situation. Previous research on dictionary use, however, teaches us to be careful with the interpretation of log file data, because the research process cannot be controlled, meaning that neither the background information on the users nor the contexts of the use or the success of a look up process can be determined exactly (Müller-Spitzer 2016; Lew 2011; Bergenholtz & Johnsen 2007). What the log files also cannot tell are the problems or the intentions the users might have had. Nevertheless, by using log file analysis, De Schryver and Joffe (2004) determined words that users searched for and that were not available in the dictionary. Subsequently, they added these missing lemmas to the dictionary, which resulted in an increase in the hit rate and certainly in greater satisfaction among the users.

Still, the use of server log files for the analysis of user behavior is problematic due to the fact that the server log files are in principle limited to what the server actually handles [...]. Only those activities of the user can be logged which are processed server-side, as opposed to those which are executed by the client (usually a web browser). Thus, the level of detail potentially included in log files is determined by the division of labor between server-side and clientside computing. Issues of data privacy can also be a limiting factor in log file analysis. (Lew 2015: 12)

Another way of collecting user data is to use web analytics systems, as was done, for example, by Lorentzen and Theilgaard (2012). They used Google's web analytics system, Google Analytics, to track the user behavior for the Danish online dictionary *ordnet.dk*, mainly to find out where their users came from (search engine, bookmarks, or from another website), and which lemmas they searched for that were not available in the dictionary. The results made it possible to improve the search process, e.g. the search for lemmas could be improved by adding further inflectional forms that could not be found before. They also combined the method with questionnaires, think-aloud-protocols and interview-based studies with selected users which additionally provided information on the users' backgrounds (intentions, satisfaction with the tool etc.). Tiberius and Niestad (2015) also used Google Analytics for the *Algemeen Nederlands Woordenboek* (ANW, Dictionary of Contemporary Dutch) to test the feature of presenting four different search possibilities to the users and the hypothesis that it would help them to better define what they are looking for, and that it would encourage them to use more than one search option. They stated that "Google Analytics is particularly useful for graphical overviews, for instance, of the types of visitors and the path most of them follow through the ANW application" (Tiberius & Niestad 2015: 29).

As using log files for the analysis of user behavior clearly has the same limitations for online grammars as for online dictionaries, we use the web tracking system Matomo (2017), which does not read the server log files, but tracks user behavior directly on the website, much like Google Analytics does in the studies mentioned above. Despite the stated disadvantages, we see a great potential in analyzing data that were created in real usage situations without overinterpreting the results. Therefore, we focus on what data can be collected with the web analytics system Matomo, and how we can make use of the collected data for the purpose of studying user behavior.

In our opinion, the biggest difference compared to dictionary research is that we cannot directly see if the users have found what they were looking for. In the case of lemma-based searches in dictionaries a log file or a web analytics system can reveal rather easily whether a searched word was found or is part of the dictionary at all. In contrast, we assume that searches in a grammar are more complex in most cases, because they are very likely to aim for the explanation of grammatical concepts or the correct use of a grammatical form within a sentence or a text. This is why we expect the searches to consist of more than only one word, although we also expect single-word searches that might either be a lemma or a grammatical term. In the first case, we assume that the user is not searching for the meaning of the respective lemma, but rather for an explanation of its rules of inflection or function within a sentence or context. The search for terminology might be similar to the usage situation of a dictionary, because the user might look for a rather short explanation for the searched term in order to understand its concept. Nevertheless, with learning more about what kinds of search string the users enter, we hope to improve the search algorithm of the database and to obtain some insight into what we should present the users as a search result. As a start, we decided to focus on the integrated dictionaries in *grammis*. In the case of searches that do not refer to terminology, we need to develop a categorization system that defines the searches, by preference automatically, in order to quantify them.

### 3 Aims and Research Questions

Having updated the system technically, our current aims are to improve the grammatical information system *grammis* for the users, and to bring more users to our site. To begin with, we analyze the data collected by the web analytics system Matomo. Some general questions we want to answer as a first attempt to analyze user behavior in *grammis* are: How can Matomo be used in the analysis of user behavior? What data need to be tracked and can be tracked with Matomo? Who is using *grammis* when, where, how often, etc.?

In order to gain more information about the users' behavior with respect to the integrated dictionaries, we focus on the following three main research questions:

**Research question 1:** What do the users search for?

By answering this question, we want to gain information on the users' intentions and interests, especially with regards to the content of our dictionaries.

**Research question 2:** Do users use the integrated dictionary links by opening the modal windows when reading the full texts?

By answering this question, we want to find out if the integrated links are used at all, and what dictionary content needs improvement.

**Research question 3:** Which results (in the ranking of the results list) do the users select after a search?

By answering this question, we want to find out if the ranking of the search results for the full text search needs to be improved.

## 4 Matomo

Matomo is a web analytics platform to track Key Performance Indicators such as: visits, search keywords, site impressions etc. (Matomo 2017). Matomo in its basic configuration is free of charge, but some premium features are only available via a yearly subscription. The collected data are fully owned by the IDS and stored on a server which is located within the IDS network. This is in line with the high standards of the German and EU privacy policies, and the reason why we chose Matomo as a tracking system for our websites instead of Google Analytics, where the data will be stored on Google servers all over the world.

Matomo is implemented via a short JavaScript code snippet that is included in the head part of every web page (Matomo 2017). It is also possible to configure the tracking code to individualize which actions should be tracked. For example, we had to customize the script to track the use of the modal windows (see Section 6.2).

We included the premium feature *Search Engine Keywords Performance* to get all keywords from external search engines like Google, because these keywords will not be passed on from Google to Matomo (Matomo 2017) in the default configuration.

As to the visitors' actions, there is a difference between *hits* and *visits* (Matomo 2017). *Hits* are the number of page impressions, showing how often a page is requested in total. A *visit* is a stay on the website of one specific user. The user is anonymous, but recognizable to Matomo. During a visit a user will perform at least one single hit on one page. So, a hit is always a part of a visit, and a visit contains at least one hit. If the user is idle for more than 30 minutes, Matomo will count their next click as a new visit.

## 5 Basic User Statistics

In this section we give a short summary of the facts and figures of *grammis* during the period under examination. We collected the data between August 21<sup>st</sup>, 2017 and March 20<sup>th</sup>, 2018. This period we refer to as the overall period. With regards to the collected data for the search keywords, it was divided into four different periods (shown in Table 3). This is due to the fact that data for the new *grammis* could be collected only after its activation in January 2018 and that the external searches could be collected only after the implementation of the necessary Matomo feature added in February 2018 (see Section 4).

During the overall period we had a total of 478,914 visits, including 475,459 visits to the old *grammis* and 3,455 visits to the new *grammis*. During this time, we had a total of 871,291 hits (i.e. page impressions). This divides into 845,863 for the old *grammis* and 25,428 for the new *grammis*. The average visit lasted 1:36 minutes with 1.8 actions per visit of the old *grammis*, and 8:49 minutes with 8.5 actions per visit of the new *grammis*. The bounce rate (i.e. the rate at which users leave *grammis* after only visiting one page) was 79% for the old *grammis* versus 35% for the new *grammis*. Most visitors came from Germany (60%), followed by Italy (13%) and Spain (3.5%). Overall, the vast majority of visitors were from Europe (94.7%), with a few from Asia (2.7%), the Americas (2.1%), or from the African continent (0.4%).

Since the redesign, fewer visitors came via search engines. Instead, they found their way to the site directly (via bookmark or from other website links), which could be an explanation for the different bounce rate values. While many visitors of the old *grammis* came via Google just to visit one single page, the new *grammis* was (and still is) not ranked highly enough in the results list to get these

“one-click-visitors”, which is why we suppose that at the moment we mostly have users who know *grammis* and use it regularly. With the final server shutdown in May 2018 (mentioned in Section 1), the old *grammis* will no longer be available, which is why we expect a better ranking in the Google search results which should lead to a rise in visitors and views. Additionally, we allowed Google to crawl our websites, a process that is not finished yet. With the completion of the Google crawl and the rising numbers of views we also expect a rise in user numbers due to a higher ranking in the Google results lists.

Looking at the internal searches in the overall period, we have 18,294 searches with 4,151 unique keywords in the old *grammis*, and 2,575 searches with 1,126 unique keywords in the new *grammis*. The keywords are case sensitive, so that the keywords “Verb” and “verb” are two unique keywords. For the actual numbers on the search data which was used in the keyword analysis, see Table 3.

## 6 Data Extraction and Analysis Tool

In order to make use of the data collected by Matomo to answer our research questions, we had to configure the data extraction and implement a unique data analysis tool into the admin backend of *grammis* (see Figure 3).

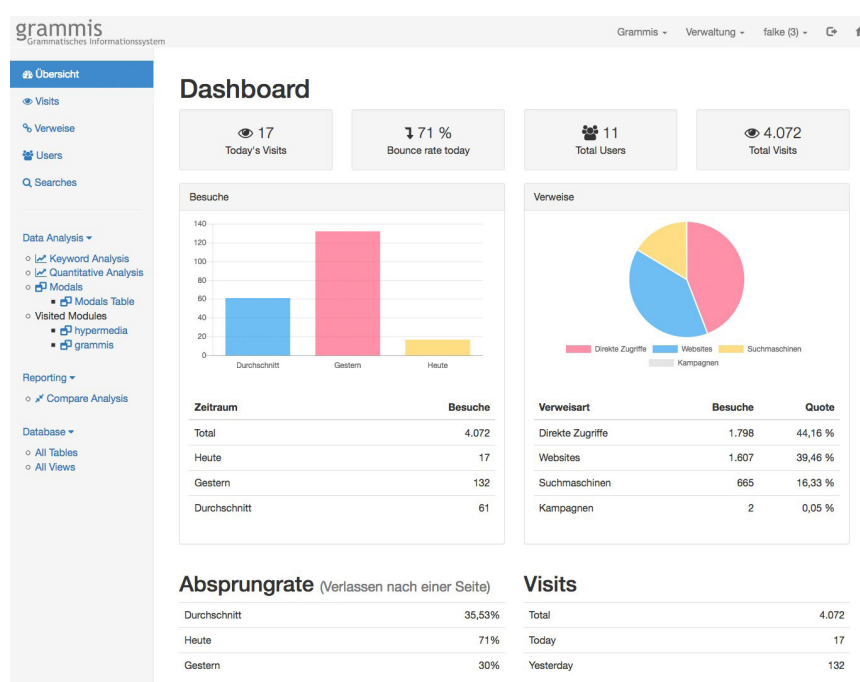


Figure 3: Data extraction and analysis tool.

The data extraction and analysis tool is written in PHP and included in our MVC-Framework. It retrieves the data from Matomo via their API, which can be called with an HTTP request including the query parameters (Matomo Tracking 2018). This makes it possible to retrieve the data for the period we want to look at and with several filter settings to retrieve only those datasets that are interesting for a given analysis. We included several charts and tables in our tool to look at the data from different points of view. Some charts and analyses made it necessary to process and convert the data from Matomo for our needs, so we implemented several methods to achieve this, e.g. for the extraction of the modals windows.



## 6.1 Tracking of the Keywords

The first research question (What do the users search for?) could not be answered with the standard configuration of Matomo, so we had to customize Matomo to make it possible to track the keywords<sup>6</sup> the users enter in both an external search engine like Google and in the *grammis* internal search field. To do this, we had to install the premium feature *Search Engine Keyword Performance* and connect it with the IDS Google Account (see Section 4), which was done on February 17<sup>th</sup> 2018. Since then, it has been possible to gain access to the keywords entered in Google<sup>7</sup>. We then extracted the keyword list, sorted by the number of hits for the old and new *grammis*, and combined the external list with the internal lists collected by Matomo. We thus had one big keyword list, including the external keywords for the old *grammis* and the new *grammis* and the two internal search keyword lists for both systems. We inserted the complete list into our database to analyze the data. The whole list consists of 6,040 data sets. The periods and number of total searches for each sub list are given in Table 3.

Table 3: Number of searches for different time periods and systems.

Platform	Type	From	To	Number of searches
old <i>grammis</i>	internal	2017/08/21	2018/03/09	4,069
	external	2018/02/17	2018/03/09	999
new <i>grammis</i>	internal	2018/01/23	2018/03/09	810
	external	2018/02/17	2018/03/09	162
<b>Total</b>				<b>6,040</b>

We then categorized the keywords the users were looking for to specify the type of the search. For this, we defined three categories: searches for terminology, searches for *object words*<sup>8</sup>, and meta searches. Search strings of more than one word of which at least one was a grammatical term were counted as terminology searches, even if an object word was included. Meta searches feature keywords that consist, for example, of the name of a component or *grammis* itself, author names, etc.

For this purpose, we wrote a script that matched the keywords automatically with our integrated dictionaries to classify the keywords into terminological and object word searches. The categorization as meta search is done by hand (see two paragraphs below). With the matching, we additionally wanted to check the coverage of our dictionaries and therefore if the user is able to find what they are looking for at all, or if we have to add more terms and lemmas to the dictionaries.

When after that matching there were still keywords left which could not be found in our dictionaries, we checked these against external sources. First, we used the terminology list of *canoo.net* to find terms that are not part of our dictionary of terminology. When an object word keyword was not contained in our internal dictionaries we implemented an alignment with the DWDS by using their API (DWDS 2017) to get the word class for this keyword. Since we do not have a full word list within *grammis*, we cannot classify nouns and other word classes automatically.

Since there were still some keywords which could not be classified automatically, we exported the list to a Microsoft Excel document and categorized the unclassified keywords by hand. The categorization of the search strings that consisted of more than one word was challenging, especially when it

<sup>6</sup> In this case *keyword* means the whole string a user is entering into the search field. The keyword can consist of a single word, multiple words or a whole phrase.

<sup>7</sup> For all other search engines this is not necessary. Since to date the Matomo data show that almost no users come from search engines other than Google, this was a necessary configuration.

<sup>8</sup> We use the term *object word* for all lemmas that are not terminology, which can be any words of a certain word class of which information is sought on. In a traditional sense the term *object language* defines what is the object of study in a certain language, while *metalanguage* defines talking about the objects of a certain language itself (Lehmann 2018).



came to interpreting the users' actual intentions. For example, the search string "*in oder auf*" ("in or on") – which turned out to be a very typical type of search – is classified as PREP, KONJ, PREP by our automatic alignment. Obviously, the user is not looking for all three word classes, but for the use of the prepositions alone, meaning that they aim to find the answer to an alternative question in which the user is looking for the correct preposition in a certain function or context. Since the classification of these object word searches consisting of more than one word is not trivial, and needs to be defined clearly, we did not include these for the analysis of the type of word classes that are searched for the most (see Section 7 and Table 5).

After the manual classification of the keywords, we had a list with two types of classifications: the type of request (terminological, object word, or meta search), and the classification of the specific word class of an object word (see Table 4). The statistics of this analysis are integrated into our admin backend of *grammis*.

## 6.2 Tracking of the Modal Windows

To answer research question 2 (Do users use the integrated dictionary links by opening the modal windows when reading the full texts?), we needed to take a look at the pages of the full text components the users visited and summed those up for each component of *grammis*. Therefore, we sorted the URLs in the hits tracker in Matomo and searched for the included component, since our URL structure is the domain followed by the name of the component, e.g. <https://grammis.ids-mannheim.de/systematische-grammatik>, where <https://grammis.ids-mannheim.de> is the domain, and *systematische-grammatik* is the component "Systematische Grammatik". A current text of this component is then browsed to via the ID of the database entry, which follows the component after a slash, e.g. <https://grammis.ids-mannheim.de/systematische-grammatik/244> to browse to entry 244 (in this case "Wortarten" (word classes)). With this, we can track and count each hit for every time a user directly navigates to a component, which can be through a link within *grammis*, a result page on an external search engine, or a bookmark.

To obtain more information about certain object words or terminology within the full texts, the texts contain links to the respective dictionary sources. These links open as modal windows, i.e. they open as a new layer on the current webpage, so that after closing the modal window the website is still open and the user does not need to click the browser's back button. Since the content of the modal window is loaded via an AJAX request, it is not counted as a hit by Matomo. Actually, in the default configuration an AJAX request is not tracked at all. To track the modal windows we had to adjust Matomo by including a few lines of JavaScript on every page to catch the event of opening a modal window. The script then sends the name of the modal window, its title, the URL of the current page the user is reading and the URL to the content which will be loaded and shown within the modal window to the Matomo database. Having done this, the use of the modal windows can be counted and the statistics were also integrated into our data extraction and analysis tool.

## 6.3 Tracking Behavior after a Search

Currently, the ranking of the search results in *grammis* is random due to the configurations of the database, which is why we want to analyze how the users interact with the results list in order to optimize the ranking of the results for the users. To answer research question 3 (Which results (in the ranking of the results list) do the users select after a search?), we planned to track the behavior of a user after a search. In the old *grammis*, the tracking of the links a user chose from the results page was not possible for Matomo due to several parameters which were sent with the results page and did not make it possible for Matomo to distinguish whether a link was a result or a search itself. With

this in mind, we configured the new *grammis* to track every internal search and the pages a user visits after a search. Still, the tracking of the chosen results causes some problems as it does not give us a complete tracking of the search-and-find process (in its standard configuration). It only tracks the keywords and the average number of hits of a page that has been chosen for a respective keyword, but it neither provides the exact pages that were chosen from the results list, nor the ranking number of the result in the list. Instead, Matomo tracks at which rate the user is leaving the page directly after the search without clicking on a result at all, and which pages exactly were clicked on after a search, but without providing the keyword that led to the result. So, for an evaluation of the result pages that are generated by our search algorithm from the database, we need to configure both the search algorithm and Matomo according to our needs and, furthermore, to combine it with a qualitative analysis of the users' decisions for certain results in the lists. Both could not be done within the timeframe for this paper but will be done in the future.

## 7 Results

### Research question 1: What do the users search for?

Concerning research question 1, Table 3 shows that 80% of the search requests come from the internal search (4,879 of 6,040).

As can be seen in Table 4, the numbers for terminological searches (46.23%) and object word searches (52.88%) are almost equal, with a slightly higher number of object word searches. Looking at the overall hit rate, the numbers change (53.15 % for terminology and 44.37% for object words). This is caused by the lower number of terms in the dictionary compared to the number of entries for the dictionaries of the four word classes (see Table 1). This shows the high interest of the users in terminology and object words and, additionally, that the dictionary of terminology is highly requested. The number of meta requests is only 0.7% of all requests, but makes up 2.41% of the total clicks. Looking into the data, it became clear that users are using Google like a bookmark for entering the site, meaning that the users search for the word “grammis” in Google and come to the website via this link.

Table 4: Types of requests.

Type of request	Number of requests	Rate	Number of hits	Rate
(undefined)	10	0.17%	10	0.04%
object word	3,194	52.88%	10,584	44.37%
terminology	2,792	46.23%	12,688	53.19%
meta search	44	0.73%	574	2.41%
<b>Total</b>	<b>6,040</b>	<b>100.00%</b>	<b>23,856</b>	<b>100.00%</b>

As a next step, we took a closer look at the object word searches and what kinds of word classes were searched for. As stated in Section 6.1, we analyzed only the object word searches that consisted of one word due to the complexity of categorizing the multiple word keywords. Of course, we also looked at the length (in words) of the keywords. The most common search string consists of only one word (66%), while the longest string contains 27 words and basically represented a whole sentence. Taking this percentage into account, we covered and thus categorized at least more than half of the searches in *grammis* with this method. Another constraint for this analysis was that we only took those object words into account that referred to one word class alone. Obviously, many words can be defined as more than one word class due to the respective function or meaning in a certain sentence. Since this categorization is a rather complex task, too, we will do this as further research as well. Nevertheless,

for the analyzed part of the data, Table 5 shows that most users are looking for verbs (4,924 hits), followed by prepositions (1,325 hits), conjunctions (951 hits), and adverbs (837).

Table 5: Word classes of the object word searches.

Word class	Number of distinct searches	Hits
Verb	1,227	4,924
Preposition	260	1,325
Conjunction	184	951
Adverb	287	837
Noun	475	829
Pronoun	123	478
Adjective	174	344
Article	15	30
<b>Total</b>	<b>2,745</b>	<b>9,718</b>

The last step was the matching of the keywords with our integrated dictionaries and the external sources (as described in Section 6.1), on the one hand to have them categorized automatically, and on the other hand to see which words are available in our dictionaries. The results in Table 6 show that more than half of the searches (58.98%) could not be classified at all, and that only 7.1% of the requests were part of our dictionary of terminology, whereas 12.06% of the requests could be found in the terminology list of *canoo.net*. This might be due to the fact that the terminology of *grammis* for the most part contains highly scientific terminology<sup>9</sup> and *canoo.net* contains more traditional German grammar terms that are used in school, for example. The data show that the terminology searches often include said traditional terms, which is an important result for us when it comes to the expansion of our dictionary of terminology. Table 6 also shows that only 3.59% of the search requests could be found within the dictionary of prepositions, 0.05% within the dictionary of affixes, 5.21% within the dictionary of connectors, and 13.01% within the dictionary of verbal valency.

Table 6: Coverage of the keywords in the dictionaries.

Dictionary	Number of requests	Rate
<i>not classifiable</i>	2,367	58.98%
Wörterbuch der Präpositionen (Dictionary of prepositions)	144	3.59%
Wissenschaftliche Terminologie ( <i>canoo.net</i> ) (Scientific terminology <i>canoo.net</i> )	484	12.06%
Wörterbuch der Affixe (Dictionary of affixes)	2	0.05%
Wörterbuch der Konnektoren (Dictionary of connectors)	209	5.21%
Wissenschaftliche Terminologie (Scientific terminology)	285	7.10%
Wörterbuch zur Verbvalenz (Dictionary on verbal valency)	522	13.01%
<b>Total</b>	<b>4,013</b>	<b>100.00%</b>

## Research question 2: Do users use the integrated dictionaries when reading the full texts?

To answer research question 2 we analyzed the use of the modal windows.<sup>10</sup> The results show that the option to open a modal window while reading a full text is used by only about 9% of the visitors (329

<sup>9</sup> Since the main part of *grammis*, the “Systematische Grammatik” (Systematic Grammar) is based on the grammar book GDS (*Grammatik der deutschen Sprache* (Grammar of the German language) by Zifonun et al. 1997), the dictionary of terminology in *grammis* is also based on the grammatical terms used and developed in that grammar.

<sup>10</sup> The duration for this analysis is from January 23<sup>rd</sup> to March 3<sup>rd</sup>, 2018, since these data could only be collected by Matomo after the update of *grammis*.

of 3,737 total visitors). Table 7 shows that, if a user is using a modal window at all, it is a grammatical term in almost 88% of the cases whereas the links to the different word class dictionaries are used only in about 12% of the cases.

Table 7: Use of the modal windows.

Modal windows	Number of visits with modal window interaction	Hits (site impressions)	Rate of hits
“Terminologie” (Terminology)	573	729	87.94%
“Grammatische Wörterbücher” (Dictionaries of word classes)	65	100	12.06%
<b>Total</b>	<b>638</b>	<b>829</b>	<b>100.00%</b>

Looking at the direct access rates to the components of *grammis*, we can see that the component “Systematische Grammatik” (Systematic Grammar) and the component “Wissenschaftliche Terminologie” (Scientific Terminology) have similar numbers of site impressions (3,265 vs. 3,142). The rate for all page impressions of these two components was 13.24% vs. 12.74% by which they have a higher ranking than the start page of *grammis* (2,616 impressions = 10.61%). The dictionary on verbal valency also ranks high with 2,201 impressions, while the dictionary of connectors has only half as many impressions (1,116). The two remaining dictionaries are ranked comparatively low with 207 hits for the dictionary of prepositions and 163 for the dictionary of affixes.

### Research question 3: Which results (in the ranking of the results list) do the users select after a search?

As stated in Section 6.3, the analysis of the results after a search is still challenging. With Matomo in its default configuration, it is not possible to find out which position in the ranking of the results the user is selecting. Instead, it is only possible to capture that the user clicked on a link of the results list, without knowing which one it was exactly. To capture the position of that respective page, we have to configure Matomo in a way that is not trivial. Although it is possible to count the pages that have been chosen after a search, it is not possible to see what the users were searching for before they chose that exact page. The configuration of Matomo to track which keyword was searched for, which ranking position the chosen results page had, and which page exactly it was, will be done in the near future.

## 8 Conclusion and Outlook

The results for research question 1 show that the scope of our dictionaries is in need of improvement, especially the dictionary of terminology, as nearly half of the requests could not be matched with a dictionary directly. This includes multiple-word keywords in which none of the given words could be found. Nevertheless, the *grammis* search algorithm that is used on the website does always find results. This is due to the programming of the search algorithm on the database, which scans through all headlines and full texts of *grammis* and also looks for words with similar spelling, synonyms, etc., if there is no direct match with the given keyword. Still, this is not a satisfactory solution, because the users do not want random results for their search, but an answer to a specific question. The results also show that there is a high interest in terminology and object words. When it comes to object words, verbs, prepositions and conjunctions in particular are very often requested, which is why we will focus on improving those dictionaries in the future.

The results for research question 2 show that only a few users use the option of opening the modal windows while reading a full text, and that if they do they are mostly interested in the terminology links. The finding that the “Systematische Grammatik” (Systematic Grammar) and the dictionary of terminology are the most used components in *grammis* confirm that the terminology is a very important part of the system, and indicate that we should have a closer look at the interaction between these two components in future research.

Since research question 3 could not be answered due to the configuration of Matomo, we need to have a closer look into how we can collect the necessary data.

Overall, we can say that the use of Matomo, especially in combination with the configuration of our analysis tool, is very helpful in analyzing user behavior, in particular for single-word searches, while the categorization of multiple-word searches is much more complex and needs to be done manually.

For future research based on the detected user behavior, we plan to improve the search results by ranking the results in order of the users’ preferences. Furthermore, the high bounce rate and high ranking of terminology and verbs suggest that many users only want to look up either a term or the spelling or inflection of a certain word. So, whenever a user is searching for a grammatical term or a word which is contained in our dictionaries, they will be provided with info boxes next to the results sets with detailed information on the specific term or word.

We also plan to analyze those searches that have been entered in search engines but did not lead to page impressions in *grammis*, thus gaining more users for our site. Nevertheless, we still have to decide how detailed the grammatical information in *grammis* should be, because we still aim to provide an academic version of our grammar that might not be useful for every type of user (e.g. students at elementary level).

## References

- Bergenholtz, H., Johnsen, M. (2007). Log Files Can and Should Be Prepared for a Functionalistic Approach. In *Lexikos*, 17, pp. 1–20.
- Bergenholtz, H., Johnsen, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. In *Hermes-Journal of Language and Communication Studies*, 34, pp. 117–141.
- canoo.net (2018). *canoonet. Deutsche Wörterbücher und Grammatik*. Accessed at: <http://canoo.net> [2018/03/27]
- De Schryver, G., Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams, S. Vessier (eds.) *Proceedings of the 11<sup>th</sup> EURALEX International Congress, EURALEX 2004, Lorient, France*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, pp. 187–196.
- Duden (2009). *Die Grammatik. Unentbehrlich für richtiges Deutsch*. 8., überarbeitete Auflage. Hrsg. von der Dudenredaktion. Mannheim et al.: Dudenverlag.
- Duden (2005). *Die Grammatik. Unentbehrlich für richtiges Deutsch*. 7., völlig neu erarbeitete und erweiterte Auflage. Hrsg. von der Dudenredaktion. Mannheim et al.: Dudenverlag.
- Dürscheid, C. et al. (2018). *Variantengrammatik des Standarddeutschen*. Accessed at: <http://www.variantengrammatik.net> [2018/03/20] (also see <http://mediawiki.ids-mannheim.de/VarGra/index.php/Hauptseite>) [2018/03/20]
- DWDS (2017). *API (Schnittstellen zum DWDS)*. Accessed at: <https://www.dwds.de/d/api> [2017/11/29]
- grammis* (2018). *grammis*. Accessed at: <https://grammis.ids-mannheim.de> [2018/03/26]
- grammis* Konnektoren (2018). *Konnektoren*. Accessed at: <https://grammis.ids-mannheim.de/systematische-grammatik/1182> [2018/03/20].
- Helbig, G. (1992). Grammatiken und ihre Benutzer. In V. Ágel, R. Hessky (eds.) *Offene Fragen – offene Antworten in der Sprachgermanistik*. Tübingen: Niemeyer (= Reihe Germanistische Linguistik 128), pp. 135–150.
- Hennig, M., Lotzow, S. (2016). Über welche grammatischen Konzepte verfügen wir? Ein empirischer Beitrag zu Grammatikbenutzungsforschung und Transferwissenschaft. In *Deutsche Sprache*, 44, pp. 1–22.



- Hennig, M., Löber, M. (2010). Benutzung und Benutzbarkeit von Grammatiken. In *Fest-Platte für Gerd Fritz. Hg. und betreut von Iris Bons, Thomas Gloning und Dennis Kaltwasser*. Gießen 23.05.2010. URL: [http://www.festschrift-gerd-fritz.de/files/hennig\\_loeber\\_2010\\_benutzung-und-benutzbarkeit-von-grammatiken.pdf](http://www.festschrift-gerd-fritz.de/files/hennig_loeber_2010_benutzung-und-benutzbarkeit-von-grammatiken.pdf)
- Hennig, M. (2010). Plädoyer für eine Grammatikbenutzungsforschung: Anliegen, Daten, Perspektiven. In *Deutsche Sprache*, 38, pp. 19–42.
- Klein, W. P. (2004). Deskriptive statt präskriptiver Sprachwissenschaft!? In *Zeitschrift für germanistische Linguistik*, 32, pp. 376–405.
- Krasner, G. E., Pope, S. T. (1988). A description of the model-view-controller user interface paradigm in the small-talk-80 system. In *Journal of object oriented programming*, 1(3), pp. 26–49.
- Lew, R. (2015). Opportunities and limitations of user studies. In C. Tiberius, C. Müller-Spitzer (eds.) *Research into dictionary use / Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. OPAL 2/2015. DOI:10.14618/opal\_02-2015, pp. 6–16.
- Lew, R. (2011). User studies: Opportunities and limitations. In K. Akasu, U. Satoru (eds.) *ASIALEX2011 Proceedings Lexicography: Theoretical and practical perspectives*. Kyoto: Asian Association for Lexicography, pp. 7–16.
- Lorentzen, H., Theilgaard, L. (2012). Online dictionaries – how do users find them and what do they do once they have? In R. V. Fjeld, J. M. Torjusen (eds.) *Proceedings of the 15<sup>th</sup> EURALEX International Congress. 7-11 August 2012. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 654–660.
- Matomo (2017). *Matomo*. Accessed at: <https://matomo.org/> [23/11/2017]
- Matomo Tracking (2018). *Tracking HTTP API*. Accessed at: <https://developer.matomo.org/api-reference/tracking-api> [23/01/2018]
- Müller-Spitzer, C. (2016). Aufgaben und Relevanz der Wörterbuchbenutzungsforschung Mitte der 2010er Jahre. In S. Schierholz et al. *Wörterbuchforschung und Lexikographie*. Berlin, Boston: de Gruyter, pp. 275–294.
- Müller-Spitzer, C. (2014a). Methoden der Wörterbuchbenutzungsforschung. In *Lexikographica*, 30(1), pp. 112–151.
- Müller-Spitzer, C. (2014b). Empirical data on contexts of dictionary use. In C. Müller-Spitzer (ed.) *Using Online Dictionaries* (= *Lexicographica: Series Maior* 145). Berlin, Boston: de Gruyter, pp. 85–126.
- Lehmann, C. (2018). *Objektsprache und Metasprache*. Accessed at: [https://www.christianlehmann.eu/ling/epistemology/techniques/redaction/Objekt&Metasprache\\_Typographie.html](https://www.christianlehmann.eu/ling/epistemology/techniques/redaction/Objekt&Metasprache_Typographie.html) [2018/03/26].
- Olanrewaju, R. F., Islam, T., & Ali, N. (2015). An Empirical Study of the Evolution of PHP MVC Framework. In H.A. Sulaiman, M.A. Othman, M.F.I. Othman, Y.A. Rahim, & N.C. Pee (eds.) *Advanced Computer and Communication Engineering Technology*. Cham: Springer, pp. 399–410.
- Schneider, R., Schwinn, H. (2014). Hypertext, Wissensnetz und Datenbank: die Webinformationssysteme Grammis und ProGr@mm. In Institut für Deutsche Sprache (eds.) *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, pp. 337–346.
- Suchowolec, K., Lang, C., Schneider, R., & Schwinn, H. (2017). Shifting Complexity from Text to Data Model. Adding Machine-Oriented Features to a Human-Oriented Terminology Resource. In J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos & S. Hellmann (eds.) *Language, Data, and Knowledge*. Cham: Springer, pp. 203–212.
- Tarp, S. (2009). Reflections on Lexicographical User Research. In *Lexikos*, 19, pp. 275–296.
- Tiberius, C., Niestad, J. (2015). Dictionary use: A case study of the ANW Dictionary. In C. Tiberius, C. Müller-Spitzer (eds.) *Research into dictionary use / Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. OPAL 2/2015. DOI:10.14618/opal\_02-2015, pp. 28–35.
- Zifonun, G. et al. (1997). *Grammatik der Deutschen Sprache*. 3 Bände. Berlin, Boston: de Gruyter.

# Combining Quantitative and Qualitative Methods in a Study on Dictionary Use

*Sascha Wolfer<sup>1</sup>, Martina Nied Curcio<sup>2</sup>, Idalete Maria Silva Dias<sup>3</sup>,  
Carolin Müller-Spitzer<sup>1</sup>, María José Domínguez Vázquez<sup>4</sup>*

<sup>1</sup>*Institut für Deutsche Sprache*, <sup>2</sup>*Universita degli Studie Roma Tre*, <sup>3</sup>*Universidade do Minho Braga*,

<sup>4</sup>*Universidade de Santiago de Compostela*

*E-mail: wolfer@ids-mannheim.de, martina.nied@uniroma3.it, idalete@ilch.uminho.pt,  
mueller-spitzer@ids-mannheim.de, majo.dominguez@usc.es*

## Abstract

Many studies on dictionary use presuppose that users do indeed consult lexicographic resources. However, little is known about what users actually do when they try to solve language problems on their own. We present an observation study where learners of German were allowed to browse the web freely while correcting erroneous German sentences. In this paper, we are focusing on the multi-methodological approach of the study, especially the interplay between quantitative and qualitative approaches. In one example study, we will show how the analysis of verbal protocols, the correction task and the screen recordings can reveal the effects of intuition, language (learning) awareness, and determination on the accuracy of the corrections. In another example study, we will show how preconceived hypotheses about the problem at hand might hinder participants from arriving at the correct solution.

**Keywords:** research into dictionary use, observation study, language learners, quantitative and qualitative methods, online lexicographic resources

## 1 Introduction

In the past two decades, more and more studies on dictionary use have been published. Most of them have investigated what users appreciate about dictionaries, which dictionaries they use, which information they need in specific situations and whether relevant information can be accessed easily and quickly within the dictionary. The lexicographic community benefited considerably from these studies (Dziemianko 2014, Lew 2015a, 2015b, Müller-Spitzer 2014a). However, most research conducted so far presupposes that users indeed do consult lexicographic resources. In contrast, language teachers and lecturers of linguistics often have the impression that students use too few high-quality dictionaries in their everyday work. As such, a lot of studies on dictionary use might start at a point that many students may never reach when dealing with everyday language problems.

Against this background, we started an international cooperation project to collect empirical data about what students (starting with students of German who are native speakers of a Romance language) actually do when they correct language problems in their second language. With this study, we want to complement results from works on a larger scale (i.e. studies reaching much more participants but collecting less detailed data), like questionnaire studies (cf. Levy/Steel 2015, Müller-Spitzer 2014b). To do so, we carried out an observation study with learners of German who are native speakers of Romance languages combining screen recordings (to observe what participants do) and verbal protocols (to get an idea of the intuitions and motivations of participants) during a correction task.

In the present paper, we want to emphasize the multi-methodological approach we took while designing the study. We chose this multi-methodological approach primarily because the study is very exploratory in nature. This means that although we followed specific research questions (e.g. whether dictionaries are used at all, and if so, how they are used), we did not test pre-formulated hypotheses. Consequently, we used a relatively “free” experimental setup (no experimental factors that are being varied systematically) with a rather open task (improving L2 sentences).

In a way, this is a risky approach, because we had to make sure that the observations we make can be compared and cross-referenced with something if there are no explicit experimental conditions that suggest certain comparisons. This is where the multi-methodological approach comes into play. The empirical data measured by the different methods (correction task results, verbal protocols, and screen recordings) can be combined with and compared to each other to gain a more complete picture of the processes that might prove pivotal for a successful correction of errors in the participants’ L2.

Before we describe the combination of methods in more detail, we want to emphasize the interplay between quantitative and qualitative research methods that proved very beneficial during data analysis and interpretation. We pursue an iterative process, which might start with a more qualitative observation (e.g. the impression that some participants behave rather unsystematically when researching in online lexicographic resources) that needs to be translated into a measurable variable (e.g. the mean number of seconds a participant spends on a resource during research), a process commonly referred to as “operationalization”. When a variable is operationalized and all relevant measurements are extracted from the data, more questions might emerge (e.g., “Although participant A stays on resource B for a long time, he does not find the correct solution to the problem. Why is that?”). Now, one might find an answer in the verbal protocol of participant A, which calls again for a more qualitative, interpretative approach. Therefore, quantitative and qualitative approaches take turns, constantly leading closer to a better understanding of the representations and processes guiding research for a language correction task. As we have already pointed out, we do not think that studies presupposing the use of dictionaries are unnecessary. Quite the contrary: these studies are indeed very useful when compiling and optimizing lexicographic resources. One aim of our work was simply to put these studies into perspective in providing an impression of how relevant lexicographic resources really are when dealing with language problems in a language-learning context.

In the next section (2), we will describe the experimental setup and the methods we applied. We will also show briefly how the data was annotated and combined to allow for the multi-methodological analyses. In Section 3, we will introduce two example studies relying more on the qualitative side of the data and analyses. In Section 4, we will discuss and sum up the results presented in this paper. An article covering more quantitative aspects of the study is currently in preparation (Müller-Spitzer et al., in prep.).

## 2 Experimental Setup and Methods

Altogether, data from 43 participants was collected. 15 university students participated in Santiago de Compostela (Spain) and 14 people in Braga (Portugal) and Rome (Italy) respectively. All participants speak German on a CEFR level between B1 and B2. For data collection, we combined a language correction task, screen recordings of all on-screen actions, and audio recordings to prepare verbal protocols after the experiment. We handed the participants a written instruction in their native language before the experiment. Along with a detailed description of the task, the instruction contained an explicit remark that they were not graded with the study. Moreover, their

university teachers were not present in the room. We found this especially important because we wanted the participants to behave as “naturally” as possible. The instruction also contained some clues on the thinking-aloud task (see below) to make it easier for them to express their thoughts during the experiment.

One or two experimenters who did not speak the participants’ native language were present in the room at all times. One additional person who spoke the respective local language natively was also present. She or he translated questions from the participants or cues from the experimenters.

The participants worked on a standard Windows 10 desktop environment on a 15-inch notebook with German keyboard layout and a wired mouse. Google Chrome (Version 57.0) and Mozilla Firefox (Version 52.0) were available for browsing. After each participant, the browser cache and history were cleaned. We used the same notebook for all participants in each country, but set the browser language to the respective local language.

## 2.1 Correction Task

The main task of the participants was to correct 18 German sentences. Each of the sentences contained one error. The word(s) that constituted the error were highlighted in bold. The sentences were constructed according to the following two criteria. i) The error is typical of early learners of German with a Romance native language. ii) The error should not be easily resolvable by simply searching the web for the stimulus sentence or parts thereof. We tested this for each stimulus sentence in all participating countries beforehand.

We used a simple Excel spreadsheet containing the stimuli sentences in one column labelled “Satz” (Eng. sentence). Next to it was a blank column labelled “Korrektur” (Eng. correction), where the participants need to type their corrected sentence. For each participant, the sequence of sentences was shuffled, and we did this to avoid position effects (e.g., the first sentence always being more likely to be correct). By using standard office software instead of special experimental software, we aimed to situate the task in an environment the participants are well acquainted with. They were not allowed to use any built-in assistance software in Windows 10. We did not give the participants any time limit beforehand, but told them after 30 minutes that they had 15 minutes left to work on the task. After 45 minutes, they were told to finish the sentence they were currently working on and terminate the experiment after that.

## 2.2 Thinking Aloud/Verbal Protocols

While working on the corrections the participants had to “think aloud”, i.e. express their thoughts on-the-fly. This is not an easy task and some participants coped with it better than others. Whenever the participants fell silent, we gave a short cue after around 10 seconds of silence. The voice signal was captured with a high-definition external microphone. After data collection, the audio track recorded by the external microphone was spliced in as the audio track of the screen recordings. The verbal protocols were transcribed by native speakers. German translations of the verbal protocols are also available.

## 2.3 Screen Recordings

We used the screen recording software ActivePresenter to record all on-screen actions. We made sure that the screen recording software did not interfere with the task in any way (e.g. screen flickering or performance drops). As indicated above, the screen recordings were later synchronized with the audio recordings to allow for easier transcription and investigation.

## 2.4 Combining the Data

The corrections provided by the participants were annotated by two native speakers of German. Each correction was classified as “correct” (all errors have been resolved and none were introduced), “correct with errors” (all errors have been resolved but other errors have been introduced), “case of doubt” (it could not be determined without a doubt whether the answer is correct or not), “wrong” (the error was not resolved or has been replaced by another one), “not dealt with” (the participant did not attempt to correct the sentence). Initial weighted kappa (Cohen, 1968) was  $\kappa = .86$ , which is typically considered as very good agreement. This is also reflected by the fact that 712 of 816 cases (87.3 %) were labelled identically by the two annotators. All disagreements were resolved through discussion.

As indicated above (cf. Section 2.2), we transcribed the voice recordings of the participants. But these verbal protocols were not the only transcriptions that had to be created. To be able to combine all data sources into one dataset, we also had to transcribe the screen recordings (cf. Section 2.3). The recordings were split into discrete “actions” by two annotators who were trained in this procedure with a number of screen recordings. Due to the large amount of data (over 30 hours of video data had to be transcribed on a second-by-second level), the annotators then worked on different subsets of recordings. The smallest units of the transcribed screen recordings are single actions like opening a webpage, returning to Excel, typing a correction, entering a search string in a resource or a search engine, clicking a hyperlink within a resource or a search engine result list, and so on. Each action is associated with the timestamp in the respective screen recording. All other types of information are on a higher level than these single actions. Table 1 gives an impression of the organization of the dataset, but only lists a subset of the available columns (= variables) and rows (= actions).

Table 1: A subset of rows and columns of our dataset to illustrate the multi-level organization of data.  
The column “VerbalProt” holds the verbal protocols (examples follow in Section 3).  
All protocols are abbreviated in Table 1 due to space limitations.

<i>Participant</i>	<i>SentPOS</i>	<i>SentID</i>	<i>Timestamp</i>	<i>Action</i>	<i>Resource</i>	<i>VerbalProt</i>	<i>Correction</i>
B-01	1	11	41	types correction	Excel	< Text >	wrong
B-01	1	11	106	opens browser	Google	< Silence >	wrong
B-01	1	11	126	opens PONS	PONS Dictionary	< Text >	wrong
...	...	...	...	...	...	...	...
S-16	13	8	1537	clicks on hyperlink “sich verfahren”	PONS Dictionary	< Text >	correct with errors
S-16	13	8	1544	switches to Excel, types correction	Excel	< Text >	correct with errors

The different levels of the variables in the dataset can be seen in Table 1. Column “Participant” is identical for all actions from this participant, “SentPOS” (the position of the sentence in the Excel spreadsheet) and “SentID” (the unique sentence ID for the whole experiment) are always identical as long as the participant works on that sentence. Columns “Timestamp” and “Action” are on the lowest level. Column “Resource” holds the lexicographic resource, a search engine (all participants used Google exclusively) or Excel. If two actions are made while the same resource is open and on-screen, the entry is repeated in successive rows. Finally, column “Correction” holds the correctness annotation described above. This is identical for all entries of the respective sentence of the respective participant. As indicated above, the real dataset we are working with during the analyses holds further columns (e.g. information regarding search strings, error type information etc.) and 7,647 rows (= actions) altogether. But the subset of columns displayed in Table 1 suffices to illustrate the potential



cross-combinations for analyses. For example, one can look at the number of different resources that have been consulted during the work on sentences that have not been corrected properly. In the next step, the time that was spent on each of these resources can be examined (via the timestamps) and the associated verbal protocols can be investigated. All of this information can be extracted in an automated way, because the dataset is organized in this multi-level table format.

The organization of our dataset enables us to implement the research approach that we outlined in the introduction. Qualitative approaches alone fail if one does not know which cases are worth a closer look, or which groups of cases exist. Such identification of interesting cases or groups is a strength of the quantitative side of the approach. As soon as these interesting cases have been investigated in a more interpretative/qualitative manner, new hypotheses can be generated that can then be investigated quantitatively again.

### 3 Example Studies

So far, most of our remarks and explanations concerning the combination of quantitative and qualitative methods have been inevitably quite abstract, because we wanted to describe the idea of the study as a whole. In this section, we are presenting two concrete example studies to make this idea clearer. In this work we are emphasizing the qualitative research approach, but also try to show how the quantitative analyses are intertwined with the qualitative one.

#### 3.1 Time, Language (Learning) Awareness and Determination

When looking at the data, we had the impression that students with less accurate sentences spent less time with the single resource. This gave us the idea of relating the time spent on a resource with sentence correction accuracy. We found out that the average time spent on resources plays a decisive role in the correctness of the final sentence (cf. Figure 1). With 2.4 seconds, the mean difference is quite small. Note that this difference means that – on average – the time spent on each single resource is 2.4 seconds longer in sentence edits that result in a correct sentence. During the course of the experiment, this difference may well mount up to a much larger overall difference between correct and wrong sentences.

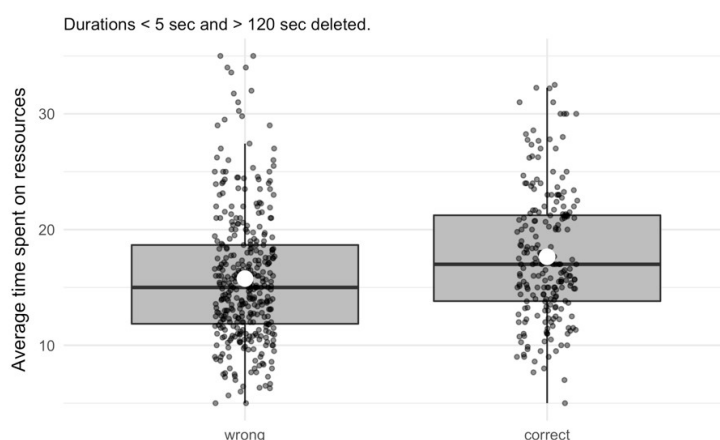


Figure 1: Boxplots for the time spent on resources for wrong and correct sentences. All times below 5 seconds have been excluded to exclude excessive switching between Excel and the browser. All times above 120 seconds have been excluded to exclude cases where participants interrupt correction of sentence A, work on sentence B (and C, D...) and then come back to the sentence A. The large white dots represent the respective mean values.

A closer look at these data indicated that the average time spent with a resource is related to frequent zapping between the resources, e. g. subject R-01 worked on sentence 1 for 3'5'', undertook 25 actions, which results in an average time of 7.65'' on a single resource (without correction time). See Table 2 for an overview of R-01's correction history.

Table 2: Timestamp, seconds per action, name of the resource and actions of participant R-01 for sentence 1 (not considering the time spent writing the correction of the sentence in the Excel file)

<i>Timestamp</i>	<i>Seconds/ Action</i>	<i>Resource</i>	<i>Action</i>
00:04:08	18	Google.it	opens Browser, opens Google, enters search string: "wenn significato"
00:04:26	2	Dicios.it	chooses the suggestion of Google, opens Dicios
00:04:28	2	Excel	opens Excel
00:04:30	10	Dicios.it	opens Dicios
00:04:40	4	Google.it	switches back to Google search results
00:04:44	3	Excel	opens Excel
00:04:47	7	Google.it	switches back to Google search results
00:04:54	3	Google.it	opens Google, looks for "leo"
00:04:57	13	LEO Dictionary	opens Leo, looks for "quando", changes the language to "it"
00:05:10	4	Excel	opens Excel
00:05:14	16	Google.it	opens new Tab, opens Google, enters search string: "costruzione con wann"
00:05:30	8	Deutsch Info	chooses the suggestion of Google, opens DeutschInfo, enters search string: "Frase condizionale con "wenn""
00:05:38	8	LEO Dictionary	opens Leo
00:05:46	12	Excel	opens Excel
00:05:58		Excel	types correction
00:06:50	6	LEO Dictionary	opens Leo
00:06:56	6	Excel	opens Excel
00:07:02	2	LEO Dictionary	opens Leo
00:07:04	3	Deutsch Info	opens DeutschInfo, enters search string: "Frase condizionale con "wenn""
00:07:07	17	Google.it	switches to the results of Google
00:07:24		Deutsches Institut Florenz.it	chooses the suggestion of Google, opens DeutschesInstitut, enters search string: "Congiunzioni: l'uso di "als" e "wenn" (quando)"
00:08:00	9	Excel	opens Excel
00:08:09	18	Deutsches Institut Florenz.it	opens DeutschesInstitut, enters search string: "Congiunzioni: l'uso di "als" e "wenn" (quando)"

Other subjects spent more time with the individual resources and achieved better results. Given this observation, we investigated the following factors: a) number of actions, b) number of search strings, c) average time spent on resources d) idea before the search<sup>1</sup> and e) accuracy of the sentences. The quantitative analyses suggest that the less time the students spent with a resource and the faster they switched between them, the more difficult it became for them to gain a clear idea

<sup>1</sup> This variable was coded by the annotators who transformed the screen recordings into the structured dataset (cf. Table 1) and is set to "true" if the participants expressed an initial idea of the problem before referring to Google or a specific resource.

of how to solve the problem. An exception to this rule is subject R-07, who undertook few actions but spent a relatively short time on the resources, namely an average of 14'6". The opposite of the general statement above would mean: The more time a participant spent on a resource and the less (s)he zapped between the resources, the better was the intuition before the search and the more sentences were correct. This statement cannot be confirmed consistently by the data; subjects with frequently correct sentences spent relatively more time on the resources (> 20') but did not necessarily have an intuition before the search for all their accurate sentences. For example, subjects R-08 and R-14 were able to correct several sentences even without any idea before the search. It can be deduced that, in addition to the average time spent with the resource and the intuition before the search, there must be other factors that predict sentence correctness. The concrete questions that arose were: What are the additional factors that play a role, and why are these factors decisive for a successful search and the correctness of the sentence? These questions cannot be answered by purely quantitative analyses. It was at exactly this point when verbal protocols came into play. This will be illustrated by examples from participant R-07<sup>2</sup>. Example 1 shows the verbal protocol for the sentence *Obwohl sich der Junge beeilt hat, hat er die U-Bahn **verloren*** (Eng. "Although the boy hurried, he missed the subway"). R-07 was aware of the polysemy of the Italian verb *perdere* (Eng. to lose, to miss) (line 8-9), which means that she had already developed a certain language awareness; she knew that in combination with a vehicle like *U-Bahn* (Eng. "subway") the German verb *verlieren* (Eng. "to lose") was not correct and that a specific verb has to be used instead (line 11-13). She was aware of words belonging together (collocations) and she consequently searched for a specific word in the resources (cf. also example 3). This is how she avoided a word-by-word translation (*perdere* – *verlieren*), which, in our stimulus sentences, usually leads to interference errors. In addition, she knew various resources and opened an appropriate resource related to the search query. In order to find out the meaning of *verloren*, she opened PONS (lines 3-4); for the conjugation of *verpassen* she opened Reverso (lines 12-14). As we can see in example 1, she used linguistic strategies: in her search, she used a synonym like *Zug* (Eng. "train") for *U-Bahn*, which she thought was more prototypical, and synonyms for *verlieren* (line 10, 12).

(1)

- 1 So I read the first sentence [she reads the sentence] (*Obwohl sich der Junge beeilt hat, hat er die U-Bahn verloren*)
- 2 eh the section in bold is *verloren* so the verb *ehm* there are two sentences therefore one is the main clause and one
- 3 is the subordinate clause *ehm* so first I look for the verb *verloren* I prefer google chrome # *ehm* # generally I use the
- 4 online dictionary PONS because perhaps it shows also the context and the use of a word and also some examples
- 5 so therefore I search from German to Italian okay so *verloren* ## which means # okay I see it's the past participle
- 6 of the verb *verlieren* so lost the past participle '*perduto*' lost ok # so I lost the train even though the boy (*hat beeilt*)
- 7 I look it up because I'm not sure what does it mean so # *beeilt* (*affrettarsi*) okay (*sich beeilen*) so even though the
- 8 boy hurried he lost the train #perhaps the error would be that *verloren* is used in other contexts so I look for other
- 9 use contexts or a synonym of the verb *verlieren* so # I look up *verlieren* and it shows me (*perdere perdere la testa* #
- 10 *dispandersi*) okay In Italian I look for a synonym of the verb *perdere* in German so I set Italian German and look for
- 11 *perdere* okay so (*verlieren verlegen smarrire*) # *eh perdere il treno* for example it shows (*verpassen*) which means
- 12 that I can use *verpassen* instead of *verlieren* and *ehm* okay it shows me (*Zug*) so okay I use *verpassen* I will look
- 13 for the past participle for being sure so I open a website with the name REVERSO ## so conjugation German verbs
- 14 I search on google and it gives me REVERSO so I look for *verpassen* # so (*verpasst*) okay I write the sentence on
- 15 the right again *obwohl sich der Junge beeilt hat hat er die U-Bahn verpasst* [she finishes the sentence] okay I go
- 16 continue with the second sentence

2 The transcription symbols are based on the *Lessico di frequenza dell'italiano parlato* (De Mauro et al. 1993): # = short break; ## = longer break; <?> word is not comprehensible; sotto<categoria> = the word was interrupted, but the reconstruction was possible; [sie liest den Satz laut vor] = extra-linguistic comment, (*hat beeilt*) = student reads a word, expression or sentence found in the resource. The original verbal protocol is in Italian. For obvious reason we have provided the translation in English.

The participant showed good metalinguistic knowledge of German and a high level of language awareness throughout the reflection process. At the same time, the subject read the grammatical annotations in the resource carefully and took them into account when finding the solution, as can be seen in example (2). The interplay of all these factors resulting from the verbal protocols might explain the longer time spent on the resources:

(2)

1 [She reads the sentence] (*An unserem Forschungsinstitut ist Ihnen unsere Bibliothek 24 Stunden zur Verfügung*)  
 2 *Verfügung* okay so *nel nostro istituto* I don't know the word so later I look for it # then *la biblioteca 24 ore*  
 3 *Stunden 24 ore* of 24 I think ok so first I look up ah *Forschungsinstitut* because I don't know the meaning so ##  
 4 *Forschungsinstitut* # okay there is no result ok so I look only for *Forschung* and there is *indagine* so I think in  
 5 our research institute *ehm* # I think that I need to say perhaps *c'è* so perhaps *gibt es* [?] I don't know so I look up  
 6 *Verfügung* ## and it shows *disposizione* so it means it is available and there is written (*jemandem zur Verfügung*  
 7 *stehen*) so instead of *ist* I might use *stehen* because it means *essere a disposizione* so I put in # but it is also wrong  
 8 (*zur Verfügung*) # *ehm jemandem essere a disposizione di qualcuno* so I write *an unserem Forschungsinstitut ehm*  
 9 *steht ihnen* because there was *jemandem* which means dative yes *ihnen unsere Bibliothek 24 Stunden* and I don't  
 10 change mmm # *zur Verfügung* [she completes the correction of the sentence] okay

R-07 connected language competence, attention, metalinguistic reflection, language awareness and dictionary use awareness to arrive at a good correction of the sentence. It is also very interesting that she often double-checked her correction proposals (example 3, lines 5-6), i.e. she changed the search direction and checked her hypothesis, although she was quite sure of the solution. This proficient use of strategies was also responsible for the high number of correct sentences of this participant.

(3)

1 [She reads the sentence] (*Wenn ich zur Schule ging habe ich viel Sport gemacht*) so when I went to school *ehm* I  
 2 did a lot of sport I did a lot of sport in this case it's wrong *wenn ehm* # because I think that eh you have to use  
 3 *als* instead of *wenn* but I still try to find out if it gives me a few examples always some context of use okay so it is  
 4 also used as conditional but it is not in this case in this case it is a temporal clause I believe yes because it is used in  
 5 the past so every time when it is used when an action of the past is repeated often so I look up the sentence *quando*  
 6 *andavo a scuola* and I have a look if it's used also from Italian # to German

Finally, the determination not to give up and to find a solution seems to be another decisive factor (cf. example 4). Other participants also had good metalinguistic knowledge, had opened a useful resource, had read the information given in the resource attentively and were close to the solution. However, they sometimes lacked the determination not to give up and to find a satisfactory solution. For example, after several unsuccessful searches (lines 1-4) and a certain insecurity (line 4), student R-14 gave up and went on to the next sentence (line 5):

(4)

1 Okay no # I can't find the solution so I usually solve it by looking for the sentence on Google and see if it is used  
 2 but in this case it doesn't give me examples I can't find anything # I try to check again on LEO dictionary if there's  
 3 something else mmm no okay I don't find don't think it is (*tauschen*) but there is (*umtauschen*) or (*gegen etwas*  
 4 *tauschen*) mmm # I'm not sure so I don't know the use of these verbs I try to search on the internet on Google okay  
 5 I can't correct it so I go ahead

As we have shown, the verbal protocols bring to light certain behavioral patterns and the reasoning of the participants. In our case, it could be shown that not only the average time with the resources

and an intuition before the search are responsible for a successful correction of the sentences, but also language competence, language awareness, and the (correct) use of strategies. Careful reading, paying attention to metalinguistic annotations in the resource and the determination to find the solution also play a fundamental role – factors that would not have come to light through a purely quantitative analysis. How much these factors interact or whether they might be arranged hierarchically or operationalized in a quantitative way remains to be explored.

### 3.2 Intuition and Focalization During Research

While analyzing the verbal protocols and the screen recordings, we observed that in some cases intuition and hypotheses formulated at the beginning of the correction task seem to influence participants' search behavior. The students initiate a search process to validate their hypothesis and miss relevant information found in the online resources. In what follows, we will give a detailed qualitative account of participants' search behavior when they show such focalization behavior. We will propose a schema based on the search patterns observed. As will be seen below, the verbal protocols and screen recordings play a crucial role in helping us to understand what may lead students to exhibit specific search behaviors.

We begin by tracing the search actions by a Portuguese participant while trying to correct the following stimulus sentence: *Er wohnt seit Jahren in Berlin und trotzdem verliert er sich immer noch* ("He has been living in Berlin for years and still gets lost"). We expected the participants (i) to identify that although the common polysemous verb *verlieren* means "to lose", the verb *sich verlieren* is not allowed in this context; (ii) to search for the correct reflexive verbs that fit the above context: *sich verlaufen*, *sich verfahren* or *sich verirren*: "*Er wohnt seit Jahren in Berlin und trotzdem verfährt/verläuft/verirrt er sich immer noch.*" From the verbal protocol it is clear that the student was not sure about the meaning of the verb *verlieren*. This led him/her to search the form "*verliert*" in Google Translate. The Google Translate result is the Portuguese verb "*perde*" (*perder* - "to lose"). The verbal protocol shows that the student correctly inferred that the equivalent Portuguese verb is the reflexive form *perde-se* (*perder-se*): "*# ahhh ele mora anos la e mesmo assim perde-se # / # ahhh he's been living there for years and still he gets lost*" (excerpt from the verbal protocol). All further search actions were directed to confirming that the correct verb in the given context is *verlieren* and that it is reflexive. Following this assumption, the student entered "*verliert sich*" in Google Translate and obtained the result "*perdido*", the Portuguese verb without the reflexive pronoun. According to the verbal protocol, the student expects Google Translate to output the Portuguese reflexive verb "*perder-se*". Since this is not the case, (s)he entered more context taken from the stimulus sentence in Google Translate "*verliert er sich immer noch*" hoping to obtain the reflexive form of the verb. Once again, the Google Translate result does not contain the reflexive form of the verb: "*ele ainda perdeu*". According to the verbal protocol, the student begins to question whether the reflexive pronoun is needed in the German stimulus sentence: "*# I think that sich is not needed here #*".

As the Portuguese equivalent *perde-se* does not appear in Google Translate and this does not correspond to the student's expectation, (s)he changed the language direction in Google Translate from Portuguese to German and added the reflexive pronoun to the Portuguese sentence: "*ele ainda se perde*". Google Translate outputs a German translation without the reflexive pronoun: "*er verliert noch*". As can be seen in the verbal protocol, this led the student to conclude that the German verb *verlieren* is not reflexive in the given context: "*# exactly I think it doesn't need sich #*".

The qualitative analysis of students' search behavior via the examination of the verbal protocols and the screen recordings has proven to be an important method for identifying search patterns common to a specific participant group and across participant groups. Regarding the German stimulus



sentence above, we observed that eight out of the 10 Portuguese participants assumed – based on intuition before initiating the search process or on hypotheses resulting from search actions – that the verb *verlieren* or *sich verlieren* is correct in this context. From this point onwards, these participants arrived at one of two hypotheses: (i) the problem lies in the reflexive pronoun – *verlieren* does not take a reflexive pronoun; (ii) the problem lies in the word order of the verb *verlieren* and the reflexive pronoun – “*verliert sich*” or “*sich verliert*”. With all further search actions, they tried to confirm or validate their respective hypothesis. Regardless of the resource used (Linguee, PONS Dictionary, Google Translate), search actions that take “*verlieren*” or “*sich verlieren*” as a starting point lead to unsuccessful results.

These qualitative observations provided us with enough evidence to formulate a focalization hypothesis search pattern that is illustrated in Figure 2.

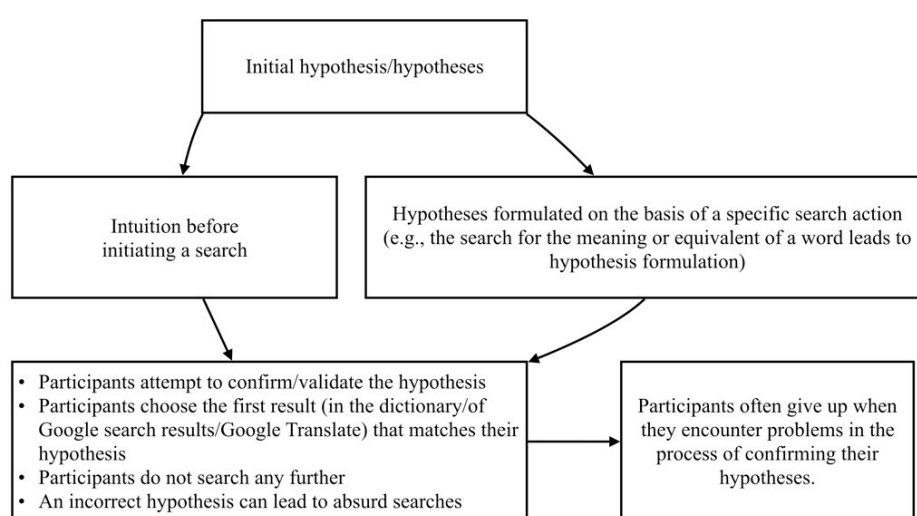


Figure 2: Schema of the focalization hypothesis search pattern

The students start off by formulating an initial hypothesis, based either on intuition before initiating a search process or on hypotheses formulated on the basis of a specific search action, such as the search for the meaning or translation equivalent of a word. The entire search process thereafter focuses on the attempt to confirm this hypothesis. The observational data seems to indicate that many students generally choose the first result they find in the resources that matches their hypothesis and do not search any further. As seen in the above example, an incorrect initial hypothesis more often than not leads to absurd search actions and results. We also observed that participants who encounter problems confirming their hypothesis usually give up on trying to correct the stimulus sentence.

In order to demonstrate the focalization hypothesis discussed above, we will provide a second example taken from the Portuguese observational data. In this example, students were asked to correct the German stimulus sentence *An unserem Forschungsinstitut ist Ihnen unsere Bibliothek 24 Stunden zur Verfügung* (“At our research center our library is at your disposal 24 hours a day”). Correcting the sentence involves identifying that the verb *ist* (Infinitive: *sein*) must be replaced by the verb *steht* (infinitive: *stehen*) in combination with *zur Verfügung* in this context. In other words, it is expected that students identify the function verb construction (German *Funktionsverbgefüge*) *zur Verfügung stehen*.

The student in question began by formulating the following hypothesis taken from the verbal protocol: “here the use of *ist* is not correct with *Verfügung* # I think it should be *gibt* instead of *ist* #”. Taking this hypothesis as a starting point, the student entered “*gibt zur Verfügung*” in the Google

search engine. The search engine outputs “zur Verfügung gibt” as the first link on the Google results page that in turn refers the user to the Linguee Dictionary. The student was so focused on validating her/his hypothesis formulated at the beginning of the search process that the simple fact that the expression “zur Verfügung gibt” appears on the results page suffices for the student to come to the conclusion: “exactly it is *gibt* that should be used because it is <...> correct #”. The search process ends at this point. The student does not select the link in the search result to access the information found in the Linguee Dictionary and does not search any further. Here, again, the exclusive focus on the initial hypothesis rules out the possibility for the student to arrive at other conclusions and following other search paths.

Although the above examples have been taken from the Portuguese participant group, the focalization hypothesis has also been observed in Italian and Spanish speaking participants. The qualitative analysis of students’ search behavior has allowed us to pick up on the focalization hypothesis. In future work we intend to complement the qualitative findings with quantitative methods in order to be able to compare the datasets in a systematic manner. The combination of qualitative and quantitative material will provide us with a more comprehensive insight into students’ search behavior.

## 4 Discussion and Summary

Given the two example studies above, we can draw a few tentative conclusions concerning the behavior of language learners when they have to resolve language problems in their L2. Generally speaking, the time that is spent consulting a resource pays off. The longer our participants stay on resources, the more likely they are to arrive at a correct final sentence. If more factors resulting from the verbal protocols are taken into consideration, the picture becomes more complicated but also clearer: thoroughness has to be accompanied by cross-checking preliminary conclusions (even when you are already quite sure about a solution), good meta-linguistic knowledge, and a strong determination to arrive at a good correction. The latter two factors are very hard to operationalize on a quantitative level. Hence, we presented evidence from the verbal protocols that allows us to infer these factors from the verbalizations of the participants. Another observation we presented in Section 3.2 is that many L2 learners start their research with a strong hypothesis in mind that guides their whole research process. This focalization, as we have called it, can be so strong that participants even ignore information that is readily available in the resource they are consulting. Alternatively, if the hypothesis guiding the search cannot be confirmed, some participants give up searching for a solution altogether.

Coming back to the main topic of this paper, we want to comment on some of the strengths of quantitative and qualitative approaches and connect those to our data. In the first example study, we started from the general quantitative observation that more time on resources leads to better results. This is a general pattern in our sample (or even larger groups of language learners) that can only be observed when the variables are clearly operationalized and analyzed by inferential statistics. Such pattern extraction and generalization on larger groups is a clear advantage of the quantitative approach. However, the qualitative approach allows us to complement the analysis with explanations for cases that do not fit the general pattern. By looking at individual search histories and the accompanying verbalizations, we can come to a more detailed understanding of the processes that help language learners to get to good solutions (or prevent them from getting to them). From there, we are able to generate further hypotheses that can be tested with our data in a quantitative way.

Finally, we want to stress that one of the main research questions of our study (What do language learners really do when they are solving language problems on their own?) can be answered in a way that is very encouraging for lexicography. In the vast majority of all sentence corrections (78.1%),

our participants used an online dictionary of some sort. Automatic translators (like Google Translate or the PONS Translator) were also widely used, but only in 21.9% of all sentence corrections. Given our sample, this suggests that language learners still rely heavily on lexicographic resources – at least on the web – even when they are allowed to use any resource they want.

## References

- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. In *Psychological Bulletin*, 70(4), pp. 213–220.
- De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*. Etas, Milano.
- Dziemianko, A. (2012). On the use(fulness) of paper and electronic dictionaries. In: S. Granger, M. Paquot (eds.) *Electronic Lexicography* (pp. 319-342). Oxford: Oxford University Press.
- Levy, M. & Steel, C. (2015). Language Learner Perspectives on the Functionality and Use of Electronic Language Dictionaries. In *ReCALL* 27(2), pp. 177-196.
- Lew, R. (2015a). Opportunities and limitations of user studies. In *OPAL – Online publizierte Arbeiten zur Linguistik*, 2/2015, pp. 6-16.
- Lew, R. (2015b). Research into the use of online dictionaries. In *International Journal of Lexicography*, 28(2), pp. 232-253.
- Müller-Spitzer, C. (ed.) (2014a). *Using Online Dictionaries*. Berlin/New York: de Gruyter. (Lexicographica: Series Maior 145).
- Müller-Spitzer, C. (2014b). Empirical data on contexts of dictionary use. In: C. Müller-Spitzer (ed.) *Using Online Dictionaries* (pp. 85-126).
- Müller-Spitzer, C., Domínguez Vázquez, M. J., Nied Curcio, M., Silva Dias, I., & Wolfer, S. (in prep.). The right hypotheses and careful reading make the difference: Results of an observation study with language learners on using language resources online.

## Acknowledgements

We want to thank the Institute for the German Language (IDS) in Mannheim, Germany, for supporting our study with funds from the institution's core budget. We also want to thank all the assistants that helped transcribing the data and supporting us with their native language skills during data acquisition.

# **Dictionary-making Process**





# Nathanaël Duez lexicographe : l'art de (re)travailler les sources

**Antonella Amatuzzi**

*Università degli Studi di Torino*

*E-mail: antonella.amatuzzi@unito.it*

## Abstract

La production lexicographique de Nathanaël Duez (1609-1660), maître de langues actif à Leyde, aux Pays Bas, comprend une édition de la *Janua linguarum reserata* de Comenius, comportant les versions française, italienne, allemande et latine, la *Nova nomenclatura quatuor linguarum, gallico, germanico, italico et latino idiomate conscripta*, le *Dictionnaire françois-alleman-latin et aleman-françois-latin* et le *Dittionario italiano e francese Dictionnaire italien et François*.

L'objectif du présent travail est de l'analyser dans son évolution (elle commence par un répertoire plurilingue pour terminer avec un véritable dictionnaire, riche d'informations et soigneusement construit) pour mettre en évidence la démarche que Duez suit pour la réalisation de ses ouvrages (notamment la sélection et le remaniement de sources préexistantes). L'étude de l'intertextualité montre qu'il intervient de plus en plus sensiblement pour créer des outils pédagogiques qui répondent aux besoins de ses élèves, clairs et facilement consultables. L'apport de Duez à l'histoire de la lexicographie devrait être réévalué.

**Keywords:** lexicographie historique, français, italien, allemande, latin, Duez, Pays Bas

## 1 Introduction

Entre le XVI<sup>e</sup> et le XVII<sup>e</sup> siècle, la lexicographie des langues européennes se développe et avance selon des lignes bien connues : les glossaires et les vocabulaires multilingues laissent place à de véritables dictionnaires, plus riches d'informations et mieux soigneusement construits (voir, comme référence générale Hausman 1989-1991). Or, le parcours que Nathanaël Duez accomplit en tant que lexicographe subit une évolution semblable et est en cela emblématique du progrès de cette discipline.

Il est l'auteur de quatre ouvrages lexicographiques et de plusieurs manuels de langue concernant le latin, le français, l'allemand et l'italien, qui eurent un succès remarquable en plusieurs pays d'Europe<sup>1</sup> mais qui demeurent cependant peu étudiés dans leur ensemble et sous-estimés, sans doute éclipsés par le rayonnement de quelques textes phare de la lexicographie qui ont davantage attiré l'attention des spécialistes.<sup>2</sup>

L'objectif du présent travail est d'analyser la production lexicographique de Duez pour mieux comprendre les relations que ses différents ouvrages entretiennent entre eux et avec d'autres écrits, afin de mettre en relief l'apport la contribution personnelle et originale de Duez à l'histoire de la lexicographie plurilingue (italien, français, allemand en particulier). Quelles sont donc les étapes de son chemin ? Sur le plan de l'intertextualité, comment intervient-il pour transformer ses sources ? Quelles sont les motivations qui le guident dans ses réélaborations ?

1 Pour une bibliographie complète voir Loonen 1995 et Jones 2000 : 298-328 ; pour l'œuvre de grammairien voir Amatuzzi (2010), Mattarucco (2003), Szoc (2009).

2 Nous pensons, entre autre, aux *Recherches italiennes et françaises* d'Antoine Oudin, voir infra, par. 3.1.

## 2 Informations biographiques

Nathanaël Duez (1609-1660) naquit à Altwiller, village d'Alsace fondé par des huguenots exilés. Son père, Samuel d'Huet, était ministre de culte protestant et l'initia au latin dès son plus jeune âge. Après avoir voyagé en Allemagne, en Italie et Angleterre et avoir vécu à Paris, en 1639 Nathanaël s'établit à Leyde où il exerça la profession de *magister linguae gallicae et italicae* jusqu'à sa mort.

Le contexte dans lequel il opère est donc celui des Pays Bas, carrefour de cultures et de langues diverses. Dans ces territoires, où plusieurs réfugiés protestants, provenant essentiellement de France et d'Allemagne, s'étaient installés, le français s'était peu à peu diffusé et avait pris un enracinement socio-culturel et religieux considérable. Il avait d'abord eu la fonction de langue de culte puis, avec l'essor des échanges commerciaux, des contacts entre personnes et des voyages, il s'affirma comme langue de communication internationale et diplomatique, et il devint la langue de distinction, qui contribuait à l'ascension sociale et à la réussite professionnelle. Il était enseigné par de nombreux maîtres de langues généralement issus du milieu des réfugiés "immigrés" de France, qui gagnaient leur vie en mettant à profit leurs compétences langagières dans le réseau d'écoles françaises qui avaient été fondées pour former les jeunes ou comme professeurs pour un public adulte (voir Dodde 1997 ; Loonen 2000 ; Van Strien-Chardonneau et Kok Escalle 2010).

Duez compte parmi ces professeurs (voir Loonen 1993) et, comme certains de ces collègues, il conçut des ouvrages pédagogiques pour ses élèves, principalement des expatriés allemands de famille aisée qui apprenaient le français ou des bourgeois francophones à qui la maîtrise de l'italien était nécessaire pour commercer ou voyager dans la péninsule.

Il collabora pour cette activité d'auteur avec les Elzevier, célèbre famille de typographes-imprimeurs, chez lesquels il commença à publier ses travaux.

## 3 La production lexicographique : évolution et intertextualité

### 3.1 *Janua aurea reserata quatuor linguarum*, Leyde, Elzevier, 1640

Le premier travail de Duez dans le domaine de la lexicographie est une édition de la *Janua linguarum reserata* de Comenius,<sup>3</sup> datant de 1640. Elle comporte les versions allemande, italienne (première traduction disponible) et française - en plus que latine - de ce texte pédagogique qui, en 1000 paragraphes numérotés en 100 "titres", fournissait des renseignements de base relativement à des sujets spécifiques et variés (par exemple : "De igne", "De fructibus", "De ulceribus et vulneribus", "De panificio", "De puerperio", "De musica", etc.) en donnant en conséquence la terminologie usuelle.

La triple traduction que propose Duez du texte latin est diligente et ponctuelle. Il ne se contente pas d'offrir des mots avec leurs traduisants en d'autres langues mais il intervient souvent avec des annotations dans les marges, l'astérisque signalant l'endroit où Duez entend les insérer. Elles ajoutent des précisions :

- (1) La partie devant du col c'est la gorge ou le gosier\*, celle de derrière le chignon. \*Où les hommes ont ordinairement cōme un os eslevé que l'on appelle le morceau d'Adam. (par. 252)
- (2) Le coste cominciata dalle ascelle si forniscono negli ipocondri\*. \*cioè quella tela che stà intorno al cuore. (par. 254)

3 La première édition de ce manuel scolaire conçu par le philosophe, grammairien et pédagogue morave, parut en Pologne en 1631 (Dantisci sumptibus et Typis Georgi Rheti).

ou elles donnent des indications sur la prononciation – de l’italien surtout - en mettant en garde contre de possibles fautes :

- (3) La calamita\* si volta dritta ó drittamente verso la tramontana. \*Auertisci ben qui l’accento nella penultima, perche [sic] hauendolo nell’ultima ; cosi calamita uorrà dire calamitas in Latino. (par. 88)

ou encore elles complètent avec des proverbes ou de la phraséologie :

- (4) L’asino & asinello chinato ó piegato, ferito dal bastone ó randello del mulattiere, ragghia\*. \*Raglio d’asino non va in cielo c. le preghiere dei tristi. (par. 180)

### 3.2 Nova nomenclatura quatuor linguarum, gallico, germanico, italico et latino idiomate conscripta, Leyde, Elzevier, 1640

La même année 1640 Duez publie une *Nova nomenclatura quatuor linguarum, gallico, germanico, italico et latino idiomate conscripta*. Dans l’“Avis au Lecteur” il cite son travail précédent en soutenant que l’entreprise de produire un ouvrage semblable à celui de Comenius serait “quod est in proverbio, Iliade post Homerum” et que dans sa *Nomenclatura* on trouvera beaucoup de mots “quae in janua neutiquam reperiuntur”.

Or, il ne fait pas de doute que le contenu lexical de l’œuvre de Comenius est sensiblement accru et que la structure de la *Nomenclatura* est différente : elle a la forme de répertoire quadrilingue, organisé en 26 chapitres, selon un classement thématique moins fragmentaire et plus fonctionnel que celui de la *Janua*. Les lemmes sont ordonnés selon une disposition qui semble répondre à des exigences didactiques : les champs lexicaux pris en considération correspondent à des situations communicatives authentiques dans lesquelles l’apprenant peut se retrouver et concernent majoritairement des objets concrets mais aussi des concepts plus abstraits<sup>4</sup>.

La *Janua* dut tout de même exercer une certaine influence sur la *Nomenclatura*. La comparaison des paragraphes traitant “De domo ejuque partibus, De la maison et de ses parties” (titre XLIX, par. 540-548) avec le chapitre XII, intitulé “De la maison” donne la mesure tout à la fois des points de contact et de la distance entre les deux ouvrages.

Sur le plan quantitatif, Duez multiplie le nombre de mots recensés.<sup>5</sup> Il inclut notamment des mots techniques utiles dans des contextes professionnels et plus rarement de la phraséologie, ce qui confirme, au moins en partie, ce qu’a affirmé (sans toutefois l’argumenter avec des exemples) Loonen (1994, s.p.) : “if we take a close look at the words selected for inclusion, we find the choices in the *Nomenclatura* to be more idiomatic and practical than those in the *Janua*, which tend to be rather stilted and uncommon”.

Pour ce qui est de sa manière de procéder, Duez semble parfois suivre de très près la *Janua*, en transformant les termes clé des différents paragraphes en mots vedette. C’est le cas par exemple de la partie relative au toit de la maison qui reproduit quasiment la même succession du texte de Comenius :

4 Voici le titres des 26 chapitres : I Des choses theologicques, II Du ciel & des elements, III Du boire & manger, IV Des habits, V Des estoffes, VI Des couleurs, VII De la marchandise, VIII Des nombres, IX Des poids & mesures, X Des drogues, XI De l’argent de France, XII De la maison, XIII Des meubles, XIV De l’homme, XV Des noms des hommes & des femmes, XVI Des noms des pays & nations, XVII Des noms des villes, XVIII Des noms des fleuves, XIX De divers estats, XX De l’estude, XXI Des jeux, XXII Du manège, XXIII De la chasse, XXIV De l’escrime, XXV De la guerre, XXVI Des amandes & supplices.

5 Une approximation effectuée sur la base de la longueur des chapitres (deux doubles pages pour la *Janua* - qui comporte sur les pages paires deux colonnes avec les textes latin et allemand et sur les pages impaires, en regard, les textes français et italien – et sept pages et demie pour la *Nomenclatura*) indique pratiquement un redoublement.

<i>Janua</i> , par. 545	<i>Nova Nomenclatura</i> édition 1652 pp. 76-77
<p>Tectum columnibus incumbit ; tignis &amp; tigellis tegulæ &amp; imbrices, scandulæ ac ardosiæ : culmen stramineum est vel lateritium.</p> <p>Das dach liget auff den balkeb oder tragn ; die ziegel und tahlziegel, schindeln und schiefferstein auff den sparren und latten : der gipfel oder forst ist strohern oder von ziegein.</p> <p>Le toit est posé et appuyé sur des chevrons, solives ou soliveaux ; les tuiles plates &amp; creuses, les esselins et les ardoises sur des lattes ou barreaux : le faiste, haut, sommet ou coupeau est de paille ou de briques.</p> <p>Il tetto giace su colonne ò sostentamenti ; le tegole e gli embrici, le gattinelle ò tavolette e le scaglie su lambrecchie sbarre o traucelli : la sommità ò cima è di strame ò di mattoni.</p>	<ul style="list-style-type: none"> <li>- Le toit ; das tach ; il tetto ; tectum</li> <li>- Le faiste ; die fürst ; la cima o vetta ; fastigium</li> <li>- Les chevrons ; die sparren oder balken ; li travetti ò travicelli ; tigna &amp; tigilla</li> <li>- Vn Pannonceau, une giroüette ; wetterhahn, tache fähulein ; una bandirouola ; coronis, idis, &amp; ventorum pinnula</li> <li>- La faistiere ou gouttiere ; die rinne oder tachrinne ; la dozza overo il canale che gitta l'acqua dal tetto ; colliciaë vel colliquiaë, canalis, stillicidium</li> <li>- Vn sommier ; ein drohm ; grosso traue ; trabs perpetua</li> <li>- Vne poutre ; ein balk ; una traue ; trabs</li> <li>- Les soliveaux ; die zwerchbaklein, so auff der drohm ligen, umb die bretter eines boden darauff zu legen ; le travicelle ; lacunaria, trabeculæ, tigilla lacunaria</li> <li>- Vne latte ; ein latte ; lambrecchia, lata, tauolette ; transversaria regula, sudes lateraria, templa, orum</li> <li>- Latter ; latten ; guernir di latte, lambrecchiare ; tectum templis instruere</li> <li>- Vne ardoise ; ein schieferstein ; scaglia, ardesia ; lithostilbe, lapis scissilis</li> <li>- Vne tuile ; ein ziegel ; tegola ; tegula</li> <li>- Des esselins ; schindeln ; gattinelle ; scandulæ</li> <li>- Vne brique ; ein gebaken stein ; un mattone ò quadrello ; later</li> </ul>

En général les échos de la *Janua* sont peu reconnaissables, étouffés dans une gamme d'autres documents difficilement discernables par l'exercice d'amplification que Duez effectue.

Dans le chapitre consacré à la chasse, par exemple, Duez greffe une longue digression (6 pages), uniquement en français, à propos du comportement de certains animaux dont nous avons pu identifier au moins une source : la *Vénerie* de Jacques de Fouilloux (première édition : Poitiers, Marnef et Bouchet frères, 1561).

Voici les passages des deux textes mis en relation, d'où il ressort que Duez suit exactement, mais sans jamais le citer, ce livre dans lequel l'auteur, passionné de chasse, avait recueilli des observations sur les animaux et leurs habitudes.

<p><i>Nova Nomenclatura</i> chapitre XXIII, p. 196 (édition Leyde, Elzevier, 1652)</p> <p>Pour faire venir en chaleur une lice prenez deux testes d'aulx, &amp; un demy couillon de castor, avec une douzaine de cantarides, faites bouillir le tout ensemble, en un pot d'une peinte, avec de la chair de mouton, &amp; en faites boire deux ou trois fois en potage à une lice, elle ne faudra jamais de venir en chaleur.</p> <p>Puis vous lairrez passer le plein decours de la lune, pour la faire couvrir sous les signes des bessons ou gemeaux &amp; du verseau. Autant en peut-on faire du masle, pour l'eschauffer.</p>	<p><i>Vénérerie</i>, p. 6r-v (édition Paris, Claude Cramoisy, 1628)</p> <p>Si vous voulez auoir de beaux chiés, il faut auoir vne belle Lyce, qui foit de bonne race, forte &amp; proportionnee de ses membres, ayás les costez &amp; les flancs grâs &amp; larges, laquelle pourrez &amp; faire venir en chaleur en ceste maniere. Prenez deux testes d'aulx, &amp; vn demy couillon d'une beste qui se nôme castor, avec du ius de cresson alenois, &amp; vne douzaine de mouches cantharides, &amp; faites bouillir le tout ensemble en vn pot tenant vne pinte, avec de la chair de mouton &amp; en faites boire par deux ou trois fois en potage à la Lyce, elle ne faudra iamais de venir en chaleur. Et autant en peut-on faite au Chien pour le rechauffer. Puis quand vous verrez que la Lice sara chaude, attendez le plein de cours de la Lune à passer, pour la faire couvrir : &amp; la faites emplir souz les signes de <i>Gemini</i> &amp; <i>Aquarius</i>, car les Chiens qui naistront en ce temps ne serôt si suiets à la rage.</p>
---	---

### 3.3 Dictionnaire françois-alleman-latin et aleman-françois-latin, Leyde, Hegher 1642

Avec le *Dictionnaire françois-alleman-latin et alleman-françois-latin* (que nous abrégeons *D FAL AFL*) Duez franchit une étape ultérieure dans sa production de lexicographe. Il ne s'agit plus d'un glossaire mais d'un vrai dictionnaire : le corpus lexical est réélaboré et organisé en ordre alphabétique, avec les entrées en français dans la première partie et en allemand dans la deuxième, accompagnées dans les deux cas du latin.

Il est évident, comme l'a montré, Von Gemmingen (1999), que Duez s'inscrit dans le sillage du *Dictionnaire François-Allemand & Allemand-François* de Levinus Hulsius (Nürnberg 1596 ; voir Jones 2010 : 419–439) et du *Dictionnaire françois allemand latin avec une briefve instruction sur la prononciation de la langue françoise* publié par Jacob Stoer (Genève, Jacob Stoer, 1610 ; voir Dubois 2010) car il tire l'essentiel de la nomenclature et des définitions de ces deux ouvrages fondateurs de la lexicographie franco-allemande.

En réalité, nous observons que la plupart des articles sont identiques dans le *Dictionnaire françois allemand* (Genève, Stoer, 1610) et dans le *D FAL AFL* (1642) mais des modifications surviennent dans les éditions successives du *D FAL AFL*, notamment dans la troisième (Amsterdam, Elzevier, 1664). Voici deux articles :

<p><i>D FAL AFL</i>, Leyde, Hegher, 1642 (copié de Stoer 1610)</p> <p>Bacon : m. Au Dauphiné est ce qu'on dit lard ailleurs. Speck, brauchet, lard ; Lardum Sabaudis.</p> <p>Brisans : m. sont les chocs &amp; froissures des vagues de la mer, escumans au hurter contre les bancs ou escueils ; Sand im Meer darwider die wasserwellen stossen : das anstossen und brechen der wasserwellen wider solche führen ; Dorsum, Pulvinus, Scopuli.</p>	<p><i>D FAL AFL</i>, Amsterdam, Elzevier, 1664</p> <p>Bacon, lard : m. Speck / Lardum.</p> <p>Brisans : les choqs ou heurts de vagues de la mer contre les bancs &amp; les escueils, m. / Das anstossen der wasserwogen oder wellen wider den sand und die klippen im meer ; Fluctus maris ad brevia &amp; scopulos maris allisi et fracti.</p>
--	---



La comparaison des deux éditions permet de découvrir comment Duez a retravaillé ses sources. Il simplifie la définition en la rendant plus intelligible (dans le premier cas, il supprime les allusions aux territoires de Dauphiné et de la Savoie, plutôt obscures pour un public non français et somme tout inutiles) et indique le genre grammatical.

Mais d'autres renseignements précieux sur manière dans laquelle Duez a composé le *Dictionnaire* nous viennent des "Avis au lecteurs".

Dans la première l'édition, aucune source n'est mentionnée. Duez précise seulement qu'il a "pris la peine de collationner tant soit peu cette Edition avec les precedentes", se référant vraisemblablement aux éditions du *Dictionnaire* édité par Stoer que, comme nous avons pu constater, il copie abondamment.

Les avis de la deuxième et de la troisième édition sont plus développés et nous donnent plus de détails :

<i>D FAL AFL</i> , Amsterdam, Elzevier, 1650	<i>D FAL AFL</i> , Amsterdam, Elzevier, 1664
<p>mais principalement ay-je encor voulu employer une singuliere diligence, en ceste seconde impression, de rendre l'œuvre plus ample et plus parfait, non seulement en le repurgeant de quelques fautes, que j'y ay remarquées pendant l'espace de plusieurs années ; mais aussy en y adjustant une infinité de paroles, que j'ay trouvé y estre necessaires , de sorte que je te puis dire avec verité, que l'œuvre a esté encor augmenté plus d'un tiers, aussi bien en l'Allemand qu'au François. Car j'y ay premierement inseré un bon nombre de remarques, que j'ay faites de temps en temps, par la pratique journaliere, depuis la premiere Impression. En apres je l'ay aussi collationné tout au long hormis les trois premieres feuilles avec le Dictionnaire François et Italien d'Antoine Oudin ; lequel est le Dictionnaire du monde le plus riche, et le plus abondant en toutes sortes de paroles, que l'on ait jamais veu ; Et puis, pour l'augmentation de l'Allemand, je l'ay pareillement conferé avec le Dictionnaire d'André Corvin, et celuy de Basile Fabre, qui m'ont semblablement fourny une infinité de mots, qui n'estoient point auparavant en cette œuvre</p>	<p>mais principalement ay-je encor voulu employer une singuliere diligence, en cette troisieme et deriniere Edition, de rendre l'œuvre plus ample et plus parfait, non seulement en le repurgeant de quelques fautes, que j'y ay remarquées pendant l'espace de plusieurs années, et en reduisant tout en beaucoup meilleur ordre, qu'il n'estoit au auparavant ; mais aussy en y adjustant une infinité de paroles, que j'ay trouvé y estre necessaires , de sorte que je te puis dire avec verité, que l'œuvre a esté encor augmenté plus d'un tiers, aussi bien en l'Allemand qu'au François. Car j'y ay premierement inseré un bon nombre de remarques, que j'ay faites de temps en temps, par la pratique journaliere, depuis la seconde Impression. En apres je l'ay aussi collationné tout avec le Dictionnaire François et Italien d'Antoine Oudin ; lequel est le Dictionnaire du monde le plus riche, et le plus abondant en toutes sortes de paroles, que l'on ait jamais veu ; comme aussi avec ceux de Monet, de Pajot, et de Kotgrave. Et puis, pour l'augmentation de l'Allemand, je l'ay pareillement conferé avec le Dictionnaire d'André Corvin, et celuy de Basile Fabre, qui m'ont semblablement fourny une infinité de mots, qui n'estoient point auparavant en cette œuvre</p>

Nous apprenons que dans la deuxième édition Duez a augmenté considérablement (de plus d'un tiers) la nomenclature et qu'ensuite (pour la troisième édition) il s'est appliqué à donner plus de cohérence et une meilleure disposition au matériel ("reduisant tout en beaucoup meilleur ordre, qu'il n'estoit au auparavant").

Ses additions dérivent de son expérience personnelle ("pratique journalière") mais il déclare aussi avoir utilisé pour enrichir la nomenclature française le *Dictionnaire François et Italien* d'Antoine

Oudin (cf. *infra*, paragraphe 3.4), et, pour l'allemand, le *Thesaurus eruditionis scholasticæ* de Basilii Faber (Leipzig, 1571 ; voir Jones 2010 : 323-345) et le *Fons latinitatis* de Andreas Corvinus (Leipzig, 1623 ; voir Jones 2010 : 252-257).

Nous avons procédé à une vérification : voici quelques cas de lemmes (parmi ceux commençant par *Ba-*) absents dans du *D FAL AFL* de 1642, qui ont été ajoutés dans la deuxième édition du *D FAL AFL* et dont la contiguïté avec ceux contenus dans le dictionnaire d'Oudin d'où ils ont vraisemblablement été tirés, est éclatante.

Oudin, <i>Dictionnaire François et Italien</i> , Paris, Sommaville, 1640	<i>D FAL AFL</i> , Amsterdam, Elzevier, 1650
Babiche, babichon : Specie di cagnolino co' peli lunghi	Babiche f. Babichon, m. Ein art von Kleinen hunden mit langen haare ; Genus catellorum crine longo
Balzan, cheval balzan : Balzano	Balzan, cheval balzan, m. ; Ein pferdt mit welken füßen
Banal : commune, di bando, della comunità	Banal m. Gemein, für einem jeden von der gemeine; Publicus, communis.
Bancasse d'une galere où couche le capitaine : bancaccia	Bancasse de galere, La couche du capitaine ; Des Hauptmans schlaffkammer ; Cubiculum seu dormitorium præfecti aliculus triremis
Barbier : barbiere Tout beau barbier la main vous tremble ; l'italien dit piano barbiere che 'l ranno è caldo	Barbier, m. ein Sherer balbierer ; Tösor Tout beau barbier la main vous tremble, Gemach verbrennet die finger nicht, vergreiffet euch nicht

Seulement dans la "Préface" à la troisième édition Duez dit avoir tiré profit également de la consultation d'autres importants ouvrages lexicographiques antécédents : l'*Invantaire des deus langues latine et françoise* de Philibert Monet (Lyon, 1635), le *Dictionnaire nouveau françois-latin* de Charles Pajot (Paris, 1636),<sup>6</sup> et *A Dictionarie of the French and English Tongues* de Randle Cotgrave (London, 1611).

### 3.4 Dittionario italiano e francese Dictionnaire italien et françois, Leyde, Elzevier, 1659-1660

La première édition du *Dittionario italiano e francese Dictionnaire italien et françois*, paraît en 1659-60.<sup>7</sup>

Duez décide de se lancer dans l'entreprise de rectifier les outils lexicographiques existants pour le français et l'italien grâce à ses compétences et à son expérience, qu'il met en avant :

Car faisant profession de la langue françoise et italienne il y a plus de vingt quatre ans et le bon Dieu m'ayant fait la grace de me donner un assez bon talent en cette vacation, j'ose bien prendre la

<sup>6</sup> Selon Jones (2010 : 311) il pourrait aussi s'agir de l'une des trois ouvrages de Pierre Rayot (recensés dans Jones 2010 : 576-577).

<sup>7</sup> Pendant près de vingt ans il a plusieurs réimpressions, quasiment identiques dans le contenu et dans la mise en page, chez des éditeurs variées, en Italie, en France et en Suisse Pour une description détaillée nous renvoyons à Lillo (2008 : fiches 32, 33, 35, 36, 37, 38, 39, 40, 41, 43, 44), Bingen (1987 : 77-849 et Van Passen (1981).

liberté de te dire (ce que soit dit sans vanterie) que je m'en suis acquis une assez bonne connoissance. Et ayant une bonne douzaine d'années commencé à remarquer plusieurs fautes et manquemens en divers Dictionnaires, qui auoient esté mis en lumiere de ces deux belles langues et particulièrement de l'Italienne, j'ay creu estre obligé de mettre en fin telles remarques et corrections en bon ordre en quelque bonne edition et de les donner au public sur tout pour le proffit et auantage des jeunes gens qui voudront apprendre l'une ou l'autre de ces deux excellentes langues et ce affin de leur y donner une plus claire et plus veritable explication des choses plus difficiles et obscures que peut-estre ils n'en trouveront pas en d'autres. ("Préface")

Il ne cite pas les sources de son *Dittionario*, se limitant à affirmer : "bien qu'il y ait bon nombre de dictionnaires italiens dont les meilleurs qui ayent jamais esté sont celuy de Crusca, d'Oudin et de Françoisin,<sup>8</sup> si est ce qu'il n'y en a point auquel il ne se trouve encor beaucoup de manquemens et d'imperfections" ("Préface").

Il ne fait pas de doute,<sup>9</sup> cependant, qu'il est une réélaboration des *Recherches italiennes et françoises* d'Antoine Oudin, ouvrage qui dominait la scène lexicographique franco-italienne (sur cet ouvrage voir Pfister 1989, Minerva 2007), jugé "le dictionnaire bilingue le plus représentatif des tendances lexicographiques [franco-italiennes] de l'époque quant à l'abondance des lemmes retenus et à leur microstructure" (Minerva 2013 : 37).<sup>10</sup>

Contrairement à des lexicographes successifs, tels Ferretti ou Veneroni, qui se situent expressément dans la continuité d'Oudin (Minerva, 1991 et 2013), Duez, semble vouloir s'en démarquer. Il ne modifie pas foncièrement la nature et l'architecture des *Recherches* mais son exercice de remaniement est tout de même consistant.

Son plus grand mérite réside dans la réorganisation des articles. Il est le premier, dans la lexicographie franco-italienne, à numéroter les différentes acceptions des mots à la suite du mot-vedette :

(5) s.v. mellone

Oudin : melon. Par similitudes, les fesses

Duez : 1 melon 2 le derriere, les fesses 3 vn niais et vn benest

Les traduisants des mots polysémiques, traités chez Oudin dans des articles autonomes, qui multipliaient les entrées, sont réunis dans un même paragraphe :

(6) s.v. verso

Oudin : vers, composition poétique

l'endroit de l'estoffe

endroit, situation, costé. La note d'un oiseau

vers, envers

Duez : subst. 1 vn vers, composition poétique 2 l'endroit ou le droit d'une estoffe 3 vn endroit, vn costé 4 vne sorte, façon ou maniere 5 la note ou le chant d'un oiseau

prepos 1 vers, deuers, 2 enuers, à l'endroit de &c.

Il intègre dans le *Dittionario* des informations concernant les parties du discours, au moins lorsqu'elles sont indispensables pour faire des distinctions et dans le choix des traduisants :

<sup>8</sup> Lorenzo Franciosini, (1600-1645), toscan, hispaniste, grammairien et traducteur, avait publié le *Vocabolario italiano, e spagnolo* (Roma : G. Angelo Ruffinelli e Angelo Manni 1620).

<sup>9</sup> Nous résumons ici quelques résultats d'une analyse plus approfondie sur le *Dittionario* contenue dans Amatuzzi (2016).

<sup>10</sup> Paru en 1640, il eut trois réédition chez le même éditeur, Antoine de Sommaville, en 1643, 1653 et 1655. En 1663 Lorenzo Ferretti, romain, secrétaire interprète et maître de langues à la Cour de Paris, en donna une nouvelle édition avec le titre *Dictionnaire italien et françois Contenant les recherches de tous les mots italiens expliquez en françois*.

## (7) s.v. manco

Oudin : defectueux : gauche. Pour foible. & manquement  
moins

Duez : adv. 1 moins 2 pas mesmes, non plus, aussi peu, encor moins  
adj. 1 defectueux 2 gauche. 3 sinistre, malheureux 4 foible  
subst. faute ou manquement

Plus généralement, les articles du *Dittionario* sont plus complets que ceux des *Recherches*. Duez enregistre souvent d'autres acceptions du même mot :

## (8) s.v. \*agghiadire

Oudin : s'engourdir de froid ; sentir vn extrême froid

Duez : 1 s'engourdir & se transir de froid, sentir vn extrême froid 2 auoir peur, estre saisi & glacé de crainte-

ou en donne une définition moins approximative et plus claire :

## (9) s.v. macaron

Oudin : cosa fatta di pasta di mandole, zucchero acqua rosa &c.

Duez : sortelletto o mostazzolo di marzapane fatto di pasta mandole zucchero acqua rosa &c.

Parfois il glose :

## (10) s.v. arazzo

Oudin : tapisseries

Duez : tapis & tapisserie figurée de diverses couleurs, appelée ainsi de la ville d'Arazzo en Perse, où il s'en fait beaucoup

## (11) s.v. abelline

Oudin : sorte d'auelines rouges au dedans

Duez : une sorte d'auelaines ou noisettes rouges au dedans appelées ainsi de la ville Auellino en la campagne romaine dont est venu le mot auelaine

L'intervention de Duez se focalise surtout sur la phraséologie, qu'il accroît sensiblement.<sup>11</sup> Il recense des collocations, des locutions figées et des proverbes absents dans les *Recherches* :

## (12) s.v. asne Le moulin est fermé les asnes se jouënt, il molino è serrato, gli asini trescano

Il ajoute des acceptions et des traductions possibles :

## (13) s.v. aggiunta

Oudin : Val più l'aggiunta che la carne, cela se dit quand vne seruante est plus belle que sa Maistresse

Duez : Val più l'aggiunta che la carne, cela se dit quand vne seruante est plus belle que sa Maistresse ; ou l'accessoire est meilleur que la chose mesme

Il étoffe les articles avec de nombreuses citations d'auteurs italiens :

## (14) s.v. accarnare

Oudin : acharner

Duez 1 acharner, ou gaigner & prendre chair, comme fait vne playe en se guerissant 2 prendre par la chair, ou penetrer dans la chair avec les ongles ou avec autre chose 3 comprendre ou entendre La piaga si accarna, la playe prend chair

Se ben l'intendimento tuo accarno, si je comprends bien ce que tu veux dire. Dante au 14. chant de purgatoire

11 Son attention pour la phraséologie est confirmée par cette affirmation de la "Préface" : "Et là où il y a quelques phrases ajoutées à vn mot c'est pour monstrier vne particulière construction et vn usage remarquable de telle parole ou pour la dignité de quelque façon de parler fort notable". Le traitement de la phraséologie dans le *Dittionario* a été étudié par Murano (2012 : 48-53).

Les modifications que Duez apporte répondent aux exigences dictées par sa fonction d'instituteur et vont dans la direction d'une simplification de la structure des articles et d'un enrichissement des contenus, ses objectifs prioritaires étant l'intelligibilité et l'exhaustivité.

La place de relief qu'il accorde à la phraséologie contribue à développer la dimension culturelle. Les articles du *Dittionario* se limitent de moins en moins au mot avec ses traduisants et ils impliquent des définitions accompagnées de paraphrases, gloses, citations et références littéraires qui favorisent la découverte et la compréhension de la culture étrangère.

Bref, Duez, en bon enseignant fait un effort de pédagogie remarquable qui consiste à rendre plus accessible et à perfectionner avec des l'insertion d'informations plus étendues et moins imprécises le matériel lexicographique d'Oudin.

## 4 Conclusions

L'itinéraire lexicographique de Duez, que nous avons retracé, est significatif pour plusieurs raisons et nous permet d'apporter des éléments de réponse aux questions initialement posées.

Tout d'abord il atteste une évolution intéressante dans l'activité de Duez, analogue à celle connue par la lexicographie de l'époque où il vit. La *Janua* n'est autre qu'une traduction (qualitativement satisfaisante, d'ailleurs) d'un manuel scolaire contenant des phrases utiles à la communication. Dans la *Nomenclatura* des sources différentes sont sélectionnées et réorganisées pour produire un glossaire quadrilingue. Le *Dictionnaire* est un véritable vocabulaire. Si la première édition ne s'éloigne pratiquement pas du *Dictionnaire françois allemand latin avec une briefve instruction sur la prononciation de la langue françoise*, publié par Stoer, les éditions successives sont amplifiées grâce à la consultation et à la "collation" avec d'autres ouvrages. Dans le *Dittionario*, enfin, la présence "auctoriale" est plus évidente.

Pour ce que est de l'intertextualité, compte tenu du fait qu'à l'époque la pratique de recopier des travaux ouvrages documents déjà existants n'était pas jugée négativement en tant que plagiat, l'étude montre que Duez effectue une opération de réélaboration ponctuelle et délicate. Fort de la maturité et de l'expérience professionnelle qu'il acquiert au fil des années, il intervient de manière de plus en plus importante sur les documents source, qu'il s'approprie, au départ, sans une conscience critique.

Les finalités de son travail sont claires : les changements qu'il introduit ont comme but de rendre la consultation des dictionnaires plus facile et d'améliorer et élucider les définitions pour mieux satisfaire aux besoins de ses élèves.

La lexicographie a pu bénéficier de ses qualités pédagogiques incontestables. Ses innovations sont loin d'être négligeables.

## Références

- A. Amatuzzi (2016). Nathanaël Duez auteur du *Guidon de la langue italienne* (1641) et du *Dittionario Francese Italiano* (1659-1660) : un maître de langues entre continuité et innovation. In *Documents pour l'histoire du français langue étrangère ou seconde*, 56, pp. 27-50.
- N. L. Dodde (1997). Franse scholen van 1482 tot 1857. In *Meesterwerk* 9, pp. 2-7.
- A. Dubois, (2010). Jacob Stoer (1542-1610), un éditeur et ses auteurs. In A. Riffaud (éd.) *L'écrivain et l'imprimeur*, Rennes : Presses Universitaires de Rennes, pp. 75-93.



- F. J. Hausmann & al. (éd.) (1989-1991). Wörterbücher Dictionaries Dictionnaires Ein internationales Handbuch zur Lexikographie An International Encyclopedia of Lexicography Encyclopédie internationale de lexicographie Berlin / New York : de Gruyter.
- W. J. Jones (2000). German Lexicography in the European Context. A descriptive bibliography of printed dictionaries and word lists containing German language (1600-1700). Berlin/New York, De Gruyter.
- J. Lillo, Jacqueline (éd.) (2008). 1583-2000 : Quattro secoli di lessicografia italo-francese. Repertorio analitico di dizionari bilingue. Berne : Peter Lang.
- J. Lillo, Jacqueline (éd.) (2013). Les best-sellers de la lexicographie franco-italienne. XVI<sup>e</sup>-XXI<sup>e</sup> siècle. Rome : Carocci.
- P. L. M. Loonen, (1993). Nathanael Duez as an example of a distinguished language master in the seventeenth century. In J. Noordegraaf & F. Vonk (éds). Five hundred Years of Foreign Language Teaching in the Netherlands 1450-1950. Amsterdam : Stichting Neerlandistiek, pp. 57-66.
- P. L. M. Loonen, (1994). The influence of Comenius on modern language teaching. In *Paradigm* 15. En ligne : <http://faculty.education.illinois.edu/westbury/paradigm/loonan.html> [15 mars 2018].
- P. L. M. Loonen (1995). Nathanael Duez. Biography and a first bibliography, In *Meesterwerk. Berichten van het Peeter Heynsgenootschap*, 3, 2-15. En ligne : <https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbXwZWV0ZXJoZXluc2dlbm9vdHNjaGFwfGd4OjRkYzE5ZTZjYzZiZWJlYyZl> [12 mars 2018].
- P. L. M. Loonen, (2000). The influence of the Huguenots on the Teaching of French in the Dutch Republic during the 17th century. In J. De Clercq, N. Lioce & P. Swiggers (éds). Grammaire et enseignement du français. 1500-1700. Louvain : Peeters, pp. 317-333.
- G. Mattarucco (2003). Prime grammatiche d'italiano per francesi (secoli XVI-XVII). Florence : Accademia della Crusca.
- N. Minerva (1991). Jean Vigneron dit Veneroni (1642-1708). In *La Lettre de la SIHFLES* 11, pp. 8-10.
- N. Minerva (2007). Représentations de l'autre. L'italien et les italiens dans quelques dictionnaires bilingues des XVII<sup>e</sup>-XVIII<sup>e</sup> siècles. In D. A. Kibbee (éd) History of linguistics 2005, selected papers from the 10th International conference on the History of the Language Sciences (ICHOLS X), 1-5 september 2005. Amsterdam-Philadelphia, John Benjamins, pp. 308-320.
- N. Minerva, (2013) Un siècle de lexicographie bilingue : le *Dictionnaire* de Giovanni Veneroni et ses adaptations. In J. Lillo (éd.) Les best-sellers de la lexicographie franco-italienne. XVI<sup>e</sup>-XXI<sup>e</sup> siècle. Rome : Carocci, pp. 33-51.
- M. Murano (2012). Des phrases aux séquences figées. La phraséologie dans les dictionnaires bilingues franco-italiens (1584-1900). Bologne : Clueb, Quaderni del Cirsil 11.
- M. Pfister (1989). L'importance d'Antoine Oudin pour la lexicographie française et italienne. In M. Giacomelli-Deslex (et al. éd.) La lingua francese del Seicento. Bari-Paris : Adriatica-Nizet, pp. 91-103.
- S. Szoc Szoc (2009). Esclaireissement sur deux maîtres plurilingues du XVII<sup>e</sup>siècle à Leyde . In *Documents pour l'histoire du français langue étrangère ou seconde* 42, 65-86. En ligne : <http://dhfles.revues.org/692> [10 mars 2018].
- M. Van Strien-Chardonneau, Madeleine & M.-C. Kok Escalle (2010). Le français aux Pays-Bas (XVII<sup>e</sup>-XIX<sup>e</sup> siècles) : de la langue du bilinguisme élitare à une langue du plurilinguisme d'éducation. In *Documents pour l'histoire du français langue étrangère ou seconde* 45, pp. 123-156. En ligne : <http://dhfles.revues.org/2448> [10 mars 2018].
- B. Von Gemmingen (1999). Hulsius - Stoer – Dhuez : Bemerkungen zur französisch-deutschen, deutsch-französischen Lexikographie in der ersten Hälfte des 17. Jahrhunderts, In H. E. Wiegand (éd.), Studien zur zweisprachigen Lexikographie mit Deutsch IV, Olm : Hildesheim/New York, pp. 81-110.



# A Workflow for Supplementing a Latvian-English Dictionary with Data from Parallel Corpora and a Reversed English-Latvian Dictionary

*Daiga Deksnē<sup>1</sup>, Andrejs Veisbergs<sup>2</sup>*

<sup>1</sup>Tilde, <sup>2</sup>University of Latvia

E-mail: [daiga.deksne@tilde.lv](mailto:daiga.deksne@tilde.lv), [andrejs.veisbergs@lu.lv](mailto:andrejs.veisbergs@lu.lv)

## Abstract

The lexicon of contemporary languages is changing rapidly, mostly by acquiring new loans and derivations. The change in lexicon is best reflected in the corpora of contemporary languages. Nowadays many collections of parallel-aligned texts are available electronically. To satisfy user needs for a modern, complete, up-to-date dictionary, we created a workflow for enriching the existing Latvian-English dictionary with data from parallel corpora containing lexis commonly used in contemporary language, as well as data from the reversed English-Latvian dictionary. While revising the existing Latvian-English dictionary, we identified some issues, for example, missing feminine forms of the nouns naming nationalities and occupations, representation of the words with optional parts or spelling variations. The task of dictionary improvement was done semi-automatically by the joint work of a lexicographer, computational linguists and programmers. Such natural language processing tools as a tokenizer, part-of-speech tagger, lemmatizer and spell-checker were used to reduce the manual work. As a result, the number of entries has increased by 32%, and the number of translations by 28%.

**Keywords:** electronic dictionaries, parallel corpora, NLP tools, XML format

## 1 Introduction

Electronic dictionaries are undergoing a huge expansion, both as concerns their production as well as their use. Though relatively new they also have to be updated regularly (Lorentzen & Trap-Jensen 2016) as the lexicon of contemporary languages is in a constant flux, with new items (especially loans and derivations) proliferating. Updating and expanding of dictionaries is a laborious and time-consuming process. However, in contrast to printed dictionaries, production of which also presumes considerable time for proofreading and printing, electronic dictionaries can be edited, supplemented and corrected promptly.

Moreover, electronic dictionaries are much less affected by space limitations (and mostly by screen size with regard to this issue). This affects some macrostructure issues, e.g. while regular derivatives (participles, verbal derivatives with prefix *non-*, agent nouns, occupation and nationality nouns in feminine and other categories) are generally avoided in printed Latvian dictionaries, these entries can be introduced in electronic one.

In order to create an electronic bilingual dictionary that corresponds to the users' current needs we created a workflow for merging three different data sources: an electronic version of the largest Latvian-English dictionary (Veisbergs 2016), the automatically reversed English-Latvian dictionary, and new entries from aligned bilingual parallel corpora. The dictionary in question is a large, general one, aimed at a relatively advanced Latvian user of English. It is mono-directional (aimed at the Latvian user) with no explanations for the Latvian part, while explanations for the English part are provided

where possible: labels, register, nuance markers, and bracketed semantic explanations. The entry structure consists of meanings, examples, and collocations subdivided by meaning, while the idiom block comes at the end of the entry.

There are different views as to the results (Newmark 1998; Geisler 2002; Veisbergs 2004; Krek, 2008) and efficiency (Geisler 1999; Tamm 2002; Veldi 2010) of bilingual dictionary reversing. Some studies and experiments are positive, others point to too much “noise” and extensive editing and proofreading that is too laborious.

The team has experience in creation of electronic dictionaries and dictionary-browsing environments on different platforms, enabling users to search in several dictionaries simultaneously. Integration of spell-checking and morphological analysis allows looking for inflected forms or finding the translation even for misspelled words. A uniform XML format for dictionary entry representation has been developed, and all dictionary resources are parsed and stored internally using this format (Deksne et al. 2013).

## 2 The Drawbacks of the Existing Latvian-English Electronic Dictionary

There were some drawbacks in the existing dictionary-browsing environment concerning representation of entries, comprehension of dictionary content by the user, and missing content.

The existing dictionary-browsing environment allowed searching for entries in several dictionaries at once; the results from different sources were presented to the user in a sequential order; translations or examples (as in Figure 1) in the direct and in the reversed dictionary tend to overlap; a user is obliged to scroll through a long list of identical results. Sometimes one source contained a hyphenated form of a compound, while another contained a non-hyphenated form; for example, there were translations ‘tom-cat’ and ‘tomcat’ in the different data sources for the same headword. There were around 11,000 entries with the same headword in the existing Latvian-English dictionary and in the reversed English-Latvian dictionary, making users wonder where the differences lay. They were thus merged in the new version of the Latvian-English dictionary.



Figure 1: Entries from two different data sources in the dictionary-browsing environment.

- Latvian bilingual dictionaries traditionally do not contain words created by regular derivation rules, e.g. participles used as adjectives, negative forms of verbs or adjectives, feminine forms of nouns naming occupation or nationality. Frequently headwords having masculine and feminine forms are given with an additional ending, as in printed dictionaries. Examples (1) and (2) show the masculine forms with additional feminine endings for the nationalities ‘Swede’ and ‘German’. Example (3) represents the masculine and feminine forms of the occupation name ‘telephone operator’. As the root of the word in the dictionary is not marked, the feminine form cannot be expanded automatically. Some words were given with a spelling variant in parentheses (see (4), Eng. ‘activation’). Such a format did not allow the user to find both forms of the word using a search engine in the dictionary-browsing environment.
  - (1) zviedrs/iete
  - (2) vācietis/e
  - (3) telefonists/e
  - (4) aktiv(iz)ēšana
- There were examples containing variations of one or several words separated by slash ‘/’ symbol. Example (5) shows a phrase which expands to four different phrases (6), (7), (8), and (9). This format does not take an extra space but is not suitable for search, and is hard to comprehend for a person who does not know the foreign language very well.
  - (5) to book/engage a season ticket/pass
  - (6) to book a season ticket
  - (7) to book a season pass
  - (8) to engage a season ticket
  - (9) to engage a season pass
- Another discrepancy was the representation of Latvian words translated in English as adjectives with an attributive meaning.
  - (10) **pilsēta** town borough; (liela) city
  - (11) **pilsētas** urban; municipal; town (attr.); towny
  - (12) olveida
- There were about fifteen hundred such entries. The Latvian word is a noun in the genitive case with or without an ending. Sometimes in Latvian dictionaries a hyphen character is used to depict the genitive. Some compounds are used only in a genitive (see (11), Eng. ‘egg-shaped’) but for the most nouns, the base form is nominative. Using the genitive case of the noun for the main entry without additional information may often be confusing, as the genitive case may coincide with the plural nominative or accusative (for example, the bold formatted headword in the entry (11) is a single genitive form of the headword in (10), but the same form is also plural nominative or accusative). A label *gen.* was added to solve this homographic issue.
- Besides, as new words proliferate any dictionary is lagging behind. There are numerous words which have appeared in English in the last few decades, such as ‘geocaching’, ‘flash mob’ or ‘raw-foodist’. Their Latvian counterparts and corresponding English equivalents were also added to the dictionary.

### 3 Compilation of Lists Containing Translation Hypotheses

A parallel corpus is a valuable resource when looking for new entries for dictionary supplementing. We compiled a corpus for possible translation extraction from several sources. The first part is a proprietary collection of parallel data used in different projects. The second part is formed by some components of an open source parallel corpus OPUS – a collection of translated texts from the web



(Tiedemann 2012). We use a collection of EU Translation Memories, documentation from the European Central Bank, documents from the European Medicines Agency, proceedings of the European Parliament and some other collections.

The Latvian text (as in (14)) was part-of-speech tagged and lemmatized (as in (15)) using NLP tools created by company *Tilde* (Deksne 2013; Pinnis & Goba 2011) while the English text (as in (13)) was left unchanged. Such a parallel-aligned corpus was passed to the next step of processing.

(13) account creation guide

(14) konta izveides norādījumi

(15) kontsN izveideN norādījumsN

The Moses toolkit used for statistical machine translation (Koehn et al. 2007) was employed for building the phrase tables. Each line in a phrase table contains a pair of Latvian and English word/phrase. These are hypothetical translations which have been obtained automatically using statistical methods. The pairs occurring only once and the stop words were filtered out. We made a list of word and translation pairs which were already present in the existing Latvian-English dictionary, and these were filtered out from the phrase table too. The rest of the lines were grouped by part-of-speech of the Latvian word. As a result of this process, we acquired lists of nouns, verbs, adverbs, adjectives, and interjections with hypothetical translations (see Table 1). Among the nouns and adjectives, there were many deverbalized derivations and compounds. Among the verbs, there were many negative verbs as well as the prefixed verbs (in the Latvian language, verb prefixation process is very productive). There are 11 prefixes used in verb formation. The verbal prefix can formally change some features of an aspect of the verb; it can also modify or change the lexical meaning of the base verb (Holvoet 2001).

Table 1: The number of entries and their hypothetical translations extracted from parallel corpora.

Word class	Number of entries	Number of hypothetical translations
noun	8,609	41,242
verb	4,928	29,617
adverb	1,483	7,847
adjective	1,025	4,247
interjection	64	523

We prepared several files in a simple tab separated format containing verbs, nouns, adjectives, adverbs and interjections. The second column contained the lemma of the word in Latvian, the third column contained the possible English translation, and the fourth column contained the frequency of the word pair in the corpus. The first column was reserved for the lexicographer's marking of possible inclusion in the dictionary (see the example in Figure 2). The lexicographer was asked to put a meaning number in the first column of the line valid for inclusion in a dictionary. The lexicographer was also instructed to manually add a comment in parentheses or some usage samples at the end of the line if necessary.

1	punktots	dotted	75
1	punktots	spotted	3
	punktots	background	2
	punktots	dotted lines	2
	punktots	green	2
1	punktots	polka-dot	2
	punktots	text	2

Figure 2: The adjective *punktots* with statistically acquired hypothetical translations from the corpus.

A total of 5,995 pairs or 7.18% of hypotheses (4,082 unique Latvian words, i.e., dictionary entries) were marked as suitable for inclusion in a dictionary.

## 4 Process Workflow

The workflow for enriching a Latvian-English dictionary consisted of several steps. Some tasks were automatized, and some involved manual work of a computational linguist or a lexicographer. The lexicographer regularly updates the dictionary in an MS Word document using rich formatting, e.g., a font style for entry title must be bold, a font style for comments or usage and grammatical information must be italic, specialized meanings have different punctuation symbols: commas, semicolons, colons, dashes. The computational linguist created scripts for transforming the content of the dictionary from the MS Word format (see Figure 4) to the proprietary dictionary XML format (see Figure 3). Specific XML tags allow marking of all parts of an entry. Special tags are reserved for headwords, translations, samples, sample translations, idioms, meaning numbers, usage information, comments, grammatical information, and punctuation symbols. It is important to scrupulously comply with formatting rules in the MS Word document, as errors in formatting can invoke errors in XML representation.

The next step was to merge the XML representation of the Latvian-English dictionary with the data from parallel corpora and from a reversed dictionary. Data from TAB separated lists of translation hypotheses from parallel corpora was converted to the XML format. A special color attribute was appended to the title and the translation tags to mark this entry as coming from a different source. As the structure of TAB separated lists is very simple, this step was easy to implement. The new entries were appended to the XML representation of the Latvian-English dictionary and all entries were sorted alphabetically.

```
<entry title="aisbergs">
  <title>aisbergs</title>
  <transl>iceberg</transl>
  <idiom />
  <from_sample>aisberga redzamā daļa</from_sample>
  <to_sample>tip of the iceberg</to_sample>
</entry>
```

Figure 3: Sample xml entry for a Latvian word *aisbergs* (Eng. ‘iceberg’).

The reversed dictionary was first filtered by removing translations and usage examples which were already present in the Latvian-English dictionary. The filtered version of the reversed dictionary was then merged with the main dictionary by including a whole entry if an entry with such title word did not exist, or by adding the translations and the usage samples at the end of the existing entry. A different color attribute was appended to the title and the translation tags to mark this entry as coming from the reversed dictionary. A new MS Word document was generated from the internal XML format. The information coming from the XML tags with a color attribute was reflected in the MS Word document (see Figure 4). The prepared document was then passed to a lexicographer for post-editing. In such a format the lexicographer could distinguish the parts of the dictionary coming from different sources and make the necessary corrections, such as, for example, reordering the translations or grouping them in a separate meaning.

**aizsargiepakojums** protective bag, protective packaging  
**aizsargierīce** safety device; protective equipment; (*ieroču*) safety-bolt; (*uz dzelzceļa*)  
 safeguard; protection device; *tehn.* **protector**  
**aizsargjosla** 1. *mil.* defence zone; 2. *bot.* **protective zone**; *meža a.* – forest shelter belt,  
 green belt  
**aizsargkārtā** coating, protective layer  
**aizsargkonstrukcija** protection structure  
**aizsargkrāsa** protective colouring; *mil.* camouflage colours; *poligr.* **safety ink**  
**aizsargkrēms** barrier cream  
**aizsargķivere** helmet, hardhat *amer.*

Figure 4: MS Word document with automatically merged entries.

## 5 Results

Editing the data extracted from parallel corpora involved the deletion of numerous entries that were grammatically erroneous, e.g. under adverbs many nouns in plural appeared, as both categories contained the ending *i*. There were also numerous gerund/participle entries that had fully predictable standard forms in both Latvian and English, which were considered not worth keeping. Likewise, Latvian verbs with a standard negative prefix which in translation would normally be equivalent to the English verb plus particle *not* were not included.

As a result, the new entries are mostly derivatives, chosen from the long list on the basis of two criteria – frequency and irregularity of English counterparts (that a Latvian user might not be able to surmise). There was also a considerable number of “missing” entries that for some reason had not been in the old dictionary. The combined version also yielded some “double entries” – either the result of wrong spelling of the Latvian word or in some cases parallel spellings. The latter were then joined in full form to the main entry. Very specialized terms or rare and obsolete words were avoided. Apart from the Latvian headwords, there was a huge number of additional English equivalents which were added to the English part, distributed among the senses or examples or added to the idiom block.

We also added gender differentiation in Latvian entries, thus having double Latvian entries for those English equivalents where gender does not differ (as in (16) and (17)), and separate entries for those where English has gender marked lexical units (the masculine forms as in (19), (20) and (23); the feminine forms as in (18), (21) and (22)).

- (16) *kinoapmeklētājs, kinoapmeklētāja* filmgoer, moviegoer *amer.*
- (17) *klasesbiedrs, klasesbiedre* classmate
- (18) *dzejniece* poetess; poet
- (19) *dzejnieks* poet
- (20) *kinoaktieris* film actor, screen actor
- (21) *kinoaktrise* film actress, screen actress
- (22) *cariene* tsarina, *tzarina, czarina*
- (23) *cars* tsar, *czar.*

Apart from these, some common abbreviations as well as some proper names were included in the entry list. While the printed dictionary traditionally had a separate appendix for geographical names, the electronic version does not differentiate between such categories. The author/lexicographer has concluded that in future the printed version will also drop the appendix and provide the geographical names in the single entry list. This seems to be a more reader-friendly approach.

## 6 Collateral Ideas and Solutions

Though most of the supplementation focused on derivatives, new meanings and extra equivalents, some issues of collocations and idioms were also brought into focus. It is well known that dictionaries often tend to compartmentalize the information by linguistic categories. This is partly inevitable as a result of the essence of dictionary – providing isolated, generalized material, not contextual use (the latter being almost indescribable in its complexity). Yet compartmentalization tends to be affected by theoretical linguistic categories, which is a somewhat scholastic exercise, trying to draw precise borders for concepts such as idiom, collocation, compound, and word. This leads to “a tendency to circumscribe the research field for purposes of consolidation” (Burger 2007: 11), while corpus linguistics produces the opposite. While theoretically usually correct, these divides are often artificial, as the strictly defined linguistic concepts do not reflect the fuzzy, blurred and scalar reality of the language, especially in a multilingual setting. Clutching at the mandatory correspondence of categories (idiom for idiom, collocation for collocation, word for word) in dictionaries is not sensible, but is often practiced and also emphasized in research. Bo Svensén plainly states “idioms in the source language must as far as possible be paralleled in the target language by idioms with the same content” (Svensen 1993: 156), and thus presented idiom for idiom. However, this is not always possible, nor should it be mandatory: language structures are different, so are ideas about some linguistic concepts, like idiom and compound.

The reversal exercise showed that some idioms in the English-Latvian dictionaries were translated as words and some words as idioms, and these were full equivalents. Should we avoid the reverse – giving some idioms as an equivalent for some Latvian words, and some words as an equivalent for some idioms – just because there is a category divide?

Sometimes an idiom was the only adequate equivalent for a word, e.g.

- (24) *sastikēt* [to together stick] to chip in;
- (25) *apmulķot* [to around fool] to make\* a fool (of), to fool, to take\* for a ride;
- (26) *apuišot* [to around boy] to fetch and carry, to wait on hand and foot.

In its turn, a Latvian idiom might have an English collocation or lexical counterpart that would be a better semantic match than an idiom with an analogous image, e.g.

- (27) *domu grauds* [thought grain] aphorism, maxim;
- (28) *ziedu laiki* [blossom days] heyday, highday, palmy days, prime, zenith;
- (29) *sarkanais gailis* [the red cock] fire; *ielaišt sarkano gaili* [to let the red cock in] – to set\* fire (to).

Sometimes an idiom would have several equivalents: words, phrases, idioms:

- (30) *tukši vārdi* [empty words] mere words, wind, hot air, lip service.

Finally, a simple entry that clearly illustrates the structural and semantic shifts between languages:

- (31) *galarezultāts* [end-result] outcome, the end result; ◇ *galarezultātā* [in the end-result] – at the end of the day; in the end.

The Latvian compound, corresponding to an English compound or collocation, deserves an English idiom when used in a declined form.

This presented a more flexible approach to the idiom-word divide, tearing down the conventional barriers of lexicographical thinking and practice. We should think more in terms of equivalence of meanings, not structures, words or phrases (Atkins & Rundell 2008). We believe that dictionaries

should be “much more phrasal than they currently are” (Granger 2008: 1353), as it is well known that “idiomaticity facilitates communication” (Bejoint 2000: 216).

## 7 Conclusions

The results of our work: the former Latvian-English dictionary contained 36,608 entries with 96,066 translations and 22,090 usage samples. The reversed dictionary contained 20,253 entries, a large part of entries partly or fully overlapped with the former Latvian-English dictionary. The new enlarged Latvian-English dictionary contains 53,867 entries, 132,481 translations and 22,277 usage samples including 4,082 new entries which were added after processing parallel aligned corpus. Despite protracted post-editing work, the accomplished end result is impressive. It not only considerably increased the number of entries, senses and equivalents, but also yielded interesting theoretical insights in the practical lexicography, like idiom treatment, genitive – attributive words, among others. One should also consider the benefits of a massive increase in the number of items for digital use and machine translation purposes.

The dictionary is available online at <https://www.letonika.lv/groups/default.aspx?g=2&r=10621033&f=1>.

## 8 References

- Atkins, B.T.S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: OUP.
- Bejoint, H. (2000). *Modern Lexicography*. Oxford: OUP.
- Burger, H. (2007). Semantic aspects of phrasemes. In D. Burger, D. Dobrovolskij, P. Kuehn, N.R. Norrick (eds.) *Phraseologie. Vol. 1*. Berlin, New York: Walter de Gruyter, pp. 90-109.
- Deksne, D., Skadina, I., & Vasiljevs, A. (2013). The modern electronic dictionary that always provides an answer. In *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. pp. 421-434.
- Geisler, C. (2002). Reversing a Swedish-English dictionary for the Internet. In L. Borin (ed.) *Language and Computers, Parallel Corpora, Parallel Worlds*. Amsterdam: Rodopi. pp. 123-133.
- Granger, S. & Paquot, M. (2008). “From dictionary to phrasebook?” In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, pp. 1345-1355.
- Holvoet, A. (2001). *Studies in the Latvian Verb*. Kraków: Wydawnictwo universitetu Jagiellońskiego.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. and Dyer, C., (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 177-180.
- Krek, S.; Šorli, M.; Kocjancic, P. (2008). The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary’. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 535-542.
- Lorentzen, H. and Trap-Jensen, L. (2016). What, When and How? – the Art of Updating an Online Dictionary. In T. Margalitadze, G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress*. Tbilisi: Ivane Javakishvili Tbilisi University Press, pp. 138-145.
- Newmark, L. (1998). Reversing a One-Way Bilingual Dictionary. In Th. Fontanelle et al. (eds.) *EURALEX 1998 Proceedings*. Liege: University of Liege.
- Svensen, B. (1993). *Practical Lexicography*. Oxford, New York: OUP.
- Tamm, A. (2002). Reversing the Dutch-Estonian Dictionary to Estonian-Dutch. In A. Braasch, C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress. Vol. 1*, Copenhagen: CST, pp. 389-399.
- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC Vol. 2012*, pp. 2214-2218.



- Veisbergs, A. (2004). Reversal as Means of Building a New Dictionary. In G. Williams, S. Vessier (eds) *Proceedings of the Eleventh EURALEX International Congress*. Lorient: UBS. Vol. I, pp. 327-332.
- Veisbergs, A. (2016). *The New Latvian English Dictionary*. Riga: Zvaigzne ABC.
- Veldi, E. (2010). Reversing a Bilingual Dictionary: a mixed blessing? In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. 6-10 July 2010. Leeuwarden/Ljouwert: Fryske Akademy – Afûk. pp. 861-865.

## Acknowledgements

The research has been supported by the European Regional Development Fund within the project “Neural Network Modelling for Inflected Natural Languages” No. 1.1.1.1/16/A/215.



# Towards a Representation of Citations in Linked Data Lexical Resources

**Anas Fahad Khan, Federico Boschetti**

*Istituto di Linguistica Computazionale “Zampolli”, CNR, Pisa*

*E-mail: fahad.khan@ilc.cnr.it, federico.boschetti@ilc.cnr.it*

## Abstract

In this article we look at the modelling of citations in lexical resources in linked data. We start by discussing the treatment of citations in linked data and in TEI; we also look at the idea of different conceptual levels as posited by models such as TEI and FRBR. We argue that in representing citations in lexical resources it is important not to confuse different levels of information, and that at least in the case of attestations it is important to model the purpose of a citation, or the claim that is being made by that citation, separately. We develop this point with two separate examples before presenting *lemonBib*, our extension of the lemon model based around the idea of a lexical attestation. We also give a treatment of part of one of the examples described previously in the article.

**Keywords:** linked data, citations, bibliographic data, lexical resources

## 1 Introduction

Up until quite recently most lexical resources published as linked (open) data have tended to be born-digital (having been developed in many cases with specific NLP tasks in mind), lately, however, there has begun to be an increased interest in the provision of retrodigitized lexical resources on the Semantic Web, and especially of those legacy resources regarded as authoritative or which are thought to hold some particular historical interest. Publishing such works as linked open data has the obvious advantage of making the information contained in them much more accessible and available to a wider public than was previously possible. At the same time, that information is structured according to a common data framework, RDF, which makes individual resources more interoperable as well as more amenable to various kinds of automated or semi-automated processing, more so than if they were text files, say. In addition, thanks to the fact that the Semantic Web offers data modelers a simple, standardized way of creating links between individual datasets, it also makes it easier to enrich an original lexical resource with links to other datasets such as, say, biographical or geographical ones. What’s more, the Semantic Web offers modelers the possibility of rendering the links between resources meaningful by giving them an explicit, formal ‘semantics’, and thus clarifying the ways in which individual datasets can help to augment the knowledge contained in other datasets. The process of converting or migrating retrodigitized lexicographic resources into RDF brings to the fore a number of different modeling challenges that concern aspects of lexicons that are usually less prominent in born-digital, NLP-oriented lexical resources. One such challenge is that of the correct modelling of lexical attestations: that is, of citations used to attest to different properties of individual lexical entries. For instance, in the case when a given lexical entry cites a particular text as exemplifying the use of a word with, say, a given sense or given orthography, it would be useful to be able to link to that text in the linked data version of the entry, and to information about the work and the author, and perhaps also to the secondary literature; the Semantic Web seems to be particularly apt in cases such as these. Clearly we would like to be able to represent as much of the information contained in the

original lexical entry as possible using the graph-based data framework of RDF, but it is also crucial (given the formal, ‘semantic’ nature of the Semantic Web) that we respect the conceptual differences between the kinds of information present in a citational act, and this calls for a more detailed and specific treatment. At the end of the day, the fact that a lexicographer or group of lexicographers decided, during the compilation of a lexicographic work, to attest to the existence of the property of a word by citing a relevant text is a salient piece of data, and one that it is worthwhile trying to model properly (even if this kind of information hasn’t featured as strongly in previous lexical linked datasets).

In this article we take a detailed look at a number of issues which arise when it comes to modelling lexical citations as linked data; we will look at examples taken from retrodigitized lexicographic resources or that concern linked data versions of print resources as contexts in which certain of these issues become much more conspicuous (although of course they also apply to born-digital resources). We will work towards a provisional set of properties and classes that, together with already existing RDF vocabularies, will help to capture some pertinent aspects of citations and attestations in lexicons. On the way we will discuss some of the pre-existing vocabularies and models for representing this kind of information with a view to their adequacy to the case at hand.

## 2 Background

In this section we will discuss related work that deals with the representation of bibliographic records and citations, both in the specialised case of computational lexicons, as well as within the general framework of linked data resource. We will begin, in Section 2.1, by looking at a useful distinction that is made within the TEI model between different ways of viewing lexical datasets (and which will be relevant for the discussion which follows in the rest of the paper) before moving on to describe the TEI approach to representing lexical citations in detail. In Section 2.2 we give a brief overview the influential FRBR model which has had an important impact on the representation of citations in linked data, as well as in the field of library science more generally. A more detailed discussion of citations in linked data is given in Section 2.3.

### 2.1 TEI: Zero, One, Two Dimensional Views and Citations

The Text Encoding Initiative (TEI) refers both to a widely used standard for encoding digital texts in XML, as well as to the consortium that maintains and develops the standard<sup>1</sup>. TEI, the standard, is available as a set of guidelines<sup>2</sup> (Burnard & Bauman 2008) which are used to define an XML schema. These guidelines are divided up into several parts and include a number of specialist modules, each of which deals with a different type of text. Dictionaries, in particular, have their own dedicated module TEI-DICT<sup>3</sup>. One interesting aspect of these dictionary guidelines, which will turn out to be extremely pertinent in what follows, is the explicit distinction that they make between three different ways of viewing dictionary data, these are: (a) **the typographic view**; (b) **the editorial view**; and (c) **the lexical view**. It will be useful to go into some detail on this threefold distinction since, at the very least, it will motivate our own separation of lexical citations into different conceptual levels below. The first view, the typographic view, essentially concerns the layout of a page – so, for instance, where the line breaks are in a text, or how the entries are arranged visually on any single page; for obvious reasons the authors of the TEI guidelines refer to this view as the ‘two-dimensional’ view. The second view, the editorial one, deals with the properties of a text modeled as a sequence of tokens. Accordingly, if

1 <http://www.tei-c.org/index.xml> [accessed 29/03/18]

2 These guidelines can be found on the TEI website: <http://www.tei-c.org/Guidelines/> [accessed 29/03/18]

3 <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html> [accessed 29/03/18]

we take this view with respect to a print dictionary, then for any specific entry we will be interested in exactly which words are used in the entry and in which order, along with the exact placement of punctuation in the text. The authors of the guidelines identify the editorial view as ‘one-dimensional’, since it is effectively concerned with a *linear* sequence of tokens. The third and final view mentioned in the guidelines, the lexical, relates to the conceptual or linguistic content of a lexicon or dictionary as well as each of its individual entries: to the fact, for instance, that a particular lexicon focuses on the medical domain or that the grammatical category of a given entry is “verb”; we might tentatively refer to this view as the ‘zero-dimensional’ view, although this term is not used in the document itself. TEI-DICT is, avowedly, a model for encoding all three views, something which has resulted in a relatively complex set of modeling guidelines, in which there exist several different ways of modelling the same information. In effect however we can lump the first two views together as dealing with the mode of presentation in a lexical resource, and isolate the last view as describing the content or meaning of the information itself. In the next subsection we will discuss the FRBR model which makes a similar classification of the different kinds of information that can be potentially referred to in a bibliography. As we shall see, this classification is not entirely orthogonal to the tripartite TEI classification discussed above. Before we move on, however, we should look at what provision the TEI-DICT guidelines offer for the encoding of citational information. In the case of citations that include quotations, the TEI guidelines recommend the use of the <cit> element; bibliographic references to other works can then be added using the <bib> element. As an illustration of the recommendations made by the TEI-DICT guidelines we will take an example from the Perseus TEI-XML encoding of the Liddell Scott Jones Ancient Greek-English lexicon (Liddell et. al. 1925), a hugely influential and authoritative Ancient Greek-English dictionary which was first published in 1843 and which is currently still in print in its 9<sup>th</sup> edition. The example in question is the following (presented in its original formatting):

**Ἀβρων**, ὠνος, ὁ,

**A.Abron**, an Argive, proverbial for luxurious living, “Ἀβρωνος βίος” Suid., Zen.1.4.

According to the TEI guidelines we can serialize the example as follows in XML:

```
<entryFree id="n210" key="*/abrwn" type="main" opt="n">
  <orth extent="full" lang="greek" opt="n">Ἀβρων</orth>
  ,
  <itype lang="greek" opt="n">ὠνος</itype>
  ,
  <gen lang="greek" opt="n">ὁ</gen>
  ,
  <sense id="n210.0" n="A" level="1" opt="n">
    <tr opt="n">Abron,</tr>
    an Argive, proverbial for luxurious living,
    <cit>
      <quote lang="greek">Ἀβρωνος βίος</quote>
      <bibl default="NO"><author>Suid.</author></bibl>
    </cit>
    ,
    <bibl n="Perseus:abo:tlg,0596,001:1:4" default="NO">
      <author>Zen.</author>
      <biblScope>1.4</biblScope>
    </bibl>
  </sense>
</entryFree>
```



Here the two citations given in the original dictionary text as “‘Ἀβρωνος βίος” Suid.’ and ‘Zen.1.4’ respectively are encoded in the first case with a <cit> element containing a <quote> and <bibl> element, and in the second case with a <bibl> element.

## 2.2 Functional Requirements for Bibliographic Records

The Functional Requirements for Bibliographic Records (FRBR) entity relationship model is perhaps the single most influential conceptual model so far devised for the representation of bibliographic data in computational resources (linked data resources being no exception to the trend). It was developed by the International Federation of Library Associations and Institutions in the early 1990s (Tillett 2007), and then in a subsequent development was harmonized with the well-known CIDOC-CRM conceptual model and published as a formal ontology called FRBR-oo (the ‘oo’ standing for object oriented) (Boeuf 2012). An expression of the core concepts of FRBR has been made available in RDF, and there also exists an RDF version of FRBR-oo, and so the FRBR model is, in effect, ready to use in the construction of RDF datasets. With respect to the contents of the model, FRBR makes a fourfold distinction in describing bibliographic entities on the basis of the particular ontological status which each entity holds. We present the classification as it pertains to texts, although these categories can just as well be applied to other kinds of bibliographically referable entity. The categories are (in ascending levels of concreteness):

- *Work*: those aspects of a text that can be abstracted away from any particular linguistic representation: so that for instance all the translations of a text, e.g., Hamlet, Amleto, हैमलेट, 哈姆雷特, etc., refer to the same Work under this view; the 0-dimensional TEI Lexical View seems to largely overlap with this category;
- *Expression*: the specific linguistic form which a Work takes, this view includes all of that which gets lost in translating a Work from one language to another;
- *Manifestation*: a physical embodiment of an Expression, e.g., the 2015 Penguin Classics edition of Hamlet;
- *Item*: a specific instance of a Manifestation, so for instance, I could use this category to refer to the copy of the 2015 Penguin Classics edition of Hamlet that is currently held in my local library.

It is clear from this description that the TEI lexical view corresponds to the FRBR concept of Work, the editorial view to Expression, and the typographical to Manifestation (and perhaps also to Item). Of course, it is important that we make the disclaimer here that the conceptual distinctions made by TEI and FRBR should not be considered as watertight, in fact they turn out to be very difficult to apply in certain kinds of concrete instance (something which we discuss in more detail in Section 4.2). In many other cases, however, they have proven to be very useful approximations.

## 2.3 Bibliographies and Citations in Linked Data

There exist a number of vocabularies that allow for, or assist in, the representation of bibliographic information as linked data; we will mention only a few of the most popular ones here and do not aim at comprehensiveness. The most well-known of these vocabularies is undoubtedly the **Dublin Core (DC)**<sup>4</sup>, which provides data modelers with a number of fundamental classes and properties allowing for the description of relations between bibliographic entities in linked data resources. However, as Peroni and Shotton (2012) point out, the generic nature of the DC vocabulary means that we are seriously restricted in the kinds of bibliographic information which we can use it to express, unless we make use of other vocabularies. Another important linked data bibliographic vocabulary is **FRBR** for which, as we mentioned in the previous section, there exist a number of versions in RDF. The

4 <http://dublincore.org/> [accessed 29/03/18]

**Bibliographic Framework (BIBFRAME)**, on the other hand, was developed as a replacement for the **MARC** standards which had been previously used in the library sector; BIBFRAME was specifically designed with linked data datasets in mind (Casalini 2017). It is interoperable with FRBR, although it uses a slightly different classification hierarchy to FRBR (the BIBFRAME concept *Work* encompasses both FRBR categories *Work* and *Expression*).

The **SPAR** suite of formal ontologies offers users a collection of vocabularies that permit them to model a wide number of different aspects of the semantic publishing and referencing domains in RDF (Peroni 2014). These ontologies have had a wide uptake in both scholarly and industrial domains, having been used by, among others, *Nature*, Europeana, and the Open University. We will single out two of these ontologies in what follows: **FRBR-aligned Bibliographic Ontology (FABiO)** and the **Citation Typing Ontology (CiTO)** (Peroni & Shotton 2012). FABiO, which carries the fact of its FRBR-aligned status in its very name, deals with RDF versions of bibliographic records; it also encompasses a number of other vocabularies, such as DC Terms and SKOS, in addition to a series of newly defined properties intended to facilitate the production of semantically rich bibliographic meta-data. CiTO on the other hand allows for the elaboration of different kinds of rhetorical and factual relationship between two or more bibliographic objects in a network of citations. Here it is important to note that the CiTO model defines a citation as “a conceptual directional link from a citing entity to a cited entity, created by a human performative act of making a citation”. This definition will have important consequences in the development of our own vocabulary for lexical attestations below. In addition, the SPAR **Document Components Ontology (DoCO)** groups together a number of vocabulary terms for describing both the structural and rhetorical makeup of a text; it will also be pertinent in what follows (Constantin et al. 2011).

So much then for the bibliographic and citational side of things, for the time being at least; when it comes to the representation of lexical information in linked data on the other hand, our options are a little bit more restricted. Indeed, the **lemon** model for representing lexical data in RDF (McCrae et. al. 2011), recently published in a updated version as **ontolex-lemon** (McCrae et. al. 2017), has come to take on the status of a de facto standard for representing lexical resources in RDF, and so, in view of its popularity, its dominance of the field as it were, we have chosen to use it as the basis of the work presented in this article. However lemon, unlike TEI-DICT, focuses on capturing the conceptual content of a lexicon; that is, it takes a primarily lexical view of lexical resources, treating them as Works according to the basic FRBR conceptual scheme. Hence there is no conflict here between the demands of fidelity to the text in its lexical view and the text in its editorial and typographical view as there is in TEI; lemon simply prioritizes the former.

Neither lemon nor its successor **ontolex-lemon** make any specific provision for lexical citations, which brings us onto one of the main arguments of our article, namely that there is a necessity for a specific vocabulary (in our case based on lemon) to do just this in the important case (and also likely the majority case when it comes to citations in lexicons) in which a citation is being used to attest a lexical entry or one of its properties. Why not, then, use the ‘citation’ class provided by CiTO to do this? The reason is that there are (at least) two ways of viewing such a citation, both of which we may want to capture separately when modelling a lexicon or a dictionary. One of these views pertains to the lexical/Work view and regards the purpose of the citation, that is, to attest to the existence, in language use, of an association of a given lexical entry with a given linguistic property; the other seems to pertain more to the Expression level or to the editorial view: to a lexicon viewed as a bibliographic entity enmeshed in a web of bibliographically-salient relations with other bibliographic resources. The object properties furnished by the CiTO vocabulary refer to this latter view. Our proposal is to create a RDF-based vocabulary that deals with the level of the former view. We elaborate on this point in the next section through the provision of two detailed examples.

### 3 Two Illustrative Examples

In this section we try and support one of the central claims of this article, namely, that a proper encoding of citations attesting to lexical properties must take into consideration at least two different kinds of conceptual entity: citations and attestations. In the following subsections, 3.1 and 3.2, we present two different examples of lexicographic encoding in which the difference between the two kinds of entity comes out as particularly transparent.

#### 3.1 If at First You Don't Succeed...

Our first example has a strong Dantesque flavor to it, and serves to illustrate how two authoritative lexical resources can completely disagree on the meaning of a citation, even one as famous as the quotation which we discuss in the example<sup>5</sup>. The example centers around the Italian word *riprovare*, which means both ‘to try something again’ (deriving in this instance from the word *provare* ‘to try’ and the prefix *ri-* which adds the sense of repetition), as well as ‘to scold, rebuke’ (in this sense it is cognate with the English verb *reprove*): we are in this case dealing with a pair of homonyms. The popular Italian dictionary *il vocabolario Treccani* (Simone et. al. 2010)<sup>6</sup> lists these as two separate entries: *riprovare*<sup>1</sup> (‘to try again’)<sup>7</sup> and *riprovare*<sup>2</sup> (‘to scold’)<sup>8</sup>; we will refer to the two homonyms in the same way in what follows. The entry for *riprovare*<sup>1</sup> makes an etymological reference to the entry for *provare* in the same dictionary and cites both the motto of the short lived 16th scientific society *L'Accademia del Cimento*, i.e., *provando e riprovando* (‘trying and trying again’), and the *terzina* of the Divine Comedy from which the motto was adapted (‘*Quel sol che pria d'amor mi scaldò 'l petto, / di bella verità m'avea scoperto, / provando e riprovando, il dolce aspetto.*’ Par. III, 1-3)<sup>9</sup> – where however, as the entry itself points out, it means *riprovare*<sup>2</sup>: that is although Dante’s use of *riprovare* is cited in the entry for *riprovare*<sup>1</sup> the entry does not make the claim that this use attests to *riprovare*<sup>1</sup>. The Treccani entry for *riprovare*<sup>2</sup> also cites the same use of *riprovando* in Dante, but in this case the claim is that it does attest to the entry in question. On the other hand *Il Grande Dizionario della Lingua Italiana* (GDLI)<sup>10</sup> (Battaglia 1961) cites both the motto of *L'Accademia del Cimento* and the *terzina* from the Divine Comedy mentioned above under its entry for *riprovare*<sup>1</sup> (which recall has the meaning ‘to try again’) – just as in Treccani – but with the contradictory claim, this time round, that both cited texts do attest to the entry in question, namely *riprovare*<sup>1</sup>.

To summarize, then: we have presented an example in which the same text is cited by two different sources and used to attest to two different homonyms of a word. The following statements describe the current example:

Treccani’s entry for *riprovare*<sup>1</sup> cites Par. III, 1-3.

1. Treccani’s entry for *riprovare*<sup>2</sup> cites Par. III, 1-3.
2. GDLI’s entry for *riprovare*<sup>1</sup> cites Par. III, 1-3.
3. *riprovare*<sup>1</sup> is attested by Par. III, 1-3.
4. *riprovare*<sup>2</sup> is attested by Par. III, 1-3.

<sup>5</sup> The example is dealt with in more detail, and an attempt at an encoding in RDF given in Bellandi et al. (2017).

<sup>6</sup> See also the online version: <http://www.treccani.it/>.

<sup>7</sup> <http://www.treccani.it/vocabolario/riprovare1/> [accessed 29/03/18]

<sup>8</sup> <http://www.treccani.it/vocabolario/riprovare2/> [accessed 29/03/18]

<sup>9</sup> Translated by Longfellow (Alighieri & Longfellow 1867) as ‘That Sun, which erst with love my bosom warmed/ Of beauteous truth had unto me discovered/By proving and reprovng, the sweet aspect.’

<sup>10</sup> The GDLI holds something like the same status and authority in the Italian language as the *Oxford English Dictionary* does in English.

The first three items in the list are true statements about the lexicons which they refer to; they describe the existence of three successful citational speech (‘performative’) acts: speech acts which can be directly represented in RDF using the *cites* object property from CiTO or one of its subproperties. These statements do not deal *directly* with words or their usages, but rather they are concerned with documents or works and the rhetorical/organizational structure pertaining to them. The other two statements, those which I have numbered 4 and 5, instead describe the direct relationship between an item in a lexicon and a text which evidences, or better, attests to its past use. These latter statements are at the level of linguistic facts about words and other lexical entries, that is, at the TEI lexical level<sup>11</sup>. The fourth statement is false but the fifth one is true; however in neither case does this follow from the truth (or falsity) of the first three statements. Both 4 and 5 are only indirectly described by CiTO’s *cites* object property; one of the core aims of the work described in this paper is to describe statements such as 4 and 5 directly in RDF. Note that this example is by no means an atypical one, as this kind of divergence between different lexical resources, for instance, is especially common when it comes to the treatment of word etymologies.

### 3.2 An Anomalous Example

The second example in this section is also our second example taken from the Liddell Scott Jones lexicon (LSJ). This time around the example entry is for the word *ἀνώμαλος*, (*anómalos*) from which the English word *anomalous* derives<sup>12</sup>.

**ἀνώμα^λ-ος** , **ον**, (ἀ- priv., ὁμαλός)

**A.** *uneven, irregular*, “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup.**, **Hp.Aēr.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι” **Id.Ph.238a22**, cf. **Pl.Ti.52e**.

**II.** of conditions, fortune, and the like , “φεῦ τῶν βροτείων ὡς ἀ. τύχαι” **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; “θέα” **Plot.6.7.34**. Adv. “-λως” **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333 D**.

**III.** of persons, *inconsistent, capricious*, “ὁμαλῶς ἀ.” **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; “πίθηκος” **Phryn. Com.20**; “τύχη” **AP10.96**. Adv. “-λως” **Isoc. 9.44**.

**IV.** Gramm., of words *which deviate from a general rule, anomalous*, **Diom.1.327 K.**; but τὸ ἀ. τῆς συντάξεως *diversity* of construction, **A.D.Synt.291.17**. Adv. -λως **Sch.Th.Oxy.853v18**.

The LSJ lists one basic sense for the entry; this single sense is then divided into a number of subsenses. Each subsense is associated with a number of citations and these (in most cases) serve to elaborate further shades of meaning with respect to their corresponding subsenses. In what follows we will concentrate on the third and fourth citations, both of which belong to the first sense, that of ‘uneven, irregular’. The third citation is interesting because of the appearance, in parentheses, of the abbreviation *cj*, which is usually found in critical apparatuses and which stands for the Latin *conicit*, ‘conjectures’. The abbreviation signifies that the citation refers to a critical reconstruction of a work and that there is, therefore, a good chance that the text referred to might not actually attest to the word or sense in question at all: all we can be sure of is that a later scholar in attempting to reconstruct the text from the fragments that were available to him or her made the decision to include the word in his or her conjectural emendation; other lexicographers may decide not to include the citation due to its conjectural status. And this is indeed the case with the third citation, which has not been included by the *Diccionario Griego-Español* (Adrados et. al. 2008), a contemporary ancient Greek-Spanish lexical

<sup>11</sup> It is not entirely clear whether all five statements belong to the TEI lexical view or not – or whether the first three regard the editorial view. Regardless we believe such examples make a strong case for defining a separate attestation relation.

<sup>12</sup> The entry can be found online here in the Perseus published version of the lexicon; <http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Da%3Dnw%2Fmalos> [accessed 29/03/2018].



resource based on the LSJ (along with a number of other Greek lexicons), in its entry for *ἀνὴρμᾶλος*<sup>13</sup>. Once again the example demonstrates the clear conceptual distinction that exists between the performative act of citing a piece of text as evidence – and there can be no reasonable doubt that the 1947 edition of the LSJ did indeed cite Thucydides 7.71 in its entry for *ἀνὴρμᾶλος* – and an instance of a word in a text attesting to a given sense – it is doubtful whether Thucydides did use the word in that sense in the passage in question: that is, the distinction between *citations* and *attestations*. Looking at the fourth citation on the other hand we see that it is prefaced by another abbreviation, *cf*, which stands, this time round, for the Latin term *confer* meaning ‘compare’ and is an instruction to readers to compare the use of the word the cited text (in this case, Aristotle’s *Prior Analytics*) with its use in the text(s) previously cited. It is underspecified whether this kind of citation attests to the same lexical sense/sense/other lexical property, or whether it only provides some interesting contrast or comparison. We cannot therefore always be sure that we are dealing with an *attestation* for, in this case, a sense; we can on the other hand be certain that we are dealing with a *citation*.

In summary then we have an example, one that is, again, by no means an exceptional or a marginal one when it comes to scholarly dictionaries like LSJ, where the idea of two different levels of description, a citational/rhetorical one and a lexical one, arises very naturally. A citation can successfully reference a text with a view to attesting a lexical property even if in reality the text does not attest that property at all: citations can also have different rhetorical purposes other than that of attestation. In the following section we try and model this notion of a lexical attestation in an RDF-based model that extends lemon.

## 4 A Proposal for a Vocabulary for Lexical Attestations

We made a number of observations in the preceding two sections with a view to motivating our definition of a specialized linked data vocabulary for representing lexical attestations. Such a vocabulary, as we hope to have shown, is useful for modeling certain kinds of linguistic claims made via the use of citations: claims which are especially common in scholarly print-born lexicographic resources. We will detail our (minimal) proposal of such a vocabulary, called lemonBib, in Section 4.2. Before that however we turn to the discussion of a modeling issue, which turns out to be very relevant to the modelling of legacy lexicographic resources in RDF, and which also relates to the TEI/FRBR classifications that we mentioned above.

### 4.1 How to Model Different Textual Views on Computational Lexica

As we mentioned above, despite the fact that the distinctions between Work and Expression and between the 0D and 1D/2D views seem to be extremely useful at first sight (and indeed they turn out to be useful in the long run, too), in practice they are often difficult to apply to numerous ambiguous or fuzzy cases. How much sense, for example, does it make to separate out the conceptual Work part of a novel like *Finnegans Wake* from its realization in any specific language<sup>14</sup>? When it comes to lexicographic resources we have to deal with an additional problem that arises from the fact that a lexical entry, as well as being a conceptual component of a lexicographic work, also happens to be a document component of a text in the same way as a chapter, a table of contents, or a bibliography are – and arguably the same

13 The DGE entry for *ἀνὴρμᾶλος* can be consulted online at <http://dge.cchs.csic.es/xdge/%E1%BC%80%CE%BD%E1%BD%BD%CE%BC%CE%B1%CE%BB%CE%BF%CF%82> [accessed 29/03/2018].

14 This is not to say that there haven’t been numerous attempts, as in the case of many other supposedly ‘untranslatable’ works of literature, at translating *Finnegans Wake* in other languages, or that these attempts were entirely fruitless. However the French translation is said to have taken its translator over 30 years to complete, and in the case of the Japanese translation the intellectual toll was so great that the first translator of the work simply disappeared and the second ended up going mad (see <http://www.mhpbooks.com/the-challenge-of-translating-finnegans-wake/> [accessed 29/3/2018]).



is also true of senses in dictionaries. This raises the issue of where we should locate lexical entries and senses in our overall classification, on the grounds of the distinctions that we've already made between Work and Expression, 0D and 1D, and attestations from citations. One option – and this is the strictly purist one – would be to extend the DoCO vocabulary with the classes Lexical Entry and Lexical Sense. Then, for instance, the order of entries in a dictionary – an ordering which, in most cases, has no systematic linguistic significance but is only there to help readers locate the word or entry that they're looking for in a physical copy of the dictionary – would be an ordering of Expression/DoCO Lexical Entries, but not of lemon/Work Lexical Entries. Of course we would want to associate corresponding Work/Expression Lexical Entries, and Lexical Senses with each other in each case. Citations would then belong to the Expression level and attestations to the Work level. Unfortunately, however, this would also lead to a doubling of entries and senses in the RDF version of a lexicon – the kind of prolixity that, conceptual purity notwithstanding, would probably make this quite an unpopular approach. We have therefore decided not to make an explicit distinction between the two views of lexical entries/senses in our example, but to merge the two conceptual levels together in the same entity.

## 4.2 LemonBib

And so it is that we finally come in the present section to the definition of our proposed extension of the ontollex-lemon model for modeling lexical attestations the ontollex-lemon model, *lemonBib*<sup>15</sup>. From our discussion above it is clear that our vocabulary should allow us to do the following:

- Relate attestations to their corresponding citations;
- Relate an attestation to the text which it refers to;
- Relate attestations to other, relevant citations.

We have decided to create a fairly minimal set of properties and classes that meet these requirements in order to make the vocabulary as re-usable as possible. Our proposed modular extension of is based around the definition of the new class Attestation. The idea is that Attestation reifies the relationship between a given lexical element in lemon – whether this is a Lexical Entry, a Lexical Sense, a Lexical Form, or something else – and a bibliographic item that contains a text exemplifying the use of the element in question; we will also be able to relate an Attestation with any citation that is associated with it. We define an object property, *isAttestedBy* relating a lexical element *e* with a member of the class Attestation *a*, with an inverse property *attests* going in the other direction. We also define the object property *involvedInAttestation* between an instance of the CiTO class Citation and Attestation with the inverse property *attestationCitation*. This allows us to relate together the entities which as we have argued in this article belong to two different conceptual levels. The object property *foundIn* relates an attestation with the bibliographic entity in which the attestation can be found. We also define two new data properties. The first, *hasContext*, relates an attestation together with the textual context in which the word is found; the second is the Boolean property *conjectural*, which is true when an attestation is based on a conjectural witness.

We now present a partial encoding of the ἀνὴρ μάλοξ example in diagrammatic form in order to illustrate the features of lemonBib listed above<sup>16</sup>. Our lexical entry has three senses<sup>17</sup>, we will focus on the first sense *sense1* (sense **A** in the original entry above), and on the third attestation of that sense *thuy\_att* (with the citation **Th.7.71**) along with its corresponding citation *thuy\_cit*. Note that *sense1* is

15 The lemonBib vocabulary is available at <http://lari-datasets.ilc.cnr.it/lemonBib>.

16 The example is available in an RDF version at [http://lari-datasets.ilc.cnr.it/ljsj\\_anomalos](http://lari-datasets.ilc.cnr.it/ljsj_anomalos).

17 Note that we have not attempted to describe the hierarchical structure of the senses in our encoding so as not to make the example overly complicated; however a lexicographic extension of ontollex-lemon that will deal with such hierarchies of senses has been proposed and is currently being discussed in the W3C ontollex mailing list (<https://www.w3.org/community/ontollex/>).

linked to `thuy_att` via the `isAttestedBy` that we have defined and that `thuy_att` is linked in turn to the relevant text (here represented using a CTS URN) by the `foundIn` relation. In addition the attestation `thuy_att` is associated with a textual context via the `hasContext` data property and is specified as referring to a conjectured text by the property `conjectural` which is set to `true`. The attestation is then linked to the citation with which it is associated, namely `thuy_cit`, using the `attestationinCit` property and vice versa using the `involvedinAttestation` property. The citation `thuy_cit` is further associated with the sense `sense1` as its citing entity as well as the cited text using object properties defined in CiTO; the type of the citation is also specified using the punned object property `citesAsEvidence`. Note that although we have not included it in the diagram, we can use the `rdfs:seeAlso` property to model the use of the `cf` abbreviation in the text.

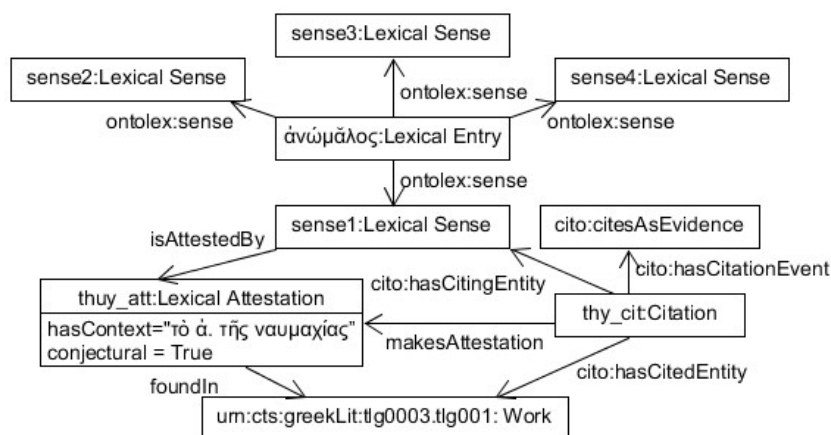


Figure 1: A (partial) encoding of the ἀνώμαλος example

## 5 Conclusion

In this article we have discussed the modeling of citations in lexical linked data resources and proposed an extension of `ontolex-lemon` for dealing with lexical attestations since, as we have argued, this case is not sufficiently covered by pre-existing vocabularies. We have concentrated on examples from traditional, print-based dictionaries because of the wealth of interesting cases that such resources offer. However, we are confident that our vocabulary will be useful for other kinds of resources, at least as a basis for the addition of further classes and properties. In further work we are planning to test the usefulness and the sufficiency of our vocabulary by using it to encode entire lexical resources.

## References

- Adrados, F. R., Elicegui, E. G., & Berenguer, J. A. (2008). *Diccionario griego-español*. Consejo Superior de Investigaciones Científicas.
- Battaglia, S. (1961). *Il Grande dizionario della lingua italiana* di Salvatore Battaglia. UTET, Torino
- Bellandi, A., Boschetti, F., Del Grosso, A. M., Khan, A. F., & Monachini, M. (2017). Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni. *Proceedings of the AIUCD*.
- Boeuf, P. L. (2012, 06). A Strange Model Named FRBROO. *Cataloging & Classification Quarterly*, 50(5-7), 422-438. doi:10.1080/01639374.2012.679222
- Casalini, M. (2017). Implications of BIBFRAME and Linked Data for Libraries and Publishers. “Roll With the Times, or the Times Roll Over You”. doi:10.5703/1288284316449

- Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F. (201). The Document Components Ontology (DoCO). In *Semantic Web – Interoperability, Usability, Applicability*, 7 (2): 167-181. Amsterdam, The Netherlands: IOS Press. <https://doi.org/10.3233/SW-150177>
- Functional requirements for bibliographic records: Final report. (2013). De Gryuter.
- Il vocabolario Treccani. (1997). Istituto della Enciclopedia italiana, Fondata da Giovanni Treccani.
- Liddell, H. G., Scott, R., & Jones, H. S. (1925). *A Greek-English lexicon*. Clarendon Press.
- McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex 2017 conference*, September (pp. 19-21).
- McCrae, J., Spohr, D., & Cimiano, P. (2011, May). Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference* (pp. 245-259). Springer, Berlin, Heidelberg.
- Peroni, S. (2014). The Semantic Publishing and Referencing Ontologies. In *Semantic Web Technologies and Legal Scholarly Publishing*: 121-193. Cham, Switzerland: Springer. [https://doi.org/10.1007/978-3-319-04777-5\\_5](https://doi.org/10.1007/978-3-319-04777-5_5)
- Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 17 (December 2012): 33-43. Amsterdam, The Netherlands: Elsevier. DOI: 10.1016/j.websem.2012.08.001
- Simone, R., Berruto, G., & D'Achille, P. (2010). *Enciclopedia dell'italiano: Il vocabolario Treccani*. Istituto della Enciclopedia italiana.
- Tillett, B. B. (2007). What is FRBR?: A conceptual model for the bibliographic universe. Library of Congress, Cataloging Distribution Service.



# The Sounds of a Dictionary: Description of Onomatopoeic Words in the Academic Dictionary of Contemporary Czech

*Magdalena Kroupová, Barbora Štěpánková, Veronika Vodrážková*

*Czech Language Institute of the Czech Academy of Sciences*

*E-mail: kroupova@ujc.cas.cz, stepankova@ujc.cas.cz, vodrazkova@ujc.cas.cz*

## Abstract

This paper is focused on the description of onomatopoeic interjections and onomatopoeic verbs in a monolingual dictionary. Compared to its predecessors, the emerging monolingual dictionary of Czech (ASSČ) provides more space for the description of all dictionary entries, including onomatopoeic interjections, and their treatment is more detailed. The study concentrates on primary meanings of words related to natural sounds which are imitated, e.g., *bú* (*moo*), *bučet* (*to moo*), as well as on secondary meanings vaguely related to the sound. The paper compares possible ways of treatment of definitions, especially determination of the defining vocabulary for the specific part of the lexicon, the reflection of imitating, and the structure of individual components of the explanation of the meaning (e.g., typical producers, supporting adjectives and adverbs etc.), in addition, proposals of the treatment in the ASSČ are presented. To provide a complex description of onomatopoeic interjections, their syntactic functions and their specific semantic features are discussed.

**Keywords:** Czech, definition, interjection, (lexical) meaning, monolingual dictionary, onomatopoeia, verb

## 1 Introduction

Onomatopoeic words constitute a specific class of words distinguished from others by specific denotation, i.e., an iconically created language sign, that is also connected with their behavior, and hence the characterization of their meaning. This paper deals with treatment of sounds in a monolingual dictionary. It mainly focuses on onomatopoeic interjections (imitations of natural sounds) and verbs (derivatives from interjections) in Czech and their treatment in the Academic Dictionary of Contemporary Czech (hereinafter ASSČ). The ASSČ has been being created in the Department of Contemporary Lexicology and Lexicography of the Czech Language Institute of the Czech Academy of Sciences since 2012. The main sources of material data are the synchronic corpora of written texts of the Czech National Corpus.<sup>1</sup> In the paper, examples of entries created to date (especially entries beginning with letters A-D) are used.

The paper consists of three main parts: 1. a brief introduction of onomatopoeic interjections and verbs; 2. description of the treatment of the definition and analyzing its particular components; 3. characterization of non-onomatopoeic meanings and specific syntactic functions of onomatopoeic words. In addition, several dictionary entries are attached that were produced as a result of the research.

### 1.1 Onomatopoeic Interjections

Interjections are words or phrases that are syntactically independent of their surroundings, they occur as separate utterances. In Czech grammars (e.g., Nekula & Rusínová 1995; Uličný 1986) interjections are usually divided into three categories: emotive interjections (expressing a spontaneous feeling or

<sup>1</sup> Especially the reference representative corpora SYN 2015, and versioned corpora SYN (versions 3 to 6), mainly SYN V4, or Web Corpora Aranea, e.g., Araneum Bohemicum Maximum (for example); see Křen et al. 2015, Křen et al. 2016, Benko 2015.



emotions), appellative interjections (managing social interaction and communication), and onomatopoeic interjections.<sup>2</sup>

Onomatopoeic interjections are words with sound denotation,<sup>3</sup> i.e., they imitate the natural sounds of humans and animals, sounds of nature, sounds produced by tools, mechanisms, machines, etc., and sounds created by humans while using various tools (see, e.g., Nekula & Rusínová 1995: 356–357; Uličný 1986: 247). Finally, baby talk forms a specific category which goes across all the above mentioned types.

### 1.1.1 Headwords

Unlike other parts of speech, interjections, notably onomatopoeic interjections, are typically represented by more than one variant of the headword.<sup>4</sup> The variant forms of the headword differ in vowel and consonant repetition: *br* / *brr* / *brrr* (*brr*); vowel quantity: *čimčarara* / *čimčarára* / *čimčarará* (*chirp*), syllable or word repetition: *cililink* / *cilililink* (*ring*); *brum* / *brum brum* (*grrr*), in variations of vocal composition *bim* / *bim bam* / *bimbam* (*ding-dong*), and so on. Sometimes, the headword of the interjection may be altered by a suffix, e.g., *-y*, *-ky*: *dup* / *dupy* / *dupky* (*stomp*). According to Fidler (2014: 38), these suffixes modify the meaning of the interjection in a specific way. When creating a dictionary the most frequent form or forms (should not be more than four) are chosen as a headword of an entry. This solution reflects the particular instability of the use of onomatopoeic interjections: on the one hand, onomatopoeic interjections are assigned the status of a lexeme (they are lexicalized), on the other hand, users are still aware of the imitative origin of these words.

## 1.2 Onomatopoeic Verbs

Onomatopoeic verbs are verbs motivated by interjections or, when the interjection is only potential or does not exist, by verbs with onomatopoeic roots (see Dokulil 1974), e.g., *bečet* (*to bleat*) – *bé*, *bručet* (*to growl*). According to Dokulil (1974), onomatopoeic imperfective verbs are mainly formed by suffixes *-a-t* (iterative verbs), e.g., *cvrkat*, *cvakat* (*to chirr*, *to click*), *-e-t* (durative verbs), e.g., *bzučet*, *crčet* (*to buzz*, *to gush*). Perfective verbs are formed by suffix *-nou-t* (momentaneous verbs), e.g., *kliknout*, *štěknout* (*to click*, *to bark*) and usually create correlation with iterative imperfective verbs, e.g., *klikat* – *kliknout*, *cvrkat* – *cvrknout*, perfective verbs are also formed by prefixes, e.g., *bečet* – *zabečet*. Less frequent suffixes are *-i-t* and *-ova-t* *chrastit*, *cukrovat* (*to rattle*, *to coo*). In the word list of the ASSČ we find the same kind of proportions of these.

Verbal nouns such as *bučení*, *cvakání* (*mooing*, *clicking*) are treated as forms of verbs in the ASSČ. In our concept, their lexical and grammatical meaning is close to the base verb meaning (see more on this in Kochová and Opavská, 2016: 43–44). Therefore they are not represented by individual entries<sup>5</sup> but they appear as part of grammatical information of verbs, and in relevant cases also in exemplification. In contrast, verbal derivatives made by subtraction (*břinkat* → *břink*) or suffixation (*bzučet* → *bzukoť*) are presented as individual entries.

2 In *Akademická gramatika spisovné češtiny* grammar, Vondráček (2013: 535) mentions Commands as the fourth type, we classify them within the Appellative interjections category.

3 For more details, see part 2.4 Imitation.

4 The entries of the ASSČ are selected from an automatically generated word list mainly based on the frequency criterion and the criterion of the commonness of their usage. The list draws on a set of balanced synchronic corpora of written texts from the Czech National Corpus SYN 2000, SYN 2010 etc. See Kochová & Opavská (2016), and Kochová et al. (2014).

5 Verbal nouns combine features of two word forms, nouns and verbs (Ružička 1966: 504). They have morphological features and syntactic functions of nouns, but their grammatical meaning is verbal, they also express verbal aspect (Petr 1986: 135). Therefore, they can either be defined as a noun or a grammatical form of verb. The defining feature of the lexicographic processing is their lexical meaning.

## 2 Definitions

The definition of a sound is based on resemblance (i.e., the basis is given by the ostensive definition), in the way similar to the meaning explanation of a color (Blatná 1995: 79), e.g.,

- (1) *bílý* ‘mající barvu sněhu, mléka ap.’ (*white* ‘having the color of snow, milk etc.’)<sup>6</sup>
- (2) *příst* ‘vydávát zvuky pod. zvukům kolovratu při předení’ (*to purr* ‘to emit sounds similar to those of a working spinning wheel’) (SSČ)
- (3) *cvakat* ‘vydávát krátké, ostré zvuky znějící jako cvak’ (*to click* ‘to emit short sharp sounds which sound like *click*’)

In some cases explanations with paraphrases are preferable, e.g.,

- (4) *cvrlikat* ‘(o drobných ptácích) vydávát jemné zvuky s rychle se měnící výškou tónu’ (*to twitter* ‘(of small birds) to emit soft sounds with a quickly changing pitch’)

The central part of the definition consists of a sound describing word. In the definition of interjections, the word *zvuk* (*sound*) is frequently used as the genus proximum, whereas in the definition of verbs, the genus is described in the verb-and-sound format (see below for the definitional patterns).

The differentia specifica, as a part of the definition which distinguishes the word from the generic term by describing its characteristic features, is usually represented by adjectives or an adjective phrase. In the ASSČ, the use of metaphorical or rarely used adjectives is avoided, and thus less metaphorical and more transparent adjectives are preferred, e.g., we prefer *hluboký* (*deep*) to *temný* (*dark*), *ostrý* (*sharp*) to *drnčivý* (*rattling*). The combination of unambiguous adjectives (and adverbs) is considered more comprehensible for users.

### 2.1 Producers

In accordance with the requirement of E. Veldi (1994) that an entry for an onomatopoeic word should make explicitly clear what kinds of objects or beings produce the sound that the word is meant to represent, special consideration is given to producers of sounds in the ASSČ. Firstly, a more precise description of producers is used compared to older dictionaries (e.g., small birds, big dog)<sup>7</sup>, and secondly, producers are introduced in a more consistent manner.

Representation of the *producer* is provided in several ways:

For interjections expressing animal sounds, the producer is the second basic point of the definition, e.g.,

- (5) *bé* 1. ‘táhlý vyražený zvuk vydávaný ovce’ (*baa* ‘prolonged blurted out sound made by sheep’)

The sounds of tools, machines, etc. are processed similarly, e.g.,

- (6) *brnk* ‘zvuk struny při doteku, úderu prstem’ (*twang* ‘sound of a string when it is touched or hit by a finger’)

A human producer is usually described less explicitly, e.g.,

- (7) *bé* 2. hlasitý, zprav. dětský pláč, naříkání (*boo-hoo* ‘loud, usually children’s crying, wailing’)

For verbs that have a producer with restricted collocability, a specific description using a semantic comment before the definition is used,<sup>8</sup> e.g.,

<sup>6</sup> Unless specified otherwise, the examples used here come from the ASSČ.

<sup>7</sup> The use of grammatical number (singular vs. plural) of the noun depends on the frequency, the occurrence in collocations, and on the dictionary tradition (convention), too.

<sup>8</sup> Cf. collocator (Atkins & Rundell 2008: 217-218).

- (8) *cvrlikat* ‘(o drobných ptácích) vydávat krátké zvuky s rychle se měnící výškou tónu’ (*to twitter* ‘(of small birds) to emit short sounds with a quickly changing pitch’)

For other onomatopoeic verbs, the producer can constitute the *differentia specifica*, e.g.,

- (9) *cinknout* ‘vydat krátký znělý zvonivý zvuk při nárazu kovových nebo skleněných předmětů’ (*to clink* ‘to emit a short resonant ringing sound by striking metal or glass objects’)

The resemblance to the prototypical producer of a sound must be sufficiently recognizable for contemporary users. For example, the definition of the verb *příst* (*to purr*), mentioned at the beginning of the discussion: ‘vydávat zvuky pod. zvukům kolovratu při předení’ (‘to emit sounds similar to sounds of the spinning wheel while spinning’) (SSČ) does not fulfill this requirement. The definition is etymologically relevant – the primary lexical meaning of the Czech verb *příst* is ‘to make, produce threads or yarn’, but for contemporary dictionary users it is not sufficiently transparent, because they are not familiar with the spinning wheel and the process of thread making anymore. It is thus a task for dictionary authors to describe the sound in a way that is appropriate for modern users, e.g.,

- (10) *příst* ‘(o kočkách) vydávat tlumené hluboké vibrující zvuky a vyjadřovat tak spokojenost’ (*to purr* ‘(of cats) to emit a muffled deep vibratory sounds expressing contentment’)

If the description of a sound expressed in a verb, adjective, adverb, or noun is too wide and vague, an interjection is used to specify the sound. (The use of the interjection also captures the word-forming relation between a motivating word and a derivative.) For example, the interjections *bú*, and *cvak* provide specification of sound qualities ‘znějící jako bú’ (‘sounds like moo’) and ‘znějící jako cvak’ (‘sounds like click’). The use of an interjection is usually combined with the description of the sound quality, e.g.,

- (11) *bučet* ‘vydávat táhlé hluboké zvuky znějící jako bú’ (*to moo* ‘to emit a deep prolonged sounds that sounds like bú’)  
 (12) *cvakat* ‘způsobovat krátké, ostré zvuky znějící jako cvak’ (*to click* ‘to cause short, sharp sounds that sound like click’)

The use of interjections in definitions depends on several conditions: firstly, the interjection must be included in the word list of the dictionary, secondly, the interjection must be relevant for the description of the verb. The interjection *cukrú* used by previous dictionaries to define the verb *cukrovat* is not a dictionary entry in the ASSČ, and therefore the definition is created in another way.

- (13) *cukrovat* ‘(o hrdličkách) vydávat opakované jemné klokotavé zvuky’ (*to coo* ‘(of dove) to emit repeated soft murmuring sounds’)

This paper concentrates on onomatopoeic interjections and verbs. However, the principles presented above with regard to the description are also relevant for their derivatives, such as an adjective *bzučivý* (*buzzing*), adverb *bzučivě* (*buzzingly*), noun *bzukot* (*buzz, buzzing*), *klik* (*click*), etc.

## 2.2 The Meaning of Onomatopoeic Interjections

In contemporary monolingual dictionaries of Czech, the explanation of the meaning of interjections is expressed in two ways – a) by using a defining verb – for example *vyjadřuje* (*expresses*), *naznačuje* (*indicates*), *označuje* (*denotes*): *bzz* ‘denotes the buzzing sound made by flying bees, flies etc.’ (SSJČ); b) by means of the genus proximum (similarly to nouns) – *pozdrav* (*greeting*), *souhlas* (*agreement*) etc.: *ahoj* ‘a greeting used especially by young people (SSČ). In the ASSČ, interjections are defined uniformly by the genus proximum, and hence by a defining noun.

As mentioned above, the most frequent defining noun is *zvuk* (*sound*), which is used for most types of onomatopoeic interjections, e.g., *baf* ‘krátký hluboký zvuk vydávaný velkým psem’ (*woof* ‘a short deep sound that a big dog makes’). Onomatopoeic interjections produced by humans are exceptional, and these words often have specific designation, e.g., *pláč*, *nařikání* (*weeping*, *moaning*), or in addition to a sound they also express a feeling, movement and so on simultaneously, e.g., *brr* ‘vyjádření pocitu zimy, nelibosti, zhnusení ap., často doprovázené otřesením se’ (*brrr* ‘an expression of feeling cold, dislike, disgust, etc., often accompanied by shuddering’).

### 2.3 The Meaning of Onomatopoeic Verbs

In the definition of verbs, the perspective of sound production is also relevant besides the sound itself. Generally, the meaning can be divided into three types, which correspond to different perspectives:<sup>9</sup>

A) ‘vydávat/vydat nějaký zvuk’<sup>10</sup> (‘to emit a sound’) – used when a producer produces the sound by himself. This type is mainly represented by sounds produced by an animal, machine, etc.: *bzučet* (*to buzz*), *bručet* (*to growl*), *čříkat* (*to chirrup*), *bouchat* (*to bang*), *cvakat* (*to click*), e.g.,

- (14) *bzučet* ‘(o hmyzu, o přístrojích ap.) vydávat znělý sykavý zvuk znějící jako bzz’ (*to buzz* ‘(of insects, machines, etc.) to emit a resonant, hissing sound that sounds like *bzz*’)
- (15) *cvaknout* ‘vydat krátký ostrý zvuk znějící jako cvak’ (*to click* ‘to emit a short sharp sound that sounds like *click*’)

or by production of nonverbal human sounds, e.g.,

- (16) *broukat* ‘(o malých dětech) vydávat neartikulované zvuky a vyjadřovat tak spokojenost’ (*to gurgle* ‘(of small children) to emit inarticulated sounds and to express contentment by means of that’)

or involuntarily emitted sounds, e.g.,

- (17) *cinknout* ‘vydat krátký znělý zvonivý zvuk při nárazu kovových nebo skleněných předmětů’ (*to clink* ‘to emit a short resonant ringing sound by striking of metal or glass objects’)

B) ‘způsobovat/způsobit nějaký zvuk’ (‘to cause a sound’) – used when a sound is emitted by an instrument, tool, device, etc. The producer of the sound is usually something in the role of an instrument, or someone using the instrument, e.g.,

- (18) *cinknout* ‘způsobit nárazem kovových nebo skleněných předmětů krátký znělý zvonivý zvuk’ (*to clink* ‘to cause a short resonant ringing sound by hitting metal or glass objects’)

The cumulation of a sound and an action or a motion often occurs in these cases. Example 18 represents a situation when a sound is primarily a result of an action or a motion. In many cases, furthermore, the sound has secondary importance, i.e., it is a side effect of an action. The explanation of the meaning is modified to ‘to do something accompanied with a sound’, the genus proximum is a verb of an action or a motion, e.g.,

- (19) *buchnout* ‘spadnout (a tím způsobit dutý zvuk)’ (*to bang* ‘to fall down (causing a hollow sound)’)

9 The definitions in the ASSČ are more uniform than those in previous Czech dictionaries. For example SSJČ uses genus proximum *ozývat se / ozvat se* (*to resound*) for type (A) in our scale, e.g., *kvičet* ‘(o zvířeti) ozývat se pronikavým, vysokým hlasem’ (*to squeal* ‘(of an animal) to sound with a piercing high voice’) or *vydávat/vydat*, e.g., *chrápat* ‘vydávat při spaní chrčivý zvuk’ (*to snore* ‘to emit raspy sound while sleeping’). Genus proximum *ozývat se / ozvat se* is also used in type (C), e.g., *bublat* ‘temně, tlumeně, přerývaně se ozývat’ (*to bubble* ‘to resound deeply, muffledly, spasmodically’). In the ASSČ genus proximum *ozývat se / ozvat se* is completely eliminated from definitions of onomatopoeic verbs for two main reasons: casus phrase or adverbial phrase with instrumental (*ozývat se / ozvat se čím*) is outdated in contemporary Czech, and primarily meaning of the verbs is ‘let sb know, respond’.

10 There are two variants of definitions due to distinction of imperfective and perfective verbs.

In our concept, a similar solution is appropriate for etymologically onomatopoeic verbs when the distinctive features of sounds are more or less irrelevant, e.g.,

(20) *bacit* ‘úderem zasáhnout’ (to hit ‘to strike with a blow’)

(21) *cákat* ‘rozstříkovat prudce tekutinu’ (to splash ‘to spray liquid fiercely around’)

C) ‘znít/zaznít nějakým zvukem’ (‘to sound with a sound’) used for a resonated sound, when the subject is not the producer of the sound, but the sound itself is, e.g., *hudba duněla městem* (music was rumbling around the town) or space, place, where the sound resonates *dlažba duněla pod kopyty koní* (the pavement was rumbling under horse’s hooves), e.g.,

(22) *dunět* ‘znít silným hlubokým dutým zvukem’ (to rumble ‘to sound with a strong deep hollow sound’)

## 2.4 Imitation

The relation between an onomatopoeic interjection and its explanation is quite specific, as the definition should describe a noise which does not consist of human speech sounds (unlike, e.g., interjections of social communication (*bravo*) etc.) and its form is always a kind of an imitation, different in each language,<sup>11</sup> in contrast to nouns, whose relation between the definiendum and the definiens is usually not reflected, probably because of the arbitrariness of the sign.<sup>12</sup>

In establishing the principles of making a dictionary, the authors have to decide whether the imitation (as one of the components of the lexical meaning) of onomatopoeic interjections will be reflected in the dictionary, and how it will be presented. There are two basic ways of defining in this context:

1. The imitation is explicitly provided by using defining words or phrases, such as *imitating of / imitates*:  
*hav, haf* ‘napodobňuje brechot psa’ (*woof* ‘imitates bark of a dog’) (SSSJ)
2. The imitation is not explicitly mentioned:  
*woof* ‘a low gruff sound typically produced by a dog’ (Merriam-Webster Dictionary),  
*woof* ‘the barking sound made by a dog’ (Oxford Dictionary).

As already indicated, onomatopoeic interjections are regarded as lexemes (lexicalized units) in the ASSČ. We consider that their imitative origin does not distort their sign integrity (i.e., icons are not completely different from other signs), and therefore the imitation is not emphasized in the ASSČ.

## 3 Specific Use of Onomatopoeic Words

### 3.1 Syntactic Functions of Onomatopoeic Interjections

Even though they usually have no syntactic role in the sentence, in several cases onomatopoeic interjections can be incorporated into the syntactic structure of a sentence. Specific semantic shifts are connected with this function. The most frequent syntactic roles are as follows:

1. An onomatopoeic interjection is modified by an adjective and plays the role of a noun as a subject or an object of a sentence. In this case, the interjection does not express any gender, and the morphologically neutral gender form of the neuter is used for the specifying adjective. If it is typical for an interjection, the use is illustrated by an example in the exemplification of the word.

<sup>11</sup> On comparing of onomatopoeia in different languages see, e.g., F. Kopečný (1957).

<sup>12</sup> Cf. Saussure (1989).



- (23) *buch* ... ve funkci podstatného jména: *když se dveře zaklapnou, udělají hlasité buch* (...in the noun function: *when the door closes, it makes a loud slam*)

2. Onomatopoeic interjections representing a motion and a sound simultaneously or interjections expressing a motion making sound can be used as the predicate of a sentence. For example, *kráva bác na zem* (cow *bác to the ground*) *bác* means *spadnout (to fall)* in this case, interpreted as grammatical categories of singular, past perfect, feminine: *spadla (she fell)* (a cow is a feminine in Czech). In the ASSČ this function is described as a separate meaning:

- (24) *bác* 2. ve funkci slovesa: *upadnout, spadnout* (in the verb function: *to fall*)

These interjections also appear in specific constructions typical for baby talk, consisting of the verb “udělat” (to do) and an interjection, e.g., *udělat bác (to fall)*.

### 3.2 The Secondary Meaning of Onomatopoeic Words

In the semantic structure of onomatopoeic words, particular changes based on metaphoric or metonymic shifts occur. We talk about the so-called semantic derivations. If the shift is lexicalized, it is represented as the secondary meaning in the dictionary, e.g.,

- (25) *cvaknout (to click)* 1. ‘způsobit krátký ostrý zvuk znějící jako cvak’ (‘to make a short sharp sound that sounds like cvak’) → 2. ‘udělat fotografii, vyfotit’ (‘to take a photograph’) – due to the characteristic clicking sound of a shutter when a photograph is being taken.

It is possible to identify certain regularity in semantic derivations, e.g.,

a) expressing motion or activity:

- (26) *bimbat* 1. ‘(o zvonu) opakovaně vydávat jednotlivé hluboké zvuky’ (*to ding* ‘(of a bell) to emit individual deep sounds repeatedly’) → 2. ‘opakovaně volně pohybovat částí těla ze strany na stranu nebo nahoru a dolů’ (*to dangle* ‘to move part of one’s body from one side to the other or up and down repeatedly and loosely’)

b) communication, e.g., speaking, singing, etc.:

- (27) *bručet (to growl)* 1. ‘(o medvědech, o strojích ap.) vydávat hluboké táhlé zvuky’ (‘of bears, machines) to make deep prolonged sounds’) → 2. ‘projevovat (slovy) nespokojenost’ (‘to express discontent (by words)’)

c) expressing abstraction:

- (28) *bác (bang)* 1. ‘zvuk při úderu, pádu, výstřelu’ (‘a sound of a hit, fall, shot’) → 2. ‘překvapení, údiv vyvolaný náhlou, zprav. nepříjemnou změnou’ (‘surprise, astonishment induced by a sudden, usually unpleasant change’)

Polysemy of some onomatopoeic words can be caused by metonymic or metaphoric shifts, as well as by random parallel origin of the meanings (it could then be regarded as homonymy), e.g., words expressing emotions, *verba dicendi*:

- (29) *bečet (to bleat)* 1. ‘(o ovcích) vydávat táhlé vyražené zvuky znějící jako bé’ (‘(of sheep) to emit prolonged bursted sounds that sounds like bé’) → 2. ‘expr. slzami a vzlykáním projevovat zármutek, bolest, rozrušení ap.’ (‘to express sorrow, pain, agitation etc. by tears and sobbing’)
- (30) *broukat (to gurgle)* 1. ‘(o malých dětech) vydávat neartikulované zvuky a vyjadřovat tak spokojenost’ (‘(of small children) to emit inarticulated sounds and to express contentment by means of that’) → 2. ‘nezřetelně, slabě zpívat’ (‘to sing weakly, inarticulately’)
- (31) *bú (moo)* 1. ‘táhlý hluboký zvuk vydávaný kravami’ (‘a prolonged deep sound emitted by cows’) → 2. ‘dětský pláč, vzlykání’ (‘children’s crying, sobbing’)

## 4 Conclusion

The dictionary treatment of onomatopoeic words requires a specific approach to the knowledge concerning their meaning, functions, and use. A thorough, detailed, and clear definition is considered necessary to make an unambiguous description of onomatopoeic words. Unfortunately, this is still rather insufficient in existing monolingual dictionaries.

While we have only dealt with onomatopoeic interjections and verbs in this paper, their treatment serves as a template for treatment of their derivatives (adjectives, adverbs, and nouns).

Onomatopoeic interjections are considered fundamental words, onomatopoeic verbs are thus, in principle, their derivatives, with which the structure of dictionary entries definition complies. Onomatopoeic interjections are regarded as independent, established signs, and lexicalized units, and therefore their imitative origin is not provided in definitions explicitly. In the definition, the resemblance of the sound is described, as well as its typical producer.

In the definition of an onomatopoeic verb, the imitation is provided explicitly to indicate its relation to the source interjection (the interjection itself can be included, too). The typical producer is also mentioned, either in the definition, or in the semantic comment.

## References

- Akademický slovník současné češtiny*. (2017). Praha: ÚJČ AV ČR. Accessed at: <http://www.slovníkcestiny.cz/> [07/02/2018].
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Benko, V. (2015) *Araneum Bohemicum Maximum, verze 15.04*. Praha: Ústav Českého národního korpusu FF UK. Accessed at: <http://www.korpus.cz> [17/01/2018].
- Blatná, R. (1995). Metajazyk v lexikografii. In F. Čermák, R. Blatná (eds.) *Manuál lexikografie*. Praha: H&H, pp. 72-89.
- Dokulil, M. (1974). K jednomu typu slovesných pojmenování. In *Naše řeč*, 57(2), pp. 57-64.
- Fidler Ueda, M. (2010). *Onomatopoeia in Czech*. Bloomington, Indiana: Slavica.
- Kochová, P., Opavská, Z. (eds.) (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Praha: ÚJČ AV ČR.
- Kochová, P., Opavská, Z. & Holcová Habrová, M. (2014). At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project). In A. Abel, C. Vettori, N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen. pp. 1145-1151.
- Kopečný, F. (1957). Slavistický příspěvek k problému tzv. elementární příbuznosti. In *Ezikovedski izsledvanija v čest na akademik Stepan Mladenov*, pp. 363-387.
- Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P. & Zasina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Accessed at: <http://www.korpus.cz> [10/12/2017].
- Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P. & Zasina, A. (2016). *Corpus SYN, version 4 from 16. 9. 2016*. Praha: Ústav Českého národního korpusu FF UK. Accessed at: <http://www.korpus.cz> [18/12/2017].
- Merriam-Webster Dictionary Online*. Accessed at: <https://www.merriam-webster.com/> [14/01/2018].
- Nekula, M., Rusínová, Z. (1995). Citoslovce. In P. Karlík, M. Nekula, Z. Rusínová (eds.) *Příruční mluvnice češtiny*. Praha: NLN, pp. 356-358.
- Oxford Dictionary Online*. Accessed at: <https://en.oxforddictionaries.com/> [14/01/2018].
- Petr, J. et al. (1986). *Mluvnice češtiny. 2. Tvarosloví*. Praha: Academia, pp. 239-250.

- Ružička, J. et al. (1966). *Morfológia slovenského jazyka*. Bratislava: Vydavateľstvo Slovenskej akadémie vied.
- Saussure, F. de (1989). *Kurz obecné lingvistiky*. Praha: Academia.
- Slovník spisovné češtiny pro školu a veřejnost*. (1994). Praha: Academia.
- Slovník spisovného jazyka českého*. (1960-1971). Praha: Academia.
- Slovník súčasného slovenského jazyka*. (2006-2015). Bratislava: Veda.
- Vondráček, M. (2013). Citoslovce. In F. Štícha (ed.) *Akademická gramatika spisovné češtiny*. Praha: Academia, pp. 532-535.

## Acknowledgements

This work has been supported by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project CZ.02.1.01/0.0/0.0/16\_013/0001781).

## Abbreviations

- ASSČ = Akademický slovník současné češtiny
- SSČ = Slovník spisovné češtiny pro školu a veřejnost
- SSJČ = Slovník spisovného jazyka českého
- SSSJ = Slovník súčasného slovenského jazyka

## Appendix

bú, búú, búúú citosl.

1. táhlý hluboký zvuk vydávaný kravami: *kravička dělá bú; kráva kývne hlavou, přežvýkne a zabučí: „Búú.“* [moo]
2. dětský pláč, vzlykání: *skočil mámě do náruče a bulí „bú, búú“; Búúú, já jsem se ztratil. – Neplač, maminku najdeme.* [boo-hoo]
3. zvolání vyjadřující nesouhlas, nespokojenost, zklamání; syn. fuj: *projev provází výkřiky nevole „fuj“ a „búúú“; diváci nečekají na mínění recenzentů a dají své pocity najevo voláním Bravo! nebo Búú!* [boo]

bučet (3. j., 3. mn. bučí, rozk. (ne)buč, čin. bučel, podst. jm. bučení) ned.

1. (kdo || ~) (o kravách) vydávat táhlé hluboké zvuky znějící jako bú: *krávy v chlévě bučely; ticho narušovalo jen bučení dobytka* [to moo]
2. expr. (kdo || ~; na koho) (o lidech) vydávat táhlé zvuky znějící jako bú a vyjadřovat tak nesouhlas, nespokojenost, zklamání ap.: *fanoušci během zápasu nespokojeně bučeli; diváci při jeho projevu pískali a bučeli; všichni na něj bučeli a křičeli, že fixloval; když režisér po premiéře vyšel na jeviště, publikum ho přivítalo bučením* [to boo]

dok. k 1 → zabučet



# Comparing Orthographies in Space and Time through Lexicographic Resources

**Christian-Emil Smith Ore<sup>1</sup>, Oddrun Grønvik<sup>2</sup>**

<sup>1</sup>University of Oslo, <sup>2</sup>University of Bergen

E-mail: [c.e.s.ore@iln.uio.no](mailto:c.e.s.ore@iln.uio.no), [Oddrun.Gronvik@uib.no](mailto:Oddrun.Gronvik@uib.no)

## Abstract

Many languages require an improved factual basis to facilitate computer-supported analysis of language variation and diachronic change. The material collections for the scholarly dictionaries of Norway serve as a platform for exploring the development and variation of Bokmål, the Norwegian written standard derived from Danish and modified towards the Norwegian vernacular through orthographic reforms that took place from 1901 to 2005. The development of modern Bokmål through usage should be analyzed by comparing corpora from different periods, lemmatized according to the then current orthography. This means building full form registers from time-bound orthographies. This plan is in process through digitizing orthographic dictionaries for Bokmål. The dictionaries are coordinated through the Dictionary Hotel, the electronic repository for retro digitized dictionaries and dialect collections at the Norwegian Language Collections, Bergen. At the lexical item level Bokmål and Nynorsk resources are coordinated through the Meta Dictionary, an electronic registry for the Norwegian lexicon. A common entry requires full identity in one headword form plus part-of-speech (POS). Preliminary results identify a core vocabulary for Bokmål of 6,900 lexical items, unchanged since 1938. More than 75,000 Meta Dictionary entries have a common identical form plus POS for Bokmål and Nynorsk. These numbers will increase when the Bokmål additions to the Meta Dictionary are quality controlled.

**Keywords:** dictionary, lexical item, full form register, computer assisted language analysis, corpus, lemmatizer, synchronic variation, diachronic change

## 1 Background, Problem, Task

The motivation behind the great historical dictionaries of the 19<sup>th</sup> and 20<sup>th</sup> centuries was to provide a full scholarly description of the single lexical item<sup>1</sup> through the dictionary entry, and through the entirety of entries, a scholarly description of whole languages through time. Their purpose was and is to replace speculation with facts, a task which overall has been completed successfully for the languages covered. The scholarly dictionaries have provided their language communities with reliable lists of lexical items, traced and identified through time, their formal qualities identified, their senses documented and ordered.

These laborious projects have been facilitated in the last decades through digitization, and many of them have been completed. In the process several lexicographical resources have been created, many of which have wider uses than dictionary production. Corpora, full form registers, spelling programs and syntactic taggers all stem from empirical linguistics and depend on lexicography to make sense of it by linking the sign – the lexical item – with sense description.

However, many languages still require a historical and comparative description to be properly documented. A case in point is Bokmål, the written standard of Norwegian which is used by the majority of the population.

<sup>1</sup> The term “lexical item” is used as defined by Atkins and Rundell (2008:163 ff).



In a European context, Norwegian is a special case. Spoken Norwegian must today be considered one language, represented through a variety of non-standard dialects, and whether a spoken standard exists is an issue of debate. There are two written standards, Bokmål and Nynorsk, one stemming from Danish as spoken in Norway, the other from a comparative analysis of 19<sup>th</sup> century Norwegian dialects. The written standards have been modified, and in many respects brought closer to one another, through a series of orthographic reforms (from 1901 until 2012). The orthographic changes have affected all aspects of language, from sign inventory (*aa* > *å* in 1917) to syntax. The history of orthographic reform is well documented. There is however limited agreement on how the different orthographic reforms have affected actual usage, especially concerning the majority language, Bokmål. The discussion on how close or separate the two standards are, and how their separate and joint development has expressed itself in usage, remains ideologically slanted and easily gets emotional (cf. Vikør 2001: 53 f.).

Moving this discussion forward requires an improved factual base. The only way of finding out exactly how present-day Norwegian, especially Bokmål, has developed, is to analyze and compare Bokmål corpora from different periods in the 20<sup>th</sup> century. This analysis requires lemmatizers providing exact coverage of the orthography for both standard languages during the selected periods. Corpora, and the comparison of corpora, will show usage and changes in this. The orthography dictionaries compiled to put the orthographic reforms into practice show what the language reformers believed and hoped for. If lemmatizers are made on the basis of contemporaneous spellers, they will identify the vocabulary used in the corpora, but they will also show non-occurrence and divergence between the corpora and the orthography dictionaries, and ultimately contribute to our general knowledge of standardization processes.

The issue of improving the factual base for exploring the development of Bokmål has gained relevance because *Bokmålsordboka* and *Nynorskordboka*, the two Norwegian general-purpose monolingual dictionaries, are to be revised through a five-year project at the University of Bergen 2018 - 2023.<sup>2</sup> The dictionaries were first edited 1974-1986, and content revision has since then been minimal. Both dictionaries qualify as scholarly dictionaries, with evidence from the Norwegian Language Collections. *Bokmålsordboka* will be expanded to match *Nynorskordboka*, from roughly 65,000 to 100,000 entries. Corpus-based linguistic studies covering the whole of the 20<sup>th</sup> century will be an important support discipline. The entries of the dictionaries will be generated from the material index of the Language Collections and the Meta Dictionary (Stortingsmelding 1 (2017–2018)), as also done for the entries of *Norsk Ordbok* (Grønvik & Ore 2013: 254).

The largest historical corpus of Norwegian text is being built by the National Library. The Library is in the process of digitizing its entire collection of text printed in Norway. The resulting corpus at present comprises more than 1 million volumes of books, newspapers and so on, and currently has more than 50 billion tokens. All texts are linked to the National Bibliography with its metadata. The digitization process consists of scanning and automatic OCR. However, the large changes in Norwegian orthography in the 20<sup>th</sup> century cause problems for the OCR-process and the lemmatization. The only full form registries with broad coverage, the Norwegian Word Banks (Bokmål & Nynorsk), have a relatively short temporal coverage. Both word banks document the current orthography from about 1990 and until today, while the text to be analyzed goes back to the 18<sup>th</sup> century. This leads to mismatches, like search finds from about 1800 for the base form “telephone”. For reliable results, lemmatizers for Norwegian in previous orthographies are needed for both the written standards of the language. The question is how best to develop them.

2 The Norwegian Language Collections, comprising lexicography, dialectology and onomastics, were moved from the University of Oslo to the University of Bergen in 2016, and are now a unit at the University Library.

## 2 Orthographies, Spellers and Digitization

There are two possible approaches to isolating orthographies within given periods: 1) Analyze a (selected, dated) corpus in terms of frequencies and distribution, and propose a list of lexical items on the basis of this analysis. 2) Start with the records of the orthographic standard of the selected period, as found in authorized (school) dictionaries, and then use grammars from the same period as a source for establishing full form schemas.

Either approach involves a lot of work, and both are useful; the corpus analysis will give the register and frequencies of forms to be identified (including plenty of homographs), while full form registers based on spellers and grammars will represent what was thought essential vocabulary at the time, in the form valid at the time. Without a valid full form register, automatic identification of word forms will be impossible.

*Bokmålsordboka* and *Nynorskordboka*, published together in 1986, were the first general purpose defining dictionaries for modern Norwegian. Earlier normative information on orthography is found in printed spellers, mostly for school, and in the official reports on suggested orthographic reforms. The most important reports cover the orthographic reforms of 1917, 1938, 1959, 2005 (Bokmål) and 2012 (Nynorsk). These are the natural pivot points for comparing before and after. For Nynorsk, ample and dated materials are present in the lexical index of the Language Collections, the Meta Dictionary, cf. section 3.1 ff. Additions to the Meta Dictionary at present therefore focus on covering the development of Bokmål.

### 2.1 Orthographic dictionaries and school spellers

Orthographic dictionaries are made to inform the public about the orthography of headwords and their inflected forms. In the Norwegian printed orthographic dictionaries and spellers of the 20<sup>th</sup> century, this information is given in a very compressed form. The first headword in base form is given in full, while derivations can be abbreviated to the word ending. Rows of compounds are often nested. Additional information, such as inflected forms, POS, sense, usage, multi-word expressions (MWEs), etymology, usage conventions are omitted, unless something more than the headword is needed to identify the headword to the user. Instances of all of the categories above can be found in school spellers, most often in an abbreviated form.

Judging by the look of the spellers, the user must be assumed to be an experienced mother tongue user, presumably a teacher, and getting children to understand and interpret the spellers correctly must have been part of their work.

The Norwegian school spellers are the work of individual compilers, often philologists with teaching experience. Unlike Sweden and Denmark, no central agency provided an authoritative orthographic dictionary. Each of the frequent orthographic reforms in the 20th century was based on recommendations from the Ministry of Church and School Affairs. The underlying reports and suggestions discuss principles and give examples. (cf. *Den nye rettskrivning* 1917: 24-25). Editors of school spellers were then tasked with fleshing out the recommendations and examples as best they could, by listing lexical items plus essential additional information. A typical example of a school speller is shown in Figure 1. School spellers needed a stamp of approval from the Ministry of Education before they could be used in school.

The first guidelines for orthographic dictionaries for school use were set up as house rules by the Norwegian Language Committee in the late 1960s. School spellers from before 1980 are geared towards saving space, omitting information that can be implied or assumed to be known by an adult literate

person. Some guidance concerning lexicographic conventions is normally found in the front matter, but no one would call these spellers explicit or learner-friendly. Conventions are expressed differently from speller to speller, and consistency levels vary.

## 2.2 Retro digitizing orthographic dictionaries

Retro digitizing 20<sup>th</sup> century orthographic dictionaries for Norwegian (Bokmål and Nynorsk) almost always involves interpretation doubts, because the original text is highly compressed and the punctuation ambiguous.

The sample shown below in Figure 1 is typical. A striking feature is the use of the typographical marker “/”. This means that the string in front of the slash can be added to following strings starting with a hyphen, and the result will be a meaningful word form. The line “akt/e; -else, -en” is to be read as a list of three lexical items: *akt*, *akte*, *aktelse*. But a lexical item with a base form *akten* does not exist. The information that “-en” in this case is to be read as information on noun inflection for the preceding lexical item *aktelse*, indicating the masculine gender, can be found in the front matter. For *akt* (noun) and *akte* (verb) no POS information is given.

The slash does not mean that the string in front represents a lexical item. The line “aksj/e, -en, -onær” does not claim that there is a lexical item *aksj* – there is not. It means that “aksj/” can be added to “-onær” to render *aksjonær*. The “-en” in between is meant to say that *aksje* is a noun with the masculine gender.

5 ]	A		[ akv
<u>A</u>	aeroplan, -er	aksent	allegori/sk
å (fem à seks)	affeksjon	aksept/ere, -ert	alle/gretto, -gro
abbed, -en	affekt/asjon, -ert	aksidens	allehelgensdag
abbedi/sse, -en l. -a	affinitet	aksiom, -et	allehånde
abborr/en = åbor	affisere, -te	aksise, -en	allemanns/eie,
abebok, -a	affære, en	aksj/e, -en, -onær	-venn
abdi/kasjon, -sere	afgan/er, -sk	aksjon, -en	aller best osv.
aber, en, et	aften/er, -sang	akt/e; -else, -en	allerede
abessin/ier, -sk	aftens/bord, -mat	akter, -skott,	alle sammen
abnorm	agat, -en	-speil, -ut	alle slags, allsl.
abonne/nt, -ment	age, holde i a.	akten/for, -om	allesteds, -nær-
abonnere, -te	agere, -te	-aktig	værende
abrupt	agg/et (nag)	aktrise, -en l. -a, -r	alle tider, vegne
absint	aggregat	aktiv/itet	all/fader, -farvei
	aggressiv	aktiv/um, fl. -a	l. -farveg
	agio	aktor, -en, -er	alli/anse, -ert

Figure 1 Eitrem 1939. The beginning of the section for the letter *a*.

Table 1 shows a simple instance of extracting lexical items from a school speller. Three lines of text contain five base forms of six different lexical items, one of them a noun with two genders.

Table 1. Example from Figure 1 of headword extraction and added POS (Eitrem 1939).

Text	Inflection	Extracted lexical item	Added POS	Sense
abdi/kasjon, -sere		abdikasjon	noun, masculine	abdication
abdi/kasjon, -sere		abdisere	verb	abdicate
aber	en, et	aber	noun, masculine or neuter	disadvantage
abessin/ier, -sk		abessinier	noun, masculine	Abyssinian
abessinsk		abessinsk	adjective	Abyssinian
abessinsk		abessinsk	noun, gender masculine	Abyssinian

In the sample above, all base forms lack explicit POS, and one, *abessinsk*, can double for two lexical items. The addition of a base form in the current orthography and POS for all represents a reasonable interpretation, but it is still an interpretation. Therefore, it is important to present the text itself in context in the electronic version, with the preceding and following entries shown. Users need to be able to compare what was printed with what is claimed to be a true interpretation.

From the point of view of digitization three concerns emerge: 1) to preserve the text of the original document as carefully as possible, so that it can be presented as it was; 2) to extract electronically the full register of lexical items with such supporting information as there is; 3) to supply essential and missing information from reliable contemporary sources.

These aims necessitate careful encoding and tagging, based on the conventions of the original document.

### 3 The case of *Norsk rettskrivningsordbok* versus the School Spellers

The purpose of discussing the inclusion of *Norsk rettskrivningsordbok* in the Dictionary Hotel is to highlight the process of computer assisted tagging of a dictionary.

*Norsk rettskrivningsordbok* (Sverdrup 1940) is the largest orthographic dictionary attempted for Norwegian Bokmål. It still exists under the title *Tanums store rettskrivningsordbok* and has been through numerous expansions and revisions. The original compiler was Jakob Sverdrup (1881–1938), professor of Germanic philology at Det Kongelige Fredriks Universitet (now the University of Oslo). In the introduction to *Norsk Rettskrivningsordbok*, it is stated that Sverdrup and his co-editor, Sandvei, drew the materials from the lexicographic excerpt collections in existence at the University, but also from catalogues and registers of all sorts, such as goods registers from the major retailers.

The orthography reform of 1938 was a major one, aimed at bringing Bokmål and Nynorsk closer together. It caused lasting excitement and resentment, especially among Bokmål users, since it went far in giving word forms from the Norwegian vernacular equal status with the traditional Danish-based forms. The vernacular forms were often identical with existing Nynorsk ones, but the influence of orthophonic ideals was stronger for Bokmål than for Nynorsk. The 1938 reform therefore includes forms like *selle* verb ‘to sell’ (equal to what is found in Swedish standard language) in addition to the traditional Bokmål form *selge*.

With more than 175,000 entries, Sverdrup’s orthographic dictionary is the most comprehensive documentation of the 1938 orthographic reform for Bokmål, and therefore a valuable measuring point for assessing the influence of orthographic reform on usage. Using it as a basis for a lemmatizer will provide a before-and-after separation mark for Bokmål text. It is also probable that a high proportion of the word forms never will be found in any corpus.

Table 2. Text sample from Sverdrup (1940). Paragraphs numbered for reference.

1	<b>adalhending</b> , -en; -er, -ene (helrim).
2	<b>adalin</b> (sovemiddel).
3	<b>adam</b> (fra Bibelens Adam); den gamle Adam; i Adams drakt.
4	<b>adamitter</b> pl. (sekt); adamittisk adj., n. -.
5	<b>adams_barn</b> , _drakt, _eple, _fiken, _hjerte, _natur, _slekt, _sønn, _tre, _ætt.
6	<b>adapsjon</b> , -en; -er, -ene (tilpasning); adapsjons_evne, _form o. fl., adaptere, -te, -t (tilpasse).

Table 2 shows a sample of the transcribed text from Sverdrup (1940), set up in table format here for reference. This text has all the categories expected in a defining dictionary, but the presentation



is far more compact and the ordering unpredictable, as only the categories deemed essential for headword identification are included in each entry. First headwords are set in bold. Meaning (1, 2, 4, and 6) and etymological information (3) both appear in parentheses. POS information appears as abbreviated inflection forms (1, 4, and 6) or an abbreviation (4). MWEs are rendered in plain text (3), after semicolons. Derived headwords in base form are rendered in plain text (4, 6). Compounds are nested (5, 6).

The four school spellers included in the Dictionary Hotel and linked to the Meta Dictionary have from 13,000 to 24,000 entries. These are basic school spellers, and the vocabulary covered can therefore be expected to overlap, and also give guidance as to what school authorities saw as the central vocabulary from 1938 to 1986. With the exception of the initial headword in each paragraph, the organization of information is as unpredictable, and the range of categories as wide, as in Sverdrup (1940).

## 4 The Dictionary Hotel

The Dictionary Hotel is the central repository of the Language Collections for searchable electronic dictionaries. It was created in 2005 as a part of the databases with background material for the Norsk Ordbok (Norwegian Dictionary) project, and linked to the material index Metaordboka (Meta Dictionary). The purpose was to create a common framework for retro-digitized dictionaries. It is an electronic library for mostly retro-digitized dictionaries, glossaries and other collections of lexical information, and is equipped with a portal for searching them in parallel and showing aligned search results (Tvedt et al. 2007). The present contents are 60 dialect dictionaries and word collections, plus a couple of large dictionaries. The school spellers and Sverdrup (1940) are stored in the Dictionary Hotel. The collection is expanded when possible, depending on capacity. The Dictionary Hotel is analogous to the German Wörterbuchnetz (Wörterbuchnetz 2018); see also Moulin and Nyhan (2014) for a discussion of this.

Each item in this library is stored as an xml-document with inline mark-up. The major purpose is to facilitate the study of lexical information. Therefore, the texts are encoded according to TEI dictionary format, which provides for dividing each text into a series of smaller text chunks or entries. For each entry, one or more headwords are marked. These are used as linking points to the common index the Meta Dictionary (see section 5 below).

Each original dictionary or collection is considered a unique document in its own right. Standardization levels vary, and all have their editorial idiosyncrasies. Therefore, every entry has an added layer containing one or more standard base forms with POS which is used in linking the individual dictionary to the corresponding Meta Dictionary entry. The standard language used for the dialect materials is Nynorsk. This is in accordance with language documentation practice in Norwegian philology since dialect studies started in the 19<sup>th</sup> century. It should be mentioned that different layers of interpretation are somewhat tricky to express in inline encoding. For an alternative approach, see Bouda and Cysouw (2012).

Every document in the Dictionary Hotel has its standard language set to either Bokmål or Nynorsk. In linking a document to the Meta Dictionary, this means that entries in a dictionary marked as “Bokmål” will be linked to a Meta Dictionary entry with an identical base form plus POS marked as “Bokmål”, and the corresponding procedure for documents set as Nynorsk documents.

### 4.1 Encoding decoded information

From an everyday point of view, encoding dictionary texts may seem a straightforward task. But the requirement is scholarly reproducible results based on the application of transparent methods.



Therefore, each dictionary must be treated as a unique document, analyzed and given a mark-up documenting the analyzer's understanding of the author's intentions and a decoding of the information found in text. The text itself must not be corrupted in the process.

Dictionaries can be very complex texts with plenty of ambiguities and lacunae. As shown above, this is true even of school spellers, as this genre shows a wide variation. In the introduction to the encoding of dictionaries in the TEI P5 guidelines (TEI 2018) two important issues are highlighted:

First, because the structure of dictionary entries varies widely both among and within dictionaries, the simplest way for an encoding scheme to accommodate the entire range of structures actually encountered is to allow virtually any element to appear virtually anywhere in a dictionary entry. It is clear, however, that strong and consistent structural principles do govern the vast majority of conventional dictionaries. [...]

Second, since so much of the information in printed dictionaries is implicit or highly compressed, their encoding requires clear thought about whether it is to capture the precise typographic form of the source text or the underlying structure of the information it presents. Since both of these views of the dictionary may be of interest, it proves necessary to develop methods of recording both, and of recording the interrelationship between them as well. [...] (TEI P5 Guidelines, Chapter 9 Dictionaries)

Both aspects are important in the Dictionary Hotel. Ideally, our goal is to preserve the typographic form at least to a degree that documents the analyzer's interpretation of the contents, that is, the information about the lexical items.

Some simplification of the original typography or character inventory may be necessary, for example removing word stress markers or changing the typeface from Black Letters (Fraktur) to Antiqua, if the typeface is irrelevant to the entry structure. The encoded text can then be used to reproduce a text with a layout close to the original. In this way we follow the basic principles for modern electronic text philology, a practice recommended for all retro digitization of dictionaries.

## 4.2 Preparations for encoding the text

The encoding and proofreading of the text is important in preparing for text analysis and tagging. The following is a run-through of the procedure established for the Dictionary Hotel:

(1) Consider the encoding a preparation for structural analysis to get the tagging as correct as possible, while preserving the original text. (2) Check the character set. A dialect dictionary may contain characters not found in Unicode to represent particular sounds (the Norwegian sound transcription alphabet Norvegia has no ISO standard). If a character of this kind is found then it must be given an established substitute for the encoding, preferably an entity. In other cases, a particular character may be used both to indicate a sound quality and a change of category in the text. This is easily interpreted by a human reader, but not by a computer program, and should be handled in connection with the encoding. (3) Decide on a system for handling the relationship between typography and field content. Typography can be over-specific or ambiguous or both. Italics for both cross references and deviant dialect forms are an example of a common occurrence. (4) Encode paragraphs as continuous strings, do not try to reproduce the line breaks of the printed page. (5) Proofread meticulously, as a missing character or space may throw the automatic tagging off course and cause failure to pick up embedded lexical items or cause strings of compounds to come out with the wrong first part.

## 4.3 Text analysis and mark-up

The most interesting step comes when the encoding is done. Analyzing the text contents: what constitutes a headword, POS, inflection, description of meaning or usage examples? In dictionaries made

by scholars the formats tend to be well defined and based on the scheme used in (old, printed) Latin dictionaries. As in these dictionaries, a head word's POS markers are often given indirectly by listing (often abbreviated) inflected forms. In spellers this practice tends to be the rule.

The dictionaries and glossaries which are candidates for the Dictionary Hotel comprise thousands of entries. A manual mark-up is therefore impossible. The mark-up is done by a script developed in an iterative process: finding general patterns, analyzing the results by the use of standard KWIC-tools, refining the script and adding exceptions until the result is satisfactory; see Christmann (2001) for a description of the similar process behind the digital *Grimm's Deutsches Wörterbuch*. Printed dictionaries are idiosyncratic, and the scholarly and philological requirement of respect for the original text makes it doubtful that there is much to gain from trying to develop an analyzer based on deep learning (that is, the use of a neural network). Depending on the required level of analysis, the process may consume several person-weeks. The resulting script documents the analysis and should be kept for future use and consultation. Manual encoding by search and replace in a text editor is not recommended. The process corresponds to the construction of a hand-tagged training corpus and the development of a statistically-based corpus analyzer. The main difference is that one has to carry out the process for each dictionary.

The output from the analyzing script is an xml-encoded text where the result of a scholarly interpretation is represented by the mark-up. In a well-structured dictionary the entry format normally has a standard structure: one or more base forms acting as headwords, POS, spelling variants and the hierarchy of definition and citations. All the information is already present in the original text, though the presentation may require a fair amount of interpretation, cf. Figure 1 and Tables 1 and 2 above. There are, however, some challenges frequently encountered in dictionaries, and especially school spellers for Germanic languages, such as Norwegian:

*Entry organization is not entirely alphabetical:* the system of creating compounds causes editors to organize dictionaries into nests, with an introductory entry for the initial part, and then a series of entries for derivations and compounds.

*Finding the POS:* Since the last part of a compound is usually an independent word with a separate entry, information about inflection and POS is often omitted in entries for compounds. For spellers these minimal compound entries are the norm. In such cases, the complete compound is the first headword of the nest plus the relevant second part, and these can be stored together in an attribute. The POS is more problematic, since one has to find a single headword corresponding to the second part, which in many cases is not possible due to the high frequency of compounds constructed from more than two parts, or homograph candidates for the second part of the compound.

As a consequence, there will at the end always be a significant number of identified headwords without POS information. In order to link a tagged dictionary to the Meta Dictionary, POS information must be added to every headword semi-automatically or manually. This information is not a part of the original dictionary. To keep the original and added information separate, an extra level is added. The result will be a "dictionary" where all headwords will have POS information either from the original analyzed dictionary or from other sources. The original analysis is put into a dictScrap element and can be extracted and displayed by the application of, say, xslt-transformations.

```
(1) <entry xml:id="SverdrNR_orig000031" n="nest_no_1"><form><orth>abelmoskus</orth><-
    gramGrp><pos> m</pos></gramGrp></form><dictScrap><form type="headword" orig="a-
    belmoskus">abelmoskus</form> bot.</dictScrap></entry>
```

## 5 Linking Base Forms and Materials

The access point of scholarly language analysis is always the word form in context, whether the raw materials are registered as sound or text. Registered raw materials must be easily accessible by solid criteria and have adequate metadata, including source information. The only functional linking of base forms on the one hand with materials on the other hand is to organize both around the lexical item as a fixed point. The lexical item has to have an established identity, and there must be sufficient materials to prove POS, inflection and at least one established sense. This information is found in the major scholarly dictionaries, and also in the traditional well-ordered collections underlying them.

In both Bokmål and Nynorsk, a lexical item can have more than one base form within a given orthography, and is almost certain to have more than one if the time span covered by the materials is long enough – something which is true of any language. The written and spoken expressions of the base forms attached to one particular lexical item can also vary considerably synchronically (through different dialect forms) and in the written standard (through orthographic variants).

In order to compare two closely related orthographies, an index is needed which allows alignment of the base forms of each orthography for the same lexical item.

### 5.1 The Meta Dictionary – headword forms and POS

For Norwegian, there is one language tool particularly suitable for indexing Bokmål and Nynorsk in parallel – the Meta Dictionary. This is an electronic registry for the Norwegian lexicon which allows linking lexicographical evidence to several base forms through one node, the Meta Dictionary entry, representing the lexical item. This node or entry may contain additional information about the lexical item, and in this way the Meta Dictionary is different from the Dictionary Hotel or the German Wörterbuchnetz (see also Ore 2000 and Ore & Ore 2010 for a shorter description in English). The Meta Dictionary format has also been adapted for the new editing and publication system for the Danish Dictionary of Old Norse (*Ordbog over det Norrøne Prosasprog*, ONP), see the related discussion in Johannsson and Battista (2014).

Since the Meta Dictionary indexes materials according to the headword form of the lexical item, the POS register is limited to forms found in headwords, i.e. the base forms of nouns, adjectives, adverbs and verbs. Nouns can be assigned gender and can have the singular or plural form. Adjectives and adverbs can be marked with degree (positive, comparative, superlative). Verbs are listed in the infinitive form. All other POS are single forms in Norwegian. This means that all pronouns have separate entries in the Meta Dictionary, irrespective of their syntactic function.

A modern language index organized around the lexical item must also handle categories with limited recognition in the lexicological literature. The written standard vocabulary contains abbreviations and symbols, and these are POS-marked as such. Prefixes are divided into two groups in Norwegian lexicography, the preformatives marked “pref” and those that represent the first part of compounds (including joining infix) marked “føreledd” ‘first section’. The Meta Dictionary also has entries for propria marked as such. The most commonly used MWEs found as headwords in orthographic dictionaries also have Meta Dictionary entries with the POS depending on syntactic function.

### 5.2 The Meta Dictionary entry

The entry head allows indexing attached materials with (1) standard language (Bokmål or Nynorsk); (2) several base forms per language form; (3) standardization status for each base form; (4) start and end date for each status; (5) POS; (6) segmentation. This entry format allows for building a diachronic

base form register by linking lexical resources and indexing them per entry for one or both of Bokmål and Nynorsk. Base forms which are or have been standard forms are used as entry headwords, while the sources express the range of forms found in spoken and written language over time. The base form schema in the Meta Dictionary is close to the format used in the Word Bank entries, but the Meta Dictionary has no information on inflection, or rules for generating full form schemas.

The Meta Dictionary entry body consists of references – hyperlinks – to various source databases. Most of these show usage examples or give definitions, but there are also materials only showing occurrence (in a given place or at a given time), pronunciation, POS, word formation potential (derivations, compounds) and so on. A typical Meta Dictionary entry is shown in Figure 2 below. The left part shows the list of links to the materials, with different icons for different source types. The middle part shows the entry tree, with segmented forms. To the right the schema for each base form is shown, with orthographic form, language choice, and status in the orthography, with start and end dates.

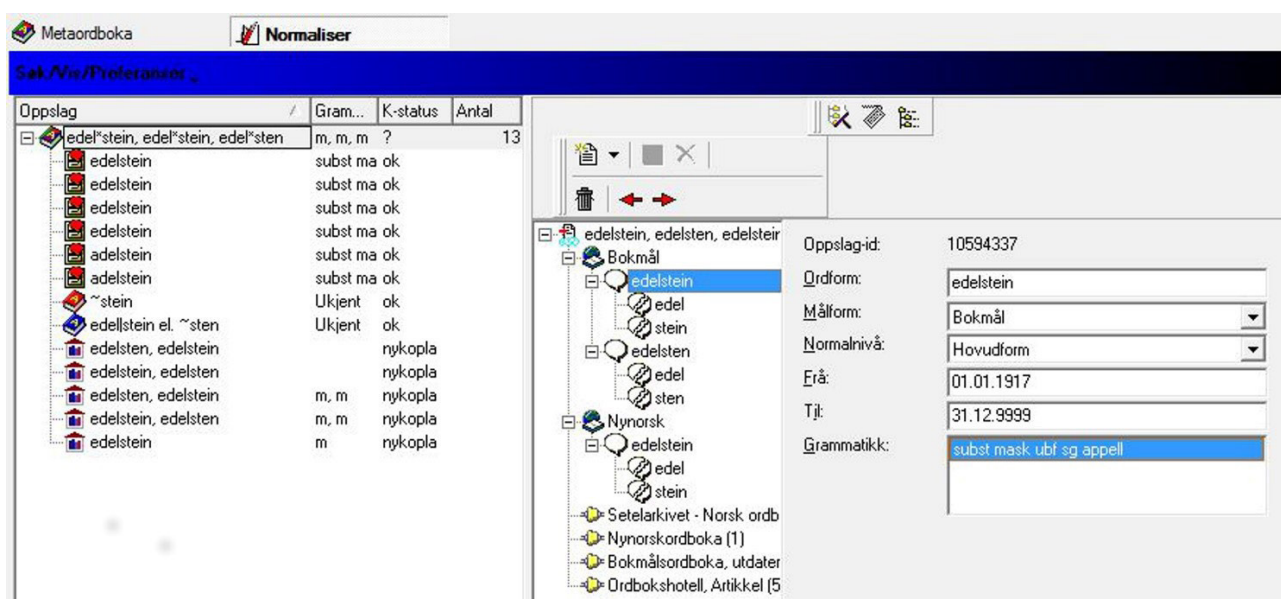


Figure 2. The Meta Dictionary entry, as shown in the editor, for the noun *edelsten* el *edelstein* m (Bokmål), *edelstein* (Nynorsk) ‘jewel’.

The orthographic differences between Bokmål and Nynorsk may consist of one character only, but that one character is important. The alignment criterion is therefore full identity in base form and POS. Since the Meta Dictionary has a time scale, it is possible to show the transition from one orthographic form to another, cf. Figure 3. The Meta Dictionary is available and freely searchable on the web.

**sjåfør** m (Bokmål, 1917-)  
**chauffeur** m (Bokmål, 1901-1917)  
**sjåfør** m (Nynorsk, 1917-)  
**chauffeur** m (Nynorsk, 1873-1917)

Figure 3. The Meta Dictionary entry for *sjåfør*, noun, masculine, with the previously valid form *chauffeur*, plus information about written standard and year.



### 5.3 Meta Dictionary Contents

The Meta Dictionary coordinates materials in standard language from different periods and transcriptions of spoken material. Nynorsk at present has a better coverage than Bokmål. Primary sources of usage for Nynorsk comprise excerpts (text, image) with metadata, or lines from corpus concordances. Secondary sources comprise electronic versions of scholarly dictionaries for Nynorsk, including major works covering the orthographies of 1873, 1917, 1938 and the current one of 2012. These dictionaries have from 40,000 to 300,000 entries, and the orthographies for Nynorsk lexical items are therefore reasonably well covered. Primary sources for Bokmål at present comprise a neologism archive (*Nyordsarkivet*, excerpts from a range of sources, ca 1970 – 1995). The secondary sources are dictionaries and spellers, ranging in size from ca 13,000 to 180,000 entries and covering the orthographies of 1938, 1959 and 2005. Many of the dictionaries and spellers are collected in the Dictionary Hotel (see section 4 above).

The Meta Dictionary was initially created as an index for the Nynorsk language collections, linking standard Nynorsk and dialect information from 1600 until today (Ore 2000). The purpose was to assist the editing of *Norsk Ordbok* (completed 2016). Through the Meta Dictionary, the Norwegian Language Collections for Nynorsk and Norwegian dialects were indexed with one base form in the Nynorsk 1938 orthography with traditional forms, plus POS. Homograph separation beyond base form plus POS has always been possible, but was not attempted in the Meta Dictionary while *Norsk Ordbok* was in production.

The Meta Dictionary is now used as a tool of coordination between Bokmål and Nynorsk. Coordination at the level of the lexical item is a long step forward towards solid ground for analysis and comparison of Bokmål and Nynorsk at different times during the 20<sup>th</sup> century and up to the present.

Linkage of a resource to the Meta Dictionary is automated; if the new entry in the resource finds a Meta Dictionary entry with the same base form and POS, it gets linked, if not it gets a new Meta Dictionary entry. The status system in the Meta Dictionary tells the human moderator – a trained lexicographer – where changes have been made, and changes are quality checked before being approved with or without adjustment.

## 6 Preliminary results

The Bokmål materials so far linked to the Meta Dictionary are limited, but some findings are nevertheless worthy of note.

### 6.1 The correlation between Bokmål and Nynorsk in the Meta Dictionary

The Bokmål additions to the Meta Dictionary have been made in the last 18 months, and there has not been time for much manual alignment, as Sverdrup (1940) is a fairly recent addition. But there are some interesting figures. Table 3 below shows contents in the Meta Dictionary after the uploading of Sverdrup 1940.

Table 3. Results after adding Sverdrup 1940 to the Dictionary Hotel, with linkage to the Meta Dictionary.

		Bokmål	Nynorsk
1	Entries in the Meta Dictionary per language	361 006	545 766
2	Unique headwords (base forms)	359 159	545 862
3	Unique combinations of headword plus POS	371 351	557 627
4	Entries with headwords in Bokmål or Nynorsk only	258 509	443 269
5	Entries with headwords in both Bokmål and Nynorsk	102 497	102 497
6	Entries with more than one headword per language	10 273	11 720



The Meta Dictionary is dynamic. Since the alignment criteria for automatic attachment to existing entries are strict (see section 5.3 the end), there will be changes as the manual alignment process gets under way. The numbers of lines 1 and 4 will decrease a little as minor divergencies get sorted; the numbers of lines 5 and 6 will increase.

The number of unique headwords (line 2) is higher than the number of entries (line 1) for both Bokmål and Nynorsk, because some entries have more than one headword (cf. line 6). The most striking piece of information is found in row 5 – between 1938 and today, which shows that more than 75,000 Meta Dictionary entries have one or more head word forms that are identical for Bokmål and Nynorsk, which is counter-intuitive to many Norwegians. This alignment does not cover inflection morphology. To some extent this sameness between headword forms has been reduced after 1980, but there are still a high number of examples, almost twice the overlap between *Bokmålsordboka* and *Nynorskordboka*.

Table 4. Overlap between the Bokmål school spellers in the Dictionary Hotel, and overlap with *Bokmålsordboka* and Sverdrup 1940.

	Sources	Number of entries	Number of (shared) headwords	Headwords found in Lexicographic Bokmål Corpus (1985 – 2005)	Hits in% of number of (shared) headwords
1	School speller 1939 I	13,046	13,119	10,380	79%
2	School speller 1939 II	17,336	18,179	13,351	73%
3	School speller 1959	17,361	17,788	14,218	79%
4	School speller 1973	23,950	25,168	20,099	80%
5	Bokmålsordboka 1986 – 2005	71,142	75,590	61,228	81%
6	Sverdrup 1940	175,948	179,000	75,500	43%
7	Sverdup + BOB		48,600	40,900	84%
8	All school spellers		7,676	7,159	93%
9	All school spellers + <i>Bokmålsordboka</i>		7,423	7,036	95%
10	All school spellers + Sverdrup (1940)		7,365	6,906	94%
11	All school spellers + BOB + Sverdrup		6,903	6,806	95%

In Table 4 explores the contents of the school spellers in the Dictionary Hotel and Sverdrup (1940) and *Bokmålsordboka*, in terms of numbers of entries and headwords. The table also shows the co-occurrence of headwords in the school spellers, and in the school spellers and Sverdrup (1940) and *Bokmålsordboka* (2005). These publications give information about the orthography of Bokmål over 80 years. A headword list of 6,906 items is common to all, and it seems safe to guess that this is a core list of essential vocabulary.

Table 3 also shows the results of testing the headword lists of the Bokmål school spellers (of 1939, 1959 and 1973) against the Lexicographic Bokmål Corpus (LBK), a 100-million-token corpus covering the period 1980 – 2005, and containing text from a variety of genres. The school spellers admittedly have a limited vocabulary, but even so, the rate of hits from the individual school spellers at 70 – 80% suggests that the orthographic changes from 1938 until today cannot have been dramatic. Line 11 shows the hit rate for headwords found in all school spellers, in Sverdrup (1940) and *Bokmålsordboka*, with a hit rate in the LBK of 95%, which seems to confirm the core list status.

## 6.2 Possibilities and future goals

What remains to be seen is how and to what extent the different groups of headwords have been used. In order to find out, full form registers for the different orthographies must be developed.

First the Norwegian case: the Language Collections at the University of Bergen has expandable full form registers for Bokmål and Nynorsk – the Word Banks. The next step should therefore be: (1) linking headwords of the Meta Dictionary to the proper Word Bank, thus equipping them with inflection paradigms for the current orthographies; (2) adding missing dated inflection paradigms to the Word Banks back to 1938 – probably few would be needed – and adjusting the Word Bank entries accordingly. With these steps taken, it will be possible to make exact and detailed examinations of all text going back to 1938, and also to examine the use of 1938 forms in text from before the orthographic reform. One could then follow on with documenting the orthography before and after 1917 and back to 1901 (Bokmål) and 1873 (Aasen's *Norsk Ordbog*).

Second, the general case: All languages have a history of language standardization. What happened to the orthography of English from 1700 to 1800? Is it possible to measure the effects of Doctor Johnson's dictionary of 1755 by comparing corpora from before 1750 with corpora from after 1755? At present, there is a widespread assumption that language standardization is something that happens, more or less spontaneously, if the language community is large and literate enough. But these assumptions have not been critically examined by direct examination of text. A Doctor Johnson full form register, with a tagger, would thus be very welcome, together with a Meta Dictionary for English.

## 7 Conclusion

Our aim has been studying how one can use lexicographic resources to measure language variation and change, and what sort of tool can give reliable results in this context. These questions arise when considering the state and history of the Norwegian written standards, Bokmål and Nynorsk. By "tool" one should not only focus on concrete software applications, and in this paper we have discussed the methods and requirements for setting up a research environment, describing a full-scale test. The major benefit for researchers and the public alike is that the contents of the dictionaries and spellers discussed in this paper are accessible for whatever searches one cares to make. Endless questions can now be met with precise answers, which in turn can be critically examined.

A less scholarly motive for getting this done is the fact that full form registers for Bokmål and Nynorsk for around 500,000 lexical items, expressed in taggers for corpora, would contribute to making Norwegian easier to use in the various language technology applications needed in a society increasingly dependent on electronic support.

## References

- Aasen, Ivar (1873): *Norsk Ordbog med dansk Forklaring*. Fjerde uforandrede udgave, (1918). Kristiania: Vestmannalaget og Cammermeyers forlag.
- Atkins, S. & Michael R. (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bouda P., Cysouw M. (2012) Treating Dictionaries as a Linked-Data Corpus. In: Chiarcos C., Nordhoff S., Hellmann S. (eds): *Linked Data in Linguistics*. Springer, Berlin, Heidelberg, 978-3-642-28248-5, Accessed at: [https://doi.org/10.1007/978-3-642-28249-2\\_2](https://doi.org/10.1007/978-3-642-28249-2_2) [30/03/2018].
- Christmann, R.(2001); Books into Bytes: Jacob and Wilhelm Grimm's Deutsches Wörterbuch on CD-ROM and on the Internet. In: *Literary and Linguistic Computing*, Volume 16, Issue 2, 1 June 2001, Pages 121–133, Accessed at: <https://doi.org/10.1093/llc/16.2.121> [18/052018].

- Den nye rettskrivning. Regler og ordlister* (The new Orthography. Rules and Word Lists). Utarbeidet ved Den departementale rettskrivningskomite. Kristiania. Det Mallingske Bogtrykkeri 1918. Fastsatt ved kgl.res. 21. desember 1917. Accessed at: [http://www.sprakradet.no/Spraka-vare/Norsk/Faksimilebiblioteket /Den\\_nye\\_rettsskrivning\\_1917](http://www.sprakradet.no/Spraka-vare/Norsk/Faksimilebiblioteket/Den_nye_rettsskrivning_1917) [25/03/2018].
- Eitrem, H. (1939): *Rettskrivning 1938. Regler for skoler og privat bruk*. Oslo: Fabritius.
- Grønvik, Oddrun & Ore, Christian-Emil Smith (2013): What should the electronic dictionary do for you – and how?, In Iztok Kosem; Jelena Kallas; Polona Gantar; Simon Krek; Margit Langements & Maria Tuulik (ed.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, Eesti Keele Instituut. ISBN 978-961-93594-0-2. pp. 243 – 260. Accessed at: [http://eki.ee/elex2013/proceedings/eLex2013\\_17\\_Gronvik+Ore.pdf](http://eki.ee/elex2013/proceedings/eLex2013_17_Gronvik+Ore.pdf) [17/05/2018].
- Hovdenak, M., Killingbergstrø, L., Lauvhjell, A., Nordlie, S., Rommetveit, M. & Worren, D. (2006): *Nynorskordboka* (4. utg.). Oslo: Samlaget.
- Johannsson, E.T. & Battista, S.(2014) A Dictionary of Old Norse Prose and its Users — Paper vs. Web-based Edition. In Abel, A., Vettori, C., Ralli, N. (eds) *Proceeding of the XVI EURALEX, The User in Focus; 15-19 July 2014, Bolzano*, Bolzano, Eurac Research, 2014, ISBN: 978-88-88906-97-3. Accessed at: <http://euralex.org/category/publications/euralex-2014/> [30/03/2018].
- Krogstad, T. & Seip, D.A (1959): *Rettskrivningsregler. Rettskrivningslære og ordliste*. Etter «Ny læreboknormal 1959». Oslo: Cappelen.
- Lange, A. (1939): *Norsk rettskrivningsordliste*. Oslo: Tiden Norsk Forlag.
- LBK (2011) Leksikografisk bokmålskorpus. Universitetet i Oslo. Accessed at: <http://www.hf.uio.no/iln/tjenester/kunnskap/samlinger/bokmal/tilgang-korpus> [25/03/2018].
- Metaordboka. Språksamlingane, Universitetet i Bergen. Accessed at: <http://usd.uib.no/perl/search/search.cgi?appid=7&tabid=57> [25/03/2018].
- Moulin, C. & Nyhan, J. (2014): The Dynamics of Digital Publications, An Exploration of Digital Lexicography. In: Davidhazi, P. (ed) *New publication cultures in the humanities: exploring the paradigm shift*, Amsterdam University Press, 2014, ISBN: 90-485-1971-3.
- Norsk ordbok*: Ordbok over det norske folkemålet og det nynorske skriftmålet (Dictionary of the Norwegian vernacular and the Nynorsk written standard). 1966-2016. Vol. 1-12. Oslo: Det Norske Samlaget.
- Nyordsarkivet. Bokmål. (The Neologism Archive. Bokmål) Språksamlingane, Universitetet i Bergen. Accessed at: <http://usd.uib.no/perl/search/search.cgi?appid=245&tabid=3521> [20/03/2018].
- Ordbokshotell. Språksamlingane, Universitetet i Bergen. Accessed at: <http://usd.uib.no/perl/search/search.cgi?appid=118&tabid=1777> [25/03/2018].
- Ore, C.-E., (2000): Metaordboken – et rammeverk for Norsk Ordbok. In: *Nordiska studier i lexikografi* 5. Göteborg: Nordisk forening for leksikografi, s. 250—270.
- Ore, C.-E. & Ore E. (2010) Re-linking a Dictionary Universe or the Metadictionary Ten Years Later In: *Conference Abstracts, King's College London, London, July 7 – 10, 2010* Published by Office for Humanities Communication Centre for Computing in the Humanities, King's College Digital Humanities 2010, London.
- Stortingsmelding 1 (2017–2018) *Nasjonalbudsjettet* 2018. Kap. 326 post 75 Tilskudd til ordboksarbeid. (Parliamentary Report 1 2017-2018. National Budget 2018. Ch. 26 subsection 75 Allocation to Lexicography) Accessed at: <https://www.regjeringen.no/no/dokumenter/prop.-1-s-kud-20172018/id2574640/sec2#match2> [12/10/2017].
- Sverdrup, J. (1940): *Norsk rettskrivningsordbok*. Bokmål. Oslo: Tanum.
- TEI (2018) *Text Encoding Initiative, Guidelines to Text Encoding (P5)*. Accessed at: <http://www.tei-c.org> [30/03/2018].
- Torvik, I. (1973): *Ordlister for alle. Bokmål*. Oslo: Universitets-forlaget.
- Tvedt, L.J, Lien, E. & Eide, Ø. (2007): Ordbokshotellet – varig lagring og formidling av norske ordsamlinger. In: Arboe, T. (red.), *Nordisk dialektologi og sociolingvistik, Foredrag på 8. Nordiske Dialektologkonferanse, Århus 2006*. Peter Skautrup Centret for Jysk Dialektforskning, Aarhus Universitet, s. 379–388.
- Vikør, L. (2001): *The Nordic Languages. Their Status and Interrelations*. Oslo. Norsk språkråd. 3. ed.
- Wangensteen, B. (2006): *Bokmålsordboka*. Definisjons- og rettskrivningsordbok. (3. utg.). Oslo: Kunnskapsforlaget.
- Wörterbuchnetz (2018) Accessed at: <http://www.woerterbuchnetz.de> [30/03/2018].

# A Universal Classification of Lexical Categories and Grammatical Distinctions for Lexicographic and Processing Purposes

*Roser Saurí, Ashleigh Alderslade, Richard Shapiro*

*Oxford University Press*

*E-mail: roser.sauri@oup.com, ashleigh.alderslade@oup.com, richard.shapiro@oup.com*

## Abstract

We introduce COMO (Compositional Morphosyntactic Ontology), a classification of part-of-speech categories and their associated grammatical features, which aims to be valid across languages of very different typology. The work has been carried out within the context of the Oxford Global Languages programme, which has the goal of developing language knowledge for 100 languages, particularly those under-represented in the digital space. The requirements around this project are: to be able to describe languages of different types while respecting their grammatical tradition, and to be able to serve two main use cases that define our typical work, namely, the labelling of linguistic information in lexicographic products, and the provision of support for language processing tools and corpus annotation processes. These requirements determined the conception and design of COMO, created as a reference model within a broader data architecture in order to address issues of syntactic and semantic interoperability. Our proposal builds on top of previous initiatives in the field aiming at the same goals, but incorporates different features in order to accommodate for the requirements in the project.

**Keywords:** part-of-speech tagging, morphosyntactic information, language modelling, interoperability, multilinguality

## 1 Introduction

The codification of each language (dictionary and grammar) has traditionally been carried out according to distinctions identified within the language, or at best in line with distinctions already used in similar ones. For example, the concepts and terms employed to account for the tense system in different Romance languages have an indisputable resemblance, but differ in some respects to the treatment of the same system in English. At the part-of-speech (POS) level, differences are also evident. Some linguistic descriptions present lexical categories specific to the language they describe (e.g., phrasal verb for English), or use different terms to refer to the same kind of units (e.g., demonstratives functioning as noun modifiers can be treated as either determiners or adjectives, depending on the linguistic tradition of the language).

This is acceptable when working within the scope of a dictionary or grammar of a particular language, and even in the case of bilingual lexicography. However, the approach soon becomes insufficient for modern multilingual usages, because there is no common ground that allows the different language descriptions to meet, thus providing little opportunity for interoperability among resources or applications of different provenance.

The problem of heterogeneous linguistic encodings is not new. It was already on the agendas of EU-funded projects in the early 1990s, when the Expert Advisory Group for Language Engineering Standards (EAGLES) was created in order to provide guidelines and standards for encoding and managing linguistic resources in different languages. At that moment, the effort was mostly



focused at ensuring *syntactic interoperability* among systems (dictionaries, corpora, etc.) of diverse origin and languages; that is, at ensuring that they could share information or communicate given a common data model and format. Syntactic interoperability, however, does not guarantee that the interpretation of the information shared by the systems is the same (Ide & Pustejovsky 2010). An example of this in the area of morphosyntactic information is brought by the term *absolute*, which in some English dictionaries is applied to pronouns like *yours* and *theirs*, to express that they only have a pronominal use (as opposed to e.g. *her*), whereas in Romanian it refers to transitive verbs when used without an object.

The limitations of systems not fully communicating at the semantic level have become apparent in more recent times, when the digitization of linguistic knowledge and the development of natural language processing (NLP) tools have started to expand to languages of very different typological affiliation (in the last decade, Chinese, Japanese, and Arabic; more recently, languages from the Indian subcontinent). Adding to this, there have also been advances in the technology supporting such digitization, which now allows for a fully-linked data paradigm. In the linguistics field this has materialized into the Linguistic Linked Data program (LLD, Chiarcos, McCrae et al. 2013), which makes it possible to gather linguistic datasets of very different languages in the same repository, connect these at a lexical level, and elicit additional knowledge by automatic means.

To allow for full communication between different linguistic resources, a number of more recent projects have been deployed specifically targeting *semantic interoperability*. All of them rely on the existence of a reference model where domain concepts are unambiguously defined, and to which the different linguistic resources defer in order to communicate amongst themselves and interpret information meaningfully. Some of these projects are: the General Ontology of Linguistic Descriptions (GOLD), the ISO TC37/SC4 Data Category Registry (ISocat), the set of Ontologies of Linguistic Annotation (OLIA), the Universal Dependencies project (UD), and the Universal Morphological Annotation (UniMorph).<sup>1</sup>

The challenges addressed by these projects are also shared at the Dictionaries Division of Oxford University Press (OUP), specifically within the context of the Oxford Global Languages (OGL) program.<sup>2</sup> OGL aims to develop linguistic knowledge for 100 languages with a particular focus on those which are under-represented in the digital sphere. This includes languages with very little codification or with features quite different from Indo-European ones, which are typically the languages that have inspired the grammatical distinctions commonly assumed in linguistic analysis. Furthermore, at the technical level the project plans to store and integrate the information for the resources of these many languages in a centralizing repository, along the lines of the Semantic Web paradigm. A key factor in this is having a model able to accommodate complexities across these many languages and enable the linking between them (Parvizi et al. 2016).

The current paper focuses on the fragment of the model concerning morphosyntactic information, that is, POS classes (aka lexical categories) and grammatical distinctions manifested by morphological and syntactic means. In particular, it presents COMO (Compositional Morphosyntactic Ontology), which aims to be a universally valid classification for distinctions at this linguistic level. We are aware that there is disagreement on whether it is possible to set a universal POS classification (Evans and Levinson 2009). However, we also appreciate that there is a set of coarse POS classes that can be commonly found across many languages.

Some of the projects listed above already present a reference model for that information level, and therefore have served as very valuable starting points for our work. However, the requirements

<sup>1</sup> Full references for each of these projects will be provided in Section 4.

<sup>2</sup> <https://www.oxforddictionaries.com/ogl>



imposed by the nature of our project and the work activity around it make them not fully useful for our purposes. The paper will thus introduce our classification by first reviewing the project requirements and subsequent design choices, and then comparing these against the morphosyntactic classifications that are most relevant here.

## 2 Requirements and System Design

Our proposal is based on a set of requirements imposed by the wide linguistic scope of the OGL program, the nature of activities and resources managed in the Dictionaries Division at OUP, and the subsequent use cases that we need to serve. These requirements determined our approach and shaped the design of the universal classification put forward in this paper. They are presented in the following subsections.

### 2.1 Cross-linguistic Validity

An OGL classification of morphosyntactic information must be able to serve across languages of very different typologies, which therefore diverge greatly in how they encode grammatical distinctions. A feature may be expressed in some languages via morphological mechanisms (for example, verbal tense in Romance languages), encoded in others using independent particles, and in other languages not expressed at all (e.g., Chinese). As a result, dictionary-based classifications of POS categories and associated grammatical information tend to be modeled on the language (or languages) targeted in each individual project. By contrast, a universal classification needs to be able to reflect the commonalities across languages in spite of the diverse means of organizing information through grammar in each language system. Two situations can be distinguished here.

#### 2.1.1 *Considerations on distinctions at the POS category level*

Grammatical distinctions are in some languages manifested at the morphological level, while in others are expressed by means of independent lexical units. An example of that involving English concerns the grammatical feature of verbal infinitive. Whereas in many languages this is realized as part of the verb morphology, in English it is expressed by means of the marker *to*. As a result, it is not uncommon that languages in which elements like these are independent lexical units classify them with an ad hoc POS category not applicable to other languages. Such POS categories therefore lack universal validity.

This problem can be avoided if POS categories are defined appealing to basic configurational properties. For instance, syntactically, what do they combine with to form more complex units (or phrases), and what are the roles that they play in the grammatical organization of the sentence; morphologically, what type of inflection process, if any, do they undergo; etc. From this perspective, independent lexical units dedicated only to encoding grammatical information can be classified as POS categories of cross-lingual validity, most often of closed-class type, such as particles or adpositions. For example, the marker for concord, commonly considered as an independent POS in Bantu languages, can actually be classified as belonging to the more general category of affix.

#### 2.1.2 *Considerations on morphosyntactic distinctions within POS categories*

Many POS classifications tend to subclassify these based on the morphosyntactic distinctions that seem more relevant in each language. Taking verbs as example, we see that in the English tradition, a common way of subclassifying them is distinguishing between lexical and auxiliary verbs. By

contrast in Russian, where aspect distinctions are fully lexicalised, verbs are subclassified into perfective and imperfective, while in the tradition of some Romance languages the direct subdivision of verbs is set in terms of their subcategorization structure (transitive, intransitive, etc.).

Nevertheless, a system where POS classes are tied to particular grammatical features will fail the purpose of being valid across languages. First, each POS will have to subdivide into as many sub-classifications as are found across the different languages covered. This situation will get even less manageable considering that for some POS categories, there may be several levels of information that apply. For example in Spanish, verbs can be simultaneously featured according to their subcategorization and pronominal properties, while in English, nouns are categorized based on their type (common, proper) and countability properties (mass, countable).

Second, the system will introduce a lot of redundancy given that many of the grammatical distinctions are shared across POS classes. For instance, degree features (positive, comparative, superlative) can be found in adjectives and adverbs. Similarly, in many languages the distinctions used in the classification of determiners (e.g., demonstrative, exclamative, indefinite, interrogative, possessive) also apply to pronouns.

Therefore, a universal POS classification that includes all POS classes and subclasses independently set for each language is not viable for our purposes. Instead, we propose a system that represents the grammatical distinctions that are possible in any language by means of a set of features complementary (and therefore orthogonal) to the basic POS ontology. Each feature is modelled as an attribute with its respective set of values, e.g., attribute *number* has as possible values: *singular*, *plural*, *dual*, *trial*, *invariable*, etc.

This results in a highly compositional approach in which the list of POS categories is kept to a minimum number of distinctions, as universally valid as possible and therefore general enough to serve languages of very different typologies. Furthermore, any POS class can in addition be qualified with several attribute-value pairs of grammatical distinctions. For example, a Spanish '*impersonal transitive verb*' will bear the pairs *pronominal\_type:impersonal*, *subcategorization:transitive*.

## 2.2 Respectful of the Grammatical Tradition In Each Language

Each language is described by its own grammatical tradition, as reflected in the way it is presented and taught in grammar books, dictionaries, etc. A major target of our work here is precisely producing and publishing dictionaries for different languages, and therefore the linguistic classifications used should be in agreement with those commonly assumed in the grammatical tradition of each language.

However, grammatical classifications in each tradition tend to be constrained to the language they describe, therefore precluding a wider, cross-linguistic view of grammar distinctions, which is what is aimed here. Two additional issues resulting from adopting the terminology in each language tradition are *feature redundancy* and *concept conflation*. The former occurs when a classification has different terms for the same notion, as the result of inheriting the terms and definitions from different language descriptions. The latter refers to a situation in which the same term is used for two different concepts, due to the fact that different languages' descriptions use it in different ways (e.g., the example with the term *absolute* presented above).

Overall, these issues derive from a wider problem, namely, that of heterogeneous linguistic descriptions constraining the reusability and interoperability of linguistic datasets and resources, which has led to a quite intense area of work in the last decade (see, e.g., Chiarcos & Erjavec 2011; Ide et al. 2017). Two solutions have been put forward in order to enhance the consistency of linguistic categorizations across languages:

- Using *cross-linguistic meta schemes* that include a fix set of content categories to be adopted by all languages. This is the approach adopted by the early initiatives in the field, such as EAGLES (Leech & Wilson 1996) and MULTEXT (Erjavec 2010, 2012).
- Providing a *reference terminology as interlingua* between the different linguistic encodings. Terms in this reference model must be defined so there is no ambiguity in their usage, and to avoid feature redundancy and concept conflation. In addition, a set of linking specifications must be developed for each language-specific morphosyntactic classification, to ensure proper mappings between each resource and the interlingua terminology. Linguistic terminologies such as GOLD (Farrar & Langendoen 2003), ISOcat (Kemps-Snijders et al. 2009), and OLIA (Chiarcos & Sukhareva 2015) assume this approach.

We adopt the second solution: COMO is conceived as a reference model facilitating the harmonization of terms and distinctions used in language-specific resources and applications, which are then able to maintain their original model and yet communicate via a set of mapping models.

### 2.3 Able to Serve Different Use Cases

COMO must be able to support different use cases: from the encoding of dictionary content to the annotation of corpus data, passing through the codification of the NLP tools used for the automatic processing of language. Because of this, the morphosyntactic classification must be able to account not only for the prototypical POS classes and their grammatical distinctions, but also for other kinds of units.

Corpus content and NLP tools, for example, require sensitivity to punctuation marks or non-standard lexical elements, such as symbols. We decided to set a category for these at the same level as more standard POS categories.

Similarly, the lexicographical use cases of OGL require the inclusion of other elements not typically considered to be POS categories. These are: parts of lexical units (i.e., affixes) as well as aggregates of lexical units, such as contractions or different types of multiword expressions (phrases, idioms, etc.).

## 3 The Compositional Morphosyntactic Ontology (COMO)

The Compositional Morphosyntactic Ontology (COMO) is a repository of linguistic terminology that provides common ground for languages of very different type, which historically have been described through diverse grammatical traditions. It enables harmonization of linguistic annotations in different types of resources, such as lexicographical datasets (including machine-readable dictionaries and dictionaries for human consumption), text corpora, and language processing tools.

COMO is not conceived as a cross-lingual meta-scheme to be assumed by all languages and resources. Rather it is an interlingua, a reference model to which morphosyntactic annotations in different resources and languages map in order to allow for maximum interoperability.

Because of this, it was vital that no language was a stronger driver than any other in the setting of a POS classification.

COMO defines and organizes morphosyntactic concepts in an ontological structure by appealing to criteria and definitions of cross-lingual validity as much as possible.

### 3.1 Lexical Categories

Table 1 presents the full set of lexical categories (aka POS classes) in COMO. It provides comments defining the category only when deemed necessary.

Table 1: POS classes and subclasses in COMO

Lexical category	Comments	
adjective		
adposition	Cover term for a closed class of words that express spatial or temporal relations, or mark various semantic roles. Typically combining with one complement, generally a noun phrase. Divided into the following three subclasses based on the position they take with respect to the complement: before (preposition), after (postposition) or surrounding it (circumposition). Possible subclasses are:	
	circumposition	Consisting of two parts that appear on each side of the complement.
	postposition	Following the complement.
	preposition	Preceding the complement, as in English.
adverb		
affix	Morpheme attached to a stem to form a new word. In some cases it is written as part of the same word whereas in others it appears as an independent element.	
	circumfix	Two separated parts appearing on each part of the stem.
	combining_form	Word normally used in compounds in combination with another element to form a word (e.g. <i>Anglo-</i> ‘English’ in <i>Anglo-Irish</i> ).
	infix	Appearing within the stem.
	prefix	Appearing before the stem.
	suffix	Appearing after the stem.
article		
conjunction		
contraction	Combination of two or more words belonging to a different POS into a single lexical unit. For example, the combinations of preposition + article in French: <i>du</i> ( <i>de+le</i> ).	
determiner		
ideophone	Ideophones are lexical units that evoke a vivid impression of certain sensations or sensory perceptions (e.g. sound meow for a cat), movement, colour (e.g., English <i>bling</i> , describing the glinting of light on things like gold), shape, action ( <i>ta-da!</i> ), etc. It is a lexical class based on the special relation between form and meaning. In some languages, ideophones correspond to common POS classes (e.g., adjectives, adverbs, etc.), but in others they are an independent POS.	
idiomatic	Multiword, phrasal or clausal expressions, generally with no compositional interpretation.	
interjection		
noun		
numeral		
particle	Particles must be associated with another word or phrase to impart meaning. They typically encode grammatical distinctions like negation, mood, tense, or case), etc. However, they cannot be classified as other main POS, including functional ones, such as prepositions, conjunctions, etc.	
predeterminer		
pronoun		
punctuation	left_parenth_punc	Left parenthetical punctuation mark, e.g., (, [.
	right_parenth_punc	Right parenthetical punctuation mark, e.g., ), }
	sentence_final	Sentence final punctuation mark
	sentence_medial	Sentence medial punctuation mark
residual	Cover class for non-standard forms, such as symbols or digits.	
verb		

POS classes have been determined according to the basic configuration (grammatical and syntactic) properties of the elements being classified. For example, syntactically, what does a lexical element combine with in order to form a more complex unit, or what is the role that it plays in the grammatical organization of the sentence; morphologically, what type of inflection process, if any, does it allow? And so on.

Since POS is essentially a configurational distinction, we have established subclasses only if they respond to strictly positional criteria that can be applied across languages. In particular, the classes *adposition*, *affix*, and *punctuation* have been subdivided based the possible placements of their elements. Note that these subclassifications are different from those in other reference models, such as GOLD or OLIA, which use grammatical distinctions to subclassify POS categories (e.g., *transitive verb* or *demonstrative determiner*). The classification also includes classes for material not traditionally considered part of the grammar, such as punctuation marks (class *punctuation*), or symbols (under class *residual*).

### 3.2 Grammatical Features

Complementing the POS classification, there is a set of features covering the different grammatical distinctions that can be manifested in a language, from those most commonly found in Indo-European languages (like tense, case, number, and gender) to others associated with languages which have less coverage in terms of linguistic encoding and language resources (for example, concord for Bantu languages, or classifier type for Chinese and other languages). We refer to these features as the set of *grammatical features*.

Previous proposals of morphosyntactic classifications also account for grammatical distinctions. EAGLES (Leech and Wilson 1996) and MULTTEXT (Erjavec 2010, 2012) are designed as *positional tagsets*, that is, as sets of tags in which each tag is modelled as an acronym, and where each piece of information (POS class and the possible grammatical features for that POS class) is represented in a specific position of the tag with a one-character code identifying the corresponding value. For example, in EAGLES, a common noun, feminine singular and in accusative case would be expressed as a tag like: *NCFSA*. The problem with this approach, however, is that it imposes a rigid structure. This makes it difficult to add new features as languages with yet uncovered properties (e.g., noun class for some Bantu languages) are added to the repository.

Other morphosyntactic classifications have adopted a more flexible approach by distinguishing between POS categories on the one hand, and morphosyntactic features on the other; e.g., GOLD (Farrar & Langendoen 2003) and OLIA (Chiarcos & Sukhareva 2015). Nevertheless, they still present at least one level of POS subcategory that tends to be language-specific, and that leads to the issues identified in section 2.1.2.

By contrast to all this previous work, COMO adopts a fully compositional approach, along the same lines as the Universal Dependencies (UD)<sup>3</sup> and the Universal Morphological Annotation (UniMorph)<sup>4</sup> proposals, where grammatical distinctions of all kinds are represented as information independent from the POS categories. In particular, each feature is modelled as an attribute with its corresponding set of values.

A key element in this approach is the level of granularity of the feature. It is important to separate into different features grammatical information that may manifest simultaneously but in fact corresponds to different notions, as some languages may present one feature but not the other. For instance, event

3 <http://universaldependencies.org/> [25/03/2018]

4 <http://unimorph.org/> [25/03/2018]



duration (distinctions of punctual vs. non-punctual) and telicity (distinctions of telic vs. atelic) should be considered as independent from each other, since they can combine in different ways.

There is no imposed hierarchy of one feature over another, thus avoiding conflicts between language traditions. Moreover, new features and values can be added as further languages are included in the OGL project, without having an impact on the previously described languages.

The rich morphosyntactic branch in OLIA was a solid base to model this kind of information. However, some of the languages that we intend to model have complex morphosyntactic features, such as Northern Sotho, wherein a single orthographic word may contain a number of morphemes; others have extensive noun systems, such as isiZulu, which has 17 different classes. OLIA was unable to fully accommodate modelling of such features.

Our proposal was also informed with the lexical and grammatical labels used in OUP monolingual and bilingual dictionaries. Finally, a further source of information was other well-tested classifications for morphosyntactic knowledge developed with multiple languages in mind or as a collaborative effort among teams in different countries (EAGLES, MULTEXT).

At the moment, COMO presents 46 grammatical features, which range from the basic nominal, verbal and adjectival morphological features present in many languages (e.g., *number, gender, case, degree, person, mood, tense*, etc.) to elements codifying syntax (e.g., *subcategorization patterns*), distinctions at the lexical semantic level (e.g., *aspect, telicity, countability*), or pragmatic information (e.g., *definiteness, referentiality, evidentiality, sentence modality*). Some of these categorizations are present in multiple POS classes (*degree, number, gender*), whereas others are particular to only one (*voice, pronoun function*). Furthermore, most categorizations are shared across several languages, although a few cases had to be tailored to specific ones, such as the classifications on *diptoticity* or *verb form type* for Arabic.

## 4 Comparison with Other Morphosyntactic Classifications

The last few decades have seen a significant effort, especially from the NLP community, to develop universal classifications for the description of different levels of linguistic information (lexical, morphological, syntactic). Early work in this respect includes initiatives on corpus annotation, like the EAGLES guidelines for annotating morphosyntactic information in corpora and lexicons (Leech & Wilson 1996), which led to the development of POS tagsets such as FreeLing<sup>5</sup> (Carreras et al. 2004) and MULTEXT-East (Erjavec 2010), among others.

This section describes the most relevant morphosyntactic classifications aiming to address interoperability issues at the level of linguistic information, and assesses them against the requirements introduced in Section 2.

### 4.1 MULTEXT-East

The MULTEXT-East morphosyntactic specifications<sup>6</sup> (Erjavec 2010, 2012) provide attributes and values for annotating at the word level, focusing in particular on 16 Eastern European languages. This system distinguishes 14 main POS classes (called morphosyntactic categories in the proposal), each of them possibly splitting into several levels of subcategories. Considering all possible categories and subcategories, there are a total of 127 morphosyntactic classes. This approach has the issues

<sup>5</sup> <https://talp-upc.gitbooks.io/freeling-4-0-user-manual/content/tagsets.html> [25/03/2018]

<sup>6</sup> <http://nl.ijs.si/ME/owl/> [25/03/2018]

and risks presented in Section 2.1.2 with regard to cross-linguistic validity. In fact, some distinctions seem quite clearly based on the grammatical traditions of different languages. For example, the differentiation between *DemonstrativeQuantifier* and *DemonstrativeDeterminer*, respectively under the *Quantifier* and *Determiner* categories, is not clear. The same can be observed with regard to the grammatical features complementing the POS classification. For example, the description of the feature *SyntacticType* is different depending on the language it is used for. In general, MULTEXT-East seems unable to successfully handle languages others than those for which it was developed.

## 4.2 General Ontology of Linguistic Descriptions (GOLD)

GOLD<sup>7</sup> (Farrar & Langendoen 2003) is a very complete ontology accounting for information at several levels of linguistic description, from phonetic to morphosemantic properties, and covering also structural unit or human language variety. The layers that are of interest here are *POS Properties* and *Morphosyntactic Properties*. POS classes are grounded on very strong language-agnostic principles that focus on the function of the item in the sentence. As a result, what are usual terms in our linguistic tradition, such as *verb*, *conjunction*, or *pronoun*, are nested under more general terms such as *Predictor*, *Functor*, or *Pro Form*, respectively. There are 19 main POS classes, which can then split into several layers of subclasses, leading to a total of 91 classes.

GOLD offers a very interesting perspective on POS classification because of its language-independent approach. It is a terminology repository very comparable to COMO in purpose, scope, and technical approach. First, it was created to address the problem of having different markup schemes for annotating linguistic data, and therefore to provide a unified description of data in different languages. More specifically, it originated to handle the annotation and description of endangered languages, a purpose very close to OGL's main goals. Furthermore, it seeks to be compatible with Semantic Web approaches, and therefore to enable automated reasoning over linguistic data.

There are, however, several reasons that prevented us from adopting it. Firstly, its linguistic terminology is quite different from that more commonly used in dictionaries and other language resources. Secondly, although it contains most of the more standard concepts in our tradition (e.g., verb, adjective, conjunction), it organizes them in a structure more complex than what is actually needed for our purposes. For example, the class for subordinating conjunctions is in level 3 of the embedding, within *Connective* within *Functor*. Finally, in spite of its wide coverage of linguistic phenomena, it does not account for categories that are important when dealing with text annotation and language processing tools, such as punctuation marks and symbols.

## 4.3 ISO TC37/SC4 Data Category Registry (ISOcat)

ISOcat (Kemps-Snijders et al. 2009) is a web-based repository of linguistic terminology providing uniform naming and semantic descriptions in order to facilitate interoperability of a wide range of resource and application types.

It took a community-driven approach from quite early in its development, allowing everybody to extend the repository based on specific languages or project needs. That approach led to the proliferation of data categories, and caused issues of redundancy (different terms defining the same notion) or concept conflation (the same term for two or more notions). See Ide, Calzolari et al. (2017:136—139) for a more comprehensive description of the situation.

With regard to the requirements of our project, these problems translate into the presence of POS categories specific for only one language, and a poor handling of the top level classification for POS

<sup>7</sup> <http://www.linguistics-ontology.org/gold/2008> [25/03/2018]

tags, with categories that can be grouped under more general classes (e.g., the different punctuation mark signs). In addition, ISOcat also includes subcategories as part of the POS classification, which leads to more than 100 POS classes and therefore carries the issues identified earlier with this type of approach (refer to section 2.1.2).

#### 4.4 Ontologies of Linguistic Annotation (OLIA)

OLIA<sup>8</sup> (Chiarcos & Sukhareva 2015) was developed to enable semantic interoperability among linguistic resources of diverse types and annotated at different levels of analysis, i.e., from phonology all the way up to discourse structure. Unlike other initiatives that aim at the same purpose (e.g., GOLD, ISOcat), OLIA is a complete architecture that includes: (a) a *reference model* with terms and their definitions, to be used as interlingua for mapping linguistic resources of different provenance and tradition; (b) specific *annotation models* for different levels of linguistic description (morphology, morphosyntax, syntax, discourse), consisting of annotation schemes and tagsets for more than 85 languages; (c) a set of *linking models* mapping concepts and properties in each of these annotation schemes to the reference terminology; and finally also (d) a set of *linking models for mapping external annotation models* to the OLIA reference model (as is the case for MULTTEXT-East, GOLD, ISOcat, or Universal Dependencies). Moreover, OLIA ontologies can serve as a centralizing hub for linguistic data categories within the Linguistic Linked Open Data (LLOD) framework<sup>9</sup> (Chiarcos et al. 2013).

OLIA is a very rich and powerful resource that already serves two of our requirements in Section 2. First, given its purpose of serving different types of linguistic resources and applications, the OLIA reference model can address the use cases we presented in Section 2.3. That is, it accounts for categories that go beyond standard POS classes but that are needed for NLP applications (e.g., punctuation marks, symbols) or lexicographic products (contractions, multiword expressions). Second, its four-tier architecture allows us to be respectful of the grammatical tradition of each language or the annotation schemes already adopted in different corpora and NLP tools (a requirement in Section 2.2).

However, OLIA cannot equally well serve our need for a classification of POS categories and grammatical features with cross-linguistic validity (a requirement in section 2.1). The POS classification in its reference model follows the standard approach of splitting POS categories into subcategories depending on the features that are considered more prominent for that category in each language. This leads to an enumerative model, where, for example, a common noun belongs to a different class (*CommonNoun*) than a mass noun (*MassNoun*), and thus the linking model must then take care of cases in which the two concepts apply to the same linguistic unit.

This enumerative approach also results in some unclear areas in the classification, especially regarding elements of lesser presence in Indo-European languages. There is for example the *Unique* class, which is conceived in accordance with the definition for this same category in EAGLES, as approximating “the linguistic concept *Particle*. It covers categories with unique or very small membership (...)”.<sup>10</sup> In OLIA, this class is split into almost 30 subcategories that would be handled more coherently and systematically using a higher compositional approach, like the one adopted in COMO. *Particle* is a key category in any morphosyntactic classification aspiring at universal validity.

8 <http://acoli.cs.uni-frankfurt.de/resources/olia/> [25/03/2018]

9 <http://www.acoli.informatik.uni-frankfurt.de/resources/lod/> [25/03/2018]

10 <http://www.ilc.cnr.it/EAGLES96/annotate/node16.html#mp> [25/03/2018]

## 4.5 Universal Dependencies (UD)

UD<sup>11</sup> is a community-based project that has the goal of developing grammatical annotations consistent across languages, with the purpose of enabling the development of NLP tools, and therefore annotated corpora, from a language typology perspective. In particular, UD provides tagsets for syntactic dependencies, POS categories, and morphosyntactic features, respectively based on the set of Stanford dependencies (de Marneffe et al. 2008, 2014), Google universal POS tags (Petrov et al. 2012), and the Intersect interlingua for morphosyntactic tagsets (Zeman 2008).

Specifically concerning POS tags in UD, the classification results from an attempt to simplify the quite fine-grained categories employed in most treebanks, after some analysis and experiments suggesting that coarser POS classes could help in the NLP tasks of POS taggers and grammar induction. Resulting from this, UD defines 12 basic POS categories and a set of grammatical attributes with their possible values. In addition, it provides mappings for a number of treebanks, accounting for more than 60 languages of very different typologies.

Our work shares the foundations and main guidelines of the UD project: cross-linguistic validity and suitability for consistent tagging in corpora and NLP tools, disregarding the grammar tradition assumed in the resource. Consequently, there are the following key common elements between the two projects:

- The interlingua architecture approach, with a terminological repository set as reference model across languages, and then a set of linking models for mapping the annotations in different resources and traditions to the concepts in the reference model.
- The compositional approach to morphosyntactic categorization, that is, by means of a set of basic cross-linguistically valid POS classes, which is further complemented with a set of grammatical features.
- A minimum set of coarse POS classes.

The main difference between UD morphosyntactic classification and COMO, however, has to do with the requirement of being able to encode content present in dictionary products, such as parts of words (affixes) and word aggregates (contractions, idioms, etc.).

## 4.6 Universal Morphological Feature Schema (UniMorph)

Similar to UD, UniMorph<sup>12</sup> is a collaborative project developing a cross-linguistically valid morphological classification to be used for training language processing tools. UniMorph aims to overcome the problems of previous classifications that are strongly driven by one, or a particular set of, languages. Thus, it is conceived as an interlingua that allows for syntactic and semantic interoperability among linguistic resources.

UniMorph does not provide POS categories. It focusses only on inflectional morphology categories that represent dimensions of meaning (e.g., number, gender, case, degree). It includes 23 dimensions (equivalent to our grammatical features), which are further specified by means of one or more features (corresponding to our values) from a set of 212 items (Sylak-Glassman 2016).

UniMorph has important similarities with our model, such as the concern for an interlingua capable of avoiding the issues of feature redundancy and concept conflation, the strongly hand-engineered approach to ensure this, and an emphasis on identifying what can be considered the semantic atoms (that is, the meaning elements that cannot be decomposed into a more fine-grained interpretation units) in order to guarantee a compositional approach.

11 <http://universaldependencies.org/> [25/03/2018]

12 <http://unimorph.org/> [25/03/2018]



Nevertheless, it also differs significantly from our proposal. First, it does not provide a classification of POS categories that can be considered valid across languages. Second, it is built top-down by reviewing the language typology literature and gathering the linguistic traits that are described for the different languages, whereas our approach is bottom-up, accounting for what it is brought in by the languages under consideration, and adapting the system as needed. Finally, it includes only distinctions introduced by inflectional morphology, whereas COMO accounts also for traits pertinent to other levels of the linguistic description as long as they are encoded as part of the information in lexical units. Examples of this include subcategorization type (e.g., transitive, intransitive, bitransitive, etc.), numeral type (cardinal, fraction, multiplier, ordinal, etc.), or pronoun scope (exclusive, inclusive).

## 5 Status and Future Work

We have introduced COMO, a classification of POS categories and their associated grammatical features, which aims to be valid across languages of very different typologies. POS classes are defined according to basic configurational properties, leaving all grammatical distinctions to be modelled by a set of complementary features, which are handled compositionally and do not bear any hierarchical organization, contrary to most previous proposals. COMO is conceived as a reference model within a wider data architecture, and therefore acts as interlingua that guarantees semantic interoperability while preserving the original encoding of linguistic resources.

Currently, COMO is deployed in RDF and JSON formats, and some language resources are modelled according to that. It is used to support lexical content relating to 20 languages of disparate families, including Indo-European (e.g. English, Spanish, Hindi, Urdu), Austronesian (Malay, Indonesian), Bantu (Swahili, isiXhosa, Setswana), and Quechuan (Southern Quechua).

COMO is still under development, as new languages are added to the OGL program. However, it already serves a wide range of morphosyntactically manifested phenomena. In fact, the classification has been actively used from its inception, and organically tested and enriched with the inclusion of new languages in the program.

In the near future COMO will be made available to the general public. Additional further work involves formalizing the mapping models between the different lexical resources and COMO reference model, as well as developing and deploying a URI strategy that allows this content to be moved into a linked data paradigm. Longer term, we plan to also generate the necessary mapping models to connect our content to external models (e.g., UD, UniMorph) and lexical resources.

## References

- Carreras, X., I. Chao, L. Padró, M. Padró (2004) FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the Language Resources and Evaluation Conference LREC 2014*: 239-242.
- Chiarcos, C., and T. Erjavec (2011) OWL/DL formalization of the MULTTEXT-EAST morphosyntactic specifications. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*: 11—20.
- Chiarcos, C., J. McCrae, P. Cimiano, C. Fellbaum (2013) Towards Open Data for Linguistics: Linguistic Linked Data. In A. Oltramari et al. (eds.) *New Trends of Research in Ontologies and Lexical Resources*. Theory and Applications of Natural language Processing. Springer-Verlag, Berlin Heidelberg.
- Chiarcos, C., and M. Sukhareva (2015) OLiA – Ontologies of Linguistic Annotation. In *Semantic Web*, vol. 6(4): 379-386.
- de Marneffe, M-C., T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C.D. Manning (2014) Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of Language Resources and Evaluation Conference 2014*.



- de Marneffe, M-C. and C.D. Manning (2008) The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.
- Erjavec, Tomaž. (2010). MULTTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17–23
- Erjavec, T. (2012) MULTTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1): 131-142.
- Evans, N. and S. Levinson (2009) The myth of language universals: Language diversity and its importance for cognitive science. In: *Behavioral and Brain Sciences*, 32(5).
- Farrar, S., T. Langendoen (2003) A linguistic ontology for the semantic web. In: *Glott International*, 7(3): 97–100.
- Ide, N., N. Calzolari, J. Eckle-Kohler, D. Gibbon, S. Hellmann, K. Lee, J. Nivre, L. Romary (2017) Community Standards for Linguistically-Annotated Resources. In: Ide N., Pustejovsky J. (eds) *Handbook of Linguistic Annotation*. Springer, Dordrecht: 113:165.
- Ide, N. and J. Pustejovsky (2010) What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proceedings of the Second International Conference Global Interoperability for Language Resources*. Hong Kong.
- Kemps-Snijders, M., M. Windhouwer, P. Wittenburg, S.E. Wright (2009) ISOcat: Remodelling metadata for language resources. In *International Journal of Metadata, Semantics and Ontologies* 8(4): 261–276.
- Leech, G., and A. Wilson (1996) *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. EAG--TCWG--MAC/R*. Version of March, 1996. <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html> [25/03/2018]
- Parvizi, A., M. Kohl, M. González, R. Saurí (2016) Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of LREC 2016*.
- Petrov, S., D. Das, and R. McDonald (2012) A universal part-of-speech tagset. In *Proceedings of Language Resources and Evaluation Conference 2012*.
- Sylak-Glassman, J. (2016) *The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema)*. Working Draft v.2. Center for Language and Speech Processing. John Hopkins University. Manuscript. <https://unimorph.github.io/doc/unimorph-schema.pdf> [23/05/2018]
- Zeman, Daniel (2008) Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of Language Resources and Evaluation Conference 2008*.

## Acknowledgements

We want to thank Imogen Foxell and Tressy Arts for their thorough considerations of various lexical and grammatical issues around languages that fall quite far from our linguistic knowledge, and which have help shape the ontology put forward here in very positive ways.



# Commonly Confused Words in Contrastive and Dynamic Dictionary Entries

**Petra Storjohann**

*Institut für Deutsche Sprache, Mannheim*

*E-mail: storjohann@ids-mannheim.de*

## Abstract

This paper discusses changes in lexicographic traditions with respect to contrastive dictionary entries and dynamic, on-demand e-lexicographic descriptions. The new German online dictionary *Paronyme - Dynamisch im Kontrast* is concerned with easily confused words (paronyms), such as *effektiv/effizient* and *sensibel/sensitiv*. New approaches to the empirical analysis and lexicographic presentation of words such as these are required, and this dictionary is committed to overcoming the discrepancy between traditional practice and insights from language use. As a corpus-guided reference work, it strives to adequately reflect not only authentic use in situations of actual communication, but also cognitive ideas such as conceptual structure, categorization and knowledge. Looking up easily confused lexical items requires contrastive entries where users can instantly compare meaning, contexts and reference. Adaptable access to lexicographic details and variable search options offer different foci and perspectives on linguistic information, and authentic examples reflect prototypical structures. These are essential in order to meet all the different interests of users. This paper will illustrate the contrastive structure of the new e-dictionary and demonstrate which information can be compared. It also focusses on various dynamic modes of dictionary consultation, which enable users to shift perspectives on paronyms accordingly.

**Keywords:** paronyms, dynamic lexicography, contrastive entries, generating information on demand

## 1 Introduction

In German, as in other languages, there are lexical items with related morphological roots and similarities in sound, spelling and/or meaning. These are easily confused by native speakers and language learners. They are referred to as paronyms, and examples include *effektiv/effizient*, *legal/legitim*, *autoritär/autoritativ* and *sensibel/sensitiv*. Paronyms have so far only been documented in two printed dictionaries (Müller 1973; Pollmann & Wolk 2010)<sup>1</sup>. Both reference books base their descriptions on introspection and arbitrary evidence taken from fiction. They are traditional dictionaries, prescriptive in nature, their entries are limited to specific aspects of meaning, and they do not cover polysemous items in detail. Their central aim is to guide users to the allegedly correct use and describe a clear distinction between the items in question.

With respect to German, there is no empirically sound and user-friendly dictionary covering paronyms in contemporary language use. The aim of the project “Paronymwörterbuch” has been to carry out a thorough examination of paronyms, based on their usage in context, using contrastive corpus-linguistic methods. It takes a descriptive and empirical approach, and documents easily confused words in German with respect to their actual use in public discourse. Its analyses and descriptions are corpus-based and cover contemporary German. The results of this work are explanatory, contrastive entries in a new dynamic and multifunctional e-dictionary called *Paronyme - Dynamisch im Kontrast* (*Paronyms*

<sup>1</sup> As well as paronyms, both dictionaries also contain a number of other cases of confusable words, e.g. homophones, synonyms, homographs etc.

– *Dynamic in Contrast*). To date, some 100 paronym pairs/sets, mostly adjectives, have been compiled. The corpus-derived list of paronyms has revealed that the phenomenon of paronymy is much larger than previously thought, including almost 2,000 cases which range from commonly known word pairs to specific technical terms and a large number of compounds (cf. Schnörch 2015).<sup>2</sup> Since 2018, the new online resource has been freely accessible via the linguistic platform OWID<sup>plus</sup>.<sup>3</sup> This reference work breaks new ground with respect to the following four issues:

Firstly, as a corpus-guided dictionary, it is descriptive in nature, documenting conventionalized patterns and use, including more recent language changes. It also provides information about preferences and tendencies rather than following prescriptive traditions. Secondly, in order to meet meta-lexicographic demands (cf. Rundell 2012; Kövecses & Csábi 2014) it combines corpus-based methods with cognitive semantics. Entries contain linguistic details which are consistently paired up with conceptual-encyclopedic information (cf. Storjohann 2017). It strives to adequately reflect ideas such as conceptual structure, categorization and knowledge. In this way, the linguistic findings correlate better with how users conceptualize language. Thirdly, an entry can consist of up to three lexical items (e.g. *effektiv/effizient/effektiv*, *unsozial/asozial/antisozial*, *praktisch/praktikabel/praktizierbar*) and it is exclusively designed for contrastive consultation processes. Paronyms are directly compared with each other in visual and explanatory ways. This enables readers to discriminate between definitions, collocates, citations, constructions, sense-related terms and conceptual categories designated by the paronyms, as well as the referential domains in which they predominantly occur. Fourthly, the project analyzed users' needs before developing the dictionary and compiling the data. These investigations revealed that usage modalities needed to be rethought and a flexible approach to information adopted, to enable different needs to be met. As a consequence, information can be flexibly adapted and dynamically generated following different navigation and menu options.<sup>4</sup>

The objective of this paper is to introduce the new dictionary and illustrate its essential functions with the help of examples. While the integration of corpus-linguistic findings and cognitive features has been explored in Storjohann (2017a, 2017b), this paper is concerned with the realization of contrastive meaning descriptions and dynamic e-lexicographic entries. The contrastive structure of the entries will be elucidated, and I will also show how this dictionary has moved away from static to dynamic e-presentation by incorporating flexible dictionary consultation options.

## 2 Why Contrastive and Dynamic Structures?

In recent years, the focus of numerous meta-lexicographic studies of German electronic dictionaries has been on users. These studies have helped lexicographers to develop a better understanding of the needs of users and their dictionary consultation behavior (e.g. Müller-Spitzer 2014). By scrutinizing a variety of language forums, the project “Paronymwörterbuch” has been able to answer numerous questions regarding its users in advance. As Storrer (2013) argues, “professional lexicographers may learn about their users' needs by studying the topics discussed in these projects” (Storrer 2013: 1251). Today, online forums are widely used social media sources, where people share their concerns about easily confused words and heterogeneous user groups consult the community about their linguistic problems. Through the study of language blogs, we have gained detailed insights into the specific linguistic problems of users, their consultation routines and their needs. They all face situations of linguistic doubt and many are familiar with well-established dictionaries. Nonetheless, in a number

2 Most paronyms are adjectives, but there are also verbs (e.g. *beenden/beendigen*) and nouns (e.g. *Methode/Methodik/Methodologie*).

3 OWID<sup>plus</sup> is a platform for multilingual lexical-lexicographic data, quantitative lexical analyses and interactive lexical applications.

4 To the best of our knowledge, there is no usage-based lexical resource in other languages that is similarly concerned with paronyms (or synonyms, antonyms for that matter) in the same comparative way.

of cases, we have learned that they consult online forums because existing dictionaries do not provide them with satisfactory answers to their linguistic problems:

[...] these days people wanting to know about word usage may well tweet their question to hundreds of followers, ask it on an internet forum, or email it to a language blogger. Sometimes they do so after encountering ‘dictionary-based problems’. (Murphy 2013: 287)

In a more or less detailed way, native users and language learners explain whole contextual situations in which their uncertainties occur. Studying these, we have learned that their interests in commonly confused words vary. They look for answers as to specific lexical use, usual contexts, possible constructions, and conceptual as well as encyclopedic issues. The answers of the language community are just as diverse and revealing (cf. Storjohann 2015). Overall, we have found that speakers have good intuitions as to what linguistic and extra-linguistic information is required to form essential parts of authentic communication. As a result, this new project was able to adjust its lexicographic concept and include features which otherwise would not have been part of the dictionary. These concern both the type of linguistic and encyclopedic information<sup>5</sup> and the forms of presentation of paronyms. In order to compile a dictionary which can answer all of the questions raised by users, a number of challenges were encountered when documenting usage-based findings. These concerned two aspects in particular: firstly, how to accommodate all interests in lexical use in an efficient contrastive description and secondly, how to flexibly account for different approaches and requirements. In the following, both aspects will be illuminated in more detail.

## 2.1 Concise Entry Overview and Detailed View

Traditionally, users have had to laboriously look up individual entries when learning about paronym behavior from general monolingual reference guides. One lexical entry had to be compared with another lexical entry in the same or another dictionary, with the user switching between the two. Alternatively, users could consult existing paronym reference books, provided they were familiar with these, where they usually encountered brief normative descriptions of single uses of two confusable words. The central aim of the new German paronym dictionary is to offer contrastive entries of two (or sometimes three) lexical items. In concise and clearly arranged comparative entries, users can instantly familiarize themselves with contextual similarities and differences.

As Figure 1 demonstrates, different types of context and degrees of similarity are visibly marked using a specific ordering system and a color scheme.<sup>6</sup> Contextual instances/uses are divided into categories based on semantic features. The more frequent term of a pair or set occurs at the top of the entry, with all its contextual semantic instances organized in a line. Directly underneath, the second, often less frequent, paronym term is listed with its contextual uses. For example, in Figure 1, each ‘box’ represents a specific contextual use of the adjectives *sportlich* (meaning sporty, sporting, sportsman-like) and *sportiv* (sporty, athletic). Identical semantic contexts (blue) between the two adjectival items occur to the left of the monitor, followed by similar contexts (green), which in turn are followed by different contexts (grey). Instantly, in this entry overview, users can capture semantic parallels (overlaps), which contexts they occur in and which meanings are attributed individually. The purpose of this meaning spectrum is to gain an immediate general overview. The information a user obtains at this point includes a minimum of lexical and encyclopedic details, such as a short definition, a conceptual reference and prototypical examples (a selection of up to five collocates), exemplifying lexical patterns and the reference itself. This concept-driven navigation structure offers a large amount of knowledge about contextual behavior, parallels or differences, meanings and concepts.

<sup>5</sup> For details and examples, see Storjohann (2017a, 2017b).

<sup>6</sup> The color scheme will not be visible in a black-and-white print of this paper.



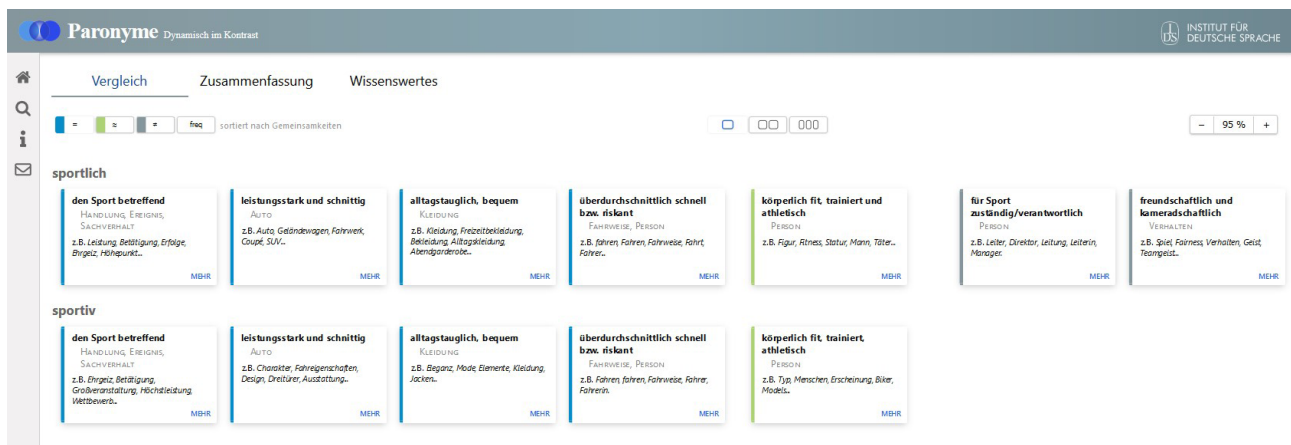


Figure 1: Concise default view with identical contexts first

Among other criteria, it is the referential/ontological categories underneath the short definitions (e.g. HANDLUNG (PROCESS), EREIGNIS (EVENT), AUTO (CAR), KLEIDUNG (CLOTHES), FAHRSTIL (STYLE OF DRIVING) PERSON (PERSON), VERHALTEN (BEHAVIOR)) that enable users to distinguish patterns and help them to encode/decode contexts and identify metonymic and metaphoric mappings (cf. Fillmore & Atkins 1992). Users can more easily relate the adjectives to their meanings and then relate these to the preferred contextual reference (here modified nouns), e.g.

- *sportlich* means ‘concerning sports’ with respect to a PROCESS, EVENT OR MATTER, for example *performance, pursuit, success, ambition, highlight*,
- *sportlich* means ‘powerful and sleek’ with respect to a CAR, for example *car, Sport Utility Vehicle, chassis, SUV, coupé*,
- *sportlich* means ‘comfortable’ with respect to CLOTHES, for example *clothing, leisure wear, apparel, casual clothes, evening dress*.

Optionally, more data can be expanded for each contextual use. In addition, an explanatory definition, more typical lexical combinations, corpus examples, synonyms/antonyms and typical constructions in different contexts can be looked up in a more complex detailed view (see Figure 2). The conceptual references representing encyclopedic ideas [e.g. EVENT, CAR, PERSON] are then explicitly integrated into the more complex definition:

***sportlich***, leistungsstark und schnittig / powerful and sleek

charakterisiert meist ein Auto bzw. dessen Erscheinungsbild dahingehend, dass es z. B. ein tiefergelegtes Fahrwerk sowie stärkere Motorleistung aufweist und optisch schnittig bzw. dynamisch wirkt

Engl.: often characterizes a car or its general appearance as being sleek or seeming dynamic, because of its lowered chassis and stronger engine output

z. B.: Auto, Geländewagen, Fahrwerk, Coupé, SUV, Limousine, Dreitürer, Flitzer, Optik, Aussehen.

e.g.: car, Sport Utility Vehicle, chassis, coupé, SUV, limousine, three-door car, sports car, look, appearance.

The relevant ontological category (in this context CAR) is specifically illustrated by up to ten collocations, which function as examples of lexical preferences in actual contexts.<sup>7</sup>

<sup>7</sup> In some cases, the conceptual/ontological reference is more a domain. This is often the case for nouns and verbs.

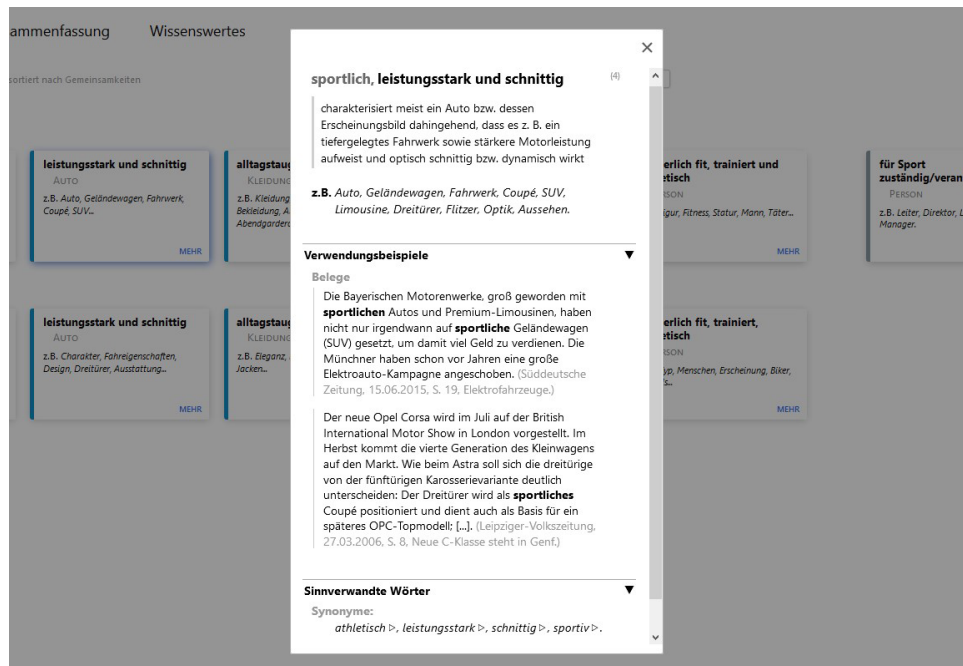


Figure 2: Detailed contextual view

The listed collocates (up to 10) are the result of the interpretation of statistically significant co-occurrences as extracted by the underlying corpus (see Section 4). In essence, they are prototypical domain elements and structured mental representations of human experience. They shed light on strong affinities to constructions and contextual preferences, and show cognitive processes in which conceptual elements motivate the configuration of another semantically related conceptual entity (cf. Kövecses & Csabi 2014). The objective of this detailed presentation is not only to provide more information, but also to combine lexical and encyclopedic details which can then be consulted and mentally stored together.<sup>8</sup>

## 2.2 Dynamic Consultation of Overview

It is commonly agreed that electronic dictionaries in particular should make more use of adaptive structures in order to solve problems of strict macrostructural ordering and to meet users' needs more effectively (e.g. Kwary 2012; Fuertes-Olivera 2013). Once we knew the needs of our potential users, a flexible presentation and visualization of linguistic data for different purposes was a central aim of our project. This meant breaking with a traditional, strictly linear and rigid organization and looking for innovative, flexible and multi-functional forms of presentation.

Different lexicographic products could be created based on one XML-data set relating its presentation for example to different user groups. Realizing tailor-made user-adaptivity is technologically feasible but only realistic once we know more about the users. Contents can be arranged dynamically changing linguistic focus to allow users to recreate and re-represent their own dictionary data. (Fuertes-Olivera 2013: 330)

As users have different concerns about typically confused words, the paronym dictionary took a crucial step back from a static reference guide and moved towards a dynamic reference guide, providing and generating specific information on demand. As can be seen in Figures 1 and 3, two different menu options are visible at the top of an entry beneath the main headings of the sections **Vergleich** (comparison), **Zusammenfassung** (summary) and **Wissenswertes** (other interesting facts). On the left, there is a choice of sorting options with regard to contextual information. These are: sort by identical use (Figure 1), sort by similar use, sort by distinct use and sort by frequency (see Figure 3).

<sup>8</sup> For other dictionary examples, see Storjohann (2017a, b).

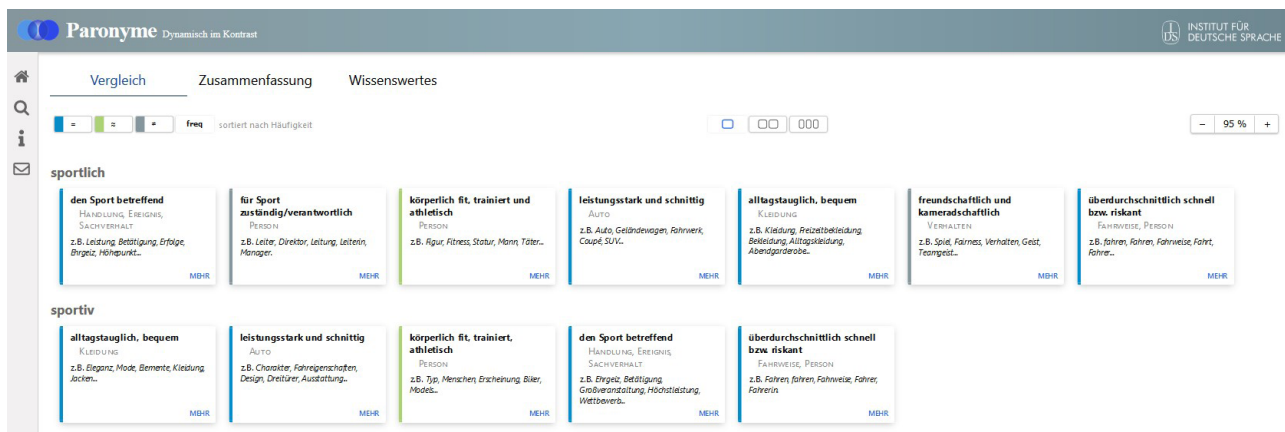


Figure 3: View of contexts listed by frequency

For instance, some users are more interested in the differences between easily confused words and address a primary interest in conceptual distinction. In such cases, it is possible to re-arrange contextual details according to distinct features first instead of semantic similarities. Alternatively, users might want to know about prototypical contexts and more infrequent occurrences of use and list all contexts according to frequency. Clicking the frequency button will display the most frequent use of each paronym first, and allows the predominance of features between the paronyms in question to be compared (cf. Figure 3). This means that different information can be obtained through the choice of different linguistic parameters. Generally, these needs to be an adaptation to dictionary consultation processes and a dynamic re-organization of information on demand.

### 2.3 Dynamic Consultation of Two Detailed Views

As a contrastive dictionary, it is also of importance to be able to compare the detailed view between two or three contexts dynamically. For this purpose, a menu feature has been implemented for user-specific selection of contexts of interest.

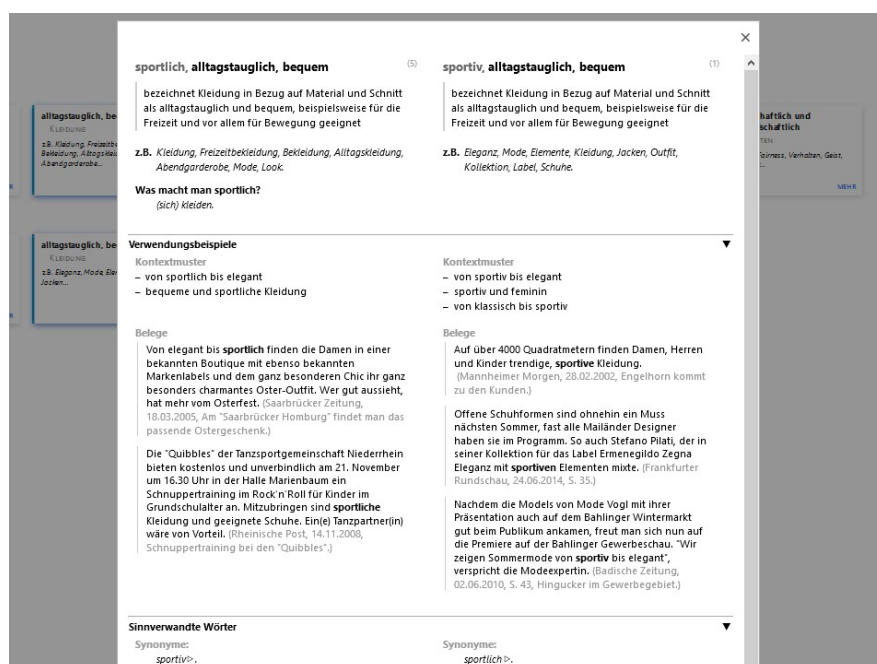


Figure 4: Two dynamically selected contexts in contrast

These contrastive views cover lexical information ranging from definitions at the top to synonyms at the bottom. Therefore, they can be studied comparatively one after another, enabling users to set both terms and their usages in direct comparison with one another. It is the user who selects the individual contexts he/she wishes to compare. Relevant topic areas, conceptual categories and prototypical domains can be consulted in parallel without switching between information or even entries. As a consequence, it is possible to answer questions such as *Can sportlich be used for people who drive a car in a high-risk way?*, *Is friendly behavior and fair play between people in competitions denoted by sportlich or sportiv?* and *Can both terms synonymously refer to people's physical appearance?* The answers to these questions can be derived successfully simply by consulting and comparing contrastive explanations.

## 2.4 Dynamic Consultation of Three Detailed Views

The contrastive detailed display is not restricted to the comparison of two contextual uses: it is also possible to individually select three contexts (see for example the triplet *regulatorisch/regulativ/regulierend* in Figure 5). Again, semantic information concerning the definition, referential domain, co-occurrences, constructional patterns and examples which are part of the detailed view are listed together in a tabular form to allow for direct comparison. Slight differences can thus be detected. These might relate, for example, to the typical domains in which the adjectives occur. *Regulatorisch* is often attested in contexts concerning the FINANCIAL ECONOMY and POLITICS, *regulativ* is frequently used in POLITICS and PHILOSOPHY, and *regulierend* typically occurs in texts reporting on POLITICS and the JUDICIARY. We also learn that all three adjectives can refer to a state of affairs or processes and denote these as controlling or intervening in terms of development and procedures. Only *regulierend* can also refer to institutions.

A contrastive dictionary consultation is, however, not restricted to contexts of different lexical items. It is also possible to individually select two or three contexts for the same word and contrast these in detail. It is then possible, for example, to examine metonymically or metaphorically related contexts of the lexical item in question.

regulatorisch, steuernd	regulativ, steuernd	regulierend, steuernd
bezeichnet einen Sachverhalt oder einen Prozess in Bezug auf Entwicklungen und Abläufe als steuernd oder als eingreifend	bezeichnet einen Sachverhalt oder einen Prozess in Bezug auf Entwicklungen und Abläufe als steuernd oder als eingreifend	bezeichnet einen Sachverhalt, einen Prozess oder eine Institution in Bezug auf Entwicklungen und Abläufe als steuernd bzw. eingreifend
häufig in FINANZWIRTSCHAFT, POLITIK	häufig in POLITIK, PHILOSOPHIE	häufig in POLITIK, JUSTIZ
z.B. Anforderungen, Rahmenbedingungen, Vorgaben, Umfeld, Maßnahmen, Eingriffe, Veränderungen, Auflagen, Vorschriften.	z.B. Idee, Eingriffe, Maßnahme, Politik, Funktion, Rahmen.	Sachverhalte und Prozesse Eingriffe, Maßnahmen, Instanz, Funktion, Rolle, Gesetze, Vorschriften. Institutionen Institutionen, Behörde, Anstalt.
<b>Verwendungsbispiele</b> Kontextmuster – durch regulatorische Eingriffe – {rechtlich / politisch} und regulatorisch – aus regulatorischer Sicht – regulatorisch eingreifen	<b>Verwendungsbispiele</b> Kontextmuster – durch regulative Eingriffe – regulativ eingreifen	
<b>Belege</b> Die WGZ Bank stellt sich auf ein schwieriges Jahr ein. 'Die neuen regulatorischen Anforderungen würden das Ergebnis ebenso belasten wie der in Teilen nach wie vor verzerrte Wettbewerb.' (Rheinische Post, 12.02.2011, WGZ Bank: 2011 wird eine große Herausforderung.) Die deutsche Stromwirtschaft hat sich gestern auf Regeln zur Preisfindung bei der Durchleitung von Strom durch fremde Leitungen verständigt. [...] Die privatwirtschaftliche Lösung zeige, daß es möglich sei, den kommenden Wettbewerb im Strommarkt ohne regulatorische Eingriffe des Staates zu gestalten, hieß es. (Mannheimer Morgen, 04.04.1998, Einigung in der Stromwirtschaft.) Auf mehr Transparenz und ein Frühwarnsystem für künftige Kreditkrisen haben sich die Staats- und Regierungschefs Deutschlands, Großbritanniens, Frankreichs und Italiens bei ihrem Treffen in London geeinigt. 'Es gibt Lücken, die geschlossen werden müssen', sagte Bundeskanzlerin Angela Merkel. Wenn	<b>Belege</b> Dass Gegenstände und Sachverhalte in unterschiedlichen Sinnfeldern erscheinen, trägt dem Umstand Rechnung, dass sie einerseits für uns real sind, andererseits vielfach interpretierbar. Es gibt für Gabriel nicht 'die Welt', sondern nur Ausschnitte. Aber in denen können wir Wahres sehr wohl erkennen. Ausdrücklich wendet sich der Neue Realist auch gegen Jürgen Habermas' regulative Idee einer universellen Vernünftigkeit, die wir angeblich immer schon voraussetzen müssen, wenn wir koordiniert handeln und dabei auch noch moralisch bleiben wollen. (Die Zeit, 10.04.2014, Die Wirklichkeit ist anders) Politische Anreize sind die wesentlichen Treiber für saubere Technologie: Emissionsvorschriften, Einspeisevergütungen und Steueranreize regulieren heute schon den Markt. Stärkere regulative Eingriffe sind auch für Transport und Mobilität zu erwarten. (Stuttgarter Zeitung, 18.05.2007, S. 2, DWS legt ersten Klimafonds auf.)	<b>Belege</b> Staatliche regulierende Eingriffe sind dort sinnvoll, wo sie nicht oder nur schlecht funktionierenden Märkten zur Funktionsfähigkeit verhelfen. (Die Zeit, 08.08.2001, Für den Drei-Generationen-Vertrag, S. 22.) Trotz der leichten Regenfälle im April stehe das Land vor der gleichen Situation wie im vergangenen Jahr, sagte Royal. Er wies darauf hin, daß in einigen Gegenden der Wasserverbrauch durch regulierende Maßnahmen in der Landwirtschaft um rund ein Drittel gedrosselt worden sei. (Nürnberger Nachrichten, 19.05.1992, S. 6, Frankreich will Dürre mit Sparmaßnahmen begegnen.) Wie an Börsen stellen Anleger Risikokapital bereit, das komplett verloren gehen kann. In den USA tobt aktuell eine Debatte darüber, wie regulierende Behörden dieses Risiko transparent machen können. (VDI nachrichten, 24.05.2013, S. 16, Crowd.)

Figure 5: Three dynamically selected contexts in contrast



### 3 Corpus, Methods and Editorial Practice

The empirical examination of paronym pairs/sets necessarily incorporates both a suitably large corpus and contrastive meaning analyses. The corpus is a purpose-built collection of data, covering a range of linguistic material between 1995 and 2015. In order to make all lexicographic description transparent, this corpus can be accessed via the corpus system COSMAS II<sup>9</sup>. All entries are based on findings from the underlying corpus. However, they are by no means computer-retrieved compilations. In fact, computational procedures performing collocation analyses support systematic structuring of linguistic data with respect to patterns and contexts. These patterns are interpreted lexicographically; citations and examples are chosen editorially. Usage-based studies and corpus-linguistic tools constantly assist editorial processes. It is argued that a reliable reference guide cannot do without comprehensive corpus material and corresponding software exploring patterns, as well as lexicographic and linguistic expertise in interpreting the retrieved data.

As well as well-established computational methods, this project also uses a corpus-linguistic method which is capable of measuring semantic similarities (or distance) between pairs of words by contrasting contextual profiles to systematically detect slight differences in terms of collocational behavior. This procedure is referred to as the Contrasting-Near-Synonyms method and was developed by Belica (2001 ff.) (see Figure 6).<sup>10</sup> It is implemented in a corpus-linguistic collocation research and development workbench (CCDB) (see Belica 2001 ff.; Keibel & Belica 2007) and it offers a visual representation of collocates which resemble the search words to varying degrees in terms of common contexts. This method assigns a color to each word. Here, for instance, yellow is assigned to *unsozial* (*unjust*), and red is assigned to *asozial* (*ruthless*, *antisocial*). On the basis of collocation profiles, semantic structures are analyzed, clustered and visualized in a two-dimensional lattice reflecting different degrees of similarity between various words by using a graded colour system. These organizing feature maps (henceforth SOMs) cluster all these items together so that proximity on the grid reflects semantic similarity between semantic profiles. The more their colors differ, the more semantic differences can be found with regard to their uses. The more similar the colors of two neighboring groups, the more similar are their collocation profiles although a strict separation is not suggested, as SOMs imply a continuum of semantic shades. The feature maps break down unstructured patterns and complex semantic properties of the two items in question and set them in relation to each other. They arrange specific aspects of meaning which the items in question share and those they do not have in common.<sup>11</sup> Essentially, this method compares how two words behave by observing all those words that show similar collocation profiles.

In 2009, Vachková and Belica suggested that this approach to collocational patterning might be applicable to the lexicographic investigation of synonyms. They argued

that salient SOM features stimulate lexicographers' associative awareness and encourage guided mental imagery leading to valuable insights into both the word semantic structure and the process of discourse-based negotiation of lexical meaning (Vachková & Belica 2009: 239).

Since 2015, this contrastive method has played an essential role in the examination of pairs of paronyms. The interpretation of such topographic maps has turned out to be a useful device for lexicographers for detecting salient thematic domains or categories of key semantic fields associated with

9 The Paronym-Corpus is freely accessible via the corpus management system COSMAS II: <http://www.ids-mannheim.de/cosmas2/>. Information on the Paronym-Corpus can be found here: <http://www1.ids-mannheim.de/lexik/paronymwoerterbuch/dasparonymkorpus.html>.

10 A number of studies of synonymy have successfully employed this method (e.g. Marková 2012).

11 A similar contrasting method is offered by the feature "Word Sketch differences" within the tool "Sketch Engine" (<https://www.sketchengine.co.uk/>).



© Cyril Belica: Modelling Semantic Proximity - Contrasting Near-Synonyms (version: 0.21, init tau: 0.4, dist: x, iter: 10000)

unsozial	asozial			
kurzsichtig	kontraproduktiv	pervers	anrühlich	feige
unlogisch	unzumutbar	widerwärtig	unhöflich	feig
unklug	untragbar	niederträchtig	respektlos	dumm
bezeichnen		herzlos	ekelhaft	blöd
schlichtweg		zynisch	irrelevant	blöde
empörend		verlogen	gedankenlos	faul
unethisch		unerträglich	landläufig	bescheuert
beschämend		Schande	kindisch	Trottel
unsinnig	unmoralisch	frauenfeindlich	renitent	desinteressiert
widersinnig	verantwortungslos	feindselig	unberechenbar	ungebildet
unvertretbar	menschenverachtend	destruktiv	vorbildlich	uncool
unwirtschaftlich	unvernünftig	korrupt	unpolitisch	abartig
unrealistisch	verwerflich	intolerant	unzuverlässig	unreif
schädlich	unsolidarisch	egoistisch	auffällig	apathisch
ineffizient	unverantwortlich	opportunistisch	egozentrisch	unfähig
vernünftig	unanständig	irrational	unauffällig	unangepasst
unausgegoren	geißeln	kriminell	anarchisch	entfremdet
unfinanzierbar	zentralistisch	faschistoid	undeutsch	zerrüttet
unsolid		gemeingefährlich	dekadent	abgleiten
unsolide		verbrecherisch	animalisch	zerrütten
familienfeindlich		rassistisch	bourgeois	entfremden
wirtschaftsfeindlich		staatsfeindlich	selbsterstörerisch	verwahrlost
Mogelpackung		konterrevolutionär	verpönen	verwahrlosen
beschäftigungspolitisch		antidemokratisch	Rücksichtslosigkeit	drogensüchtig
Bundesregierung	Haushaltssanierung	rigid	diskriminieren	stigmatisieren
rundweg	Gesundheitspolitik	rigide	benachteiligt	stigmatisiert
strikt	Steuerpolitik	solidarisch	Ausgrenzung	Zigeuner
Sozialverband	Sparpolitik	abwürgen	Unterschicht	ausgegrenzt
Bundesfinanzminister	neoliberal	sozial	benachteiligen	Asoziale
Flickwerk	Rentenpolitik	radikal	erniedrigen	rassistisch
Bundesgesundheitsmin	Haushaltskonsolidierung			minderwertig
Abschiebep Praxis	Arbeitsmarktpolitik			diskriminiert
Sparpaket	mittragen	Budgetsanierung		Feigling
Regierungsplan	unpopulär	Gesundheitswesen		Verräter
Sozialreform	Sozialabbau	Leistungskürzung		diffamiert
Steuerplan	Bildungsbereich	Gesundheitssystem		Schmarotzer
Eichel	Sparbeschluß	Wohlfahrtsstaat		abgestempelt
Sparplan	Sparbeschluß	Kündigungsschutz		abstempeln
Gesundheitsreform	rigoros	Grundrente		Untermensch
Reformplan	Sozialbereich	einkommensabhängig		Psychopath

Figure 6: Contrasting *unsozial* and *asozial* with SOMs

paronyms (for an example of this, see Teichmann, forthcoming). The major advantage is that we gain an immediate insight into the thematic topics or contextual domains in which the lexical items predominantly occur. As outlined in Storjohann and Schnörch (2014) and in Teichmann (forthcoming) in more detail, feature maps guide lexicographers to those contextual patterns which will provide further evidence, for example, through the study of collocations that can be attributed to specific thematic domains. Generally speaking, the investigation of paronyms often results in a fruitful methodological combination of SOM-based analysis and collocation analyses.

To have the corpus material, tools and corpus-analytic methods at hand, it was, however, still necessary to develop a lexicographic concept with an underlying linguistic theory, as well as a manageable editorial workload. In particular, the appropriate description of a two/three-lemma entry, together with a dynamic reshuffling of lexicographic information, required an ambitious lexicographic concept. Access to information and navigation structures needed to be straightforward, well thought-through and intuitive. In addition to reliability of information, we wanted to guarantee comfortable usability, both for users in their consultation routines and for lexicographers in their daily practical work. For this, the writing system presents the lexicographers with a number of different tasks within a complex, elaborate XML architecture. The most demanding task is to unite two or three single-lemma descriptions into one homogenous contrastive entry with all its relational elements for comparing and optional (re-)arranging. After all, it is the interaction between editorial conventions,

data requirements and lexicographic expertise that makes the perfect dictionary entry. Compiling and analyzing the data as well as writing the dictionary quickly settled into a daily routine. In the end, the structures, menu options and lexicographers' ideas about the design still held numerous challenges for the hypertext programming, some of which are still unresolved.

## 4 Summary

The new dictionary *Paronyme – Dynamisch im Kontrast* represents a radical change from existing German lexicographic conventions. Users can expect direct access to meaning and use of two or three easily confused words together in concise contrastive overviews or in detailed contrastive descriptions. As an e-dictionary, the new guide can go far beyond the depth of information found in the two existing printed paronym dictionaries. In situations of linguistic doubt, native speakers and learners can learn about thematic domains and semantic environments in which readily confusable words are likely to occur, together with their natural lexical preferences. Synonyms and corpus samples illustrate the information provided, while explicit definitions present details which are important for understanding and encoding. The depth of information is realized as a two-level consultation view, i.e. a short overview and an optional detailed view.

It has implemented dynamic search options, which have replaced rigid structures. With regard to the dynamicity of lexical details, we have shown the options offered by the dictionary in order to flexibly adapt information within an entry. Indeed, "an online dictionary can be adapted to the needs of each dictionary user" (Kwary 2012: 35). Adaptive access and variable search options allow different foci and perspectives on paronymy. Accordingly, the organization of elements changes, and different facets of structural knowledge can be activated. In this way, this reference guide includes alternative access routes to language in authentic usage events involving paronyms. As a digital resource, the aim of the new dictionary is to exploit the possibilities of the electronic medium in order to create a flexible, informative and user-friendly instrument which will enable users to make correct choices. Our dictionary will reflect language-oriented descriptions which show how paronymy works in real communication. We hope to create a reliable source of linguistic and encyclopedic information for situations of language doubt.

## References

- Belica, Cyril (1995). *Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethoden*. Mannheim: Institut für Deutsche Sprache.
- Belica, Cyril (2001ff.). *Kookkurrenzdatenbank CCDB - V3.3. Eine korpuslinguistische Denk- und Experimentierplattform*. Institut für Deutsche Sprache: Mannheim. Accessed at: <http://corpora.ids-mannheim.de/ccdb/> [30/03/2018].
- COSMAS II: *Corpus Search, Management and Analysis System*. Mannheim: IDS. Accessed at: <https://www.ids-mannheim.de/cosmas2/web-app/> [30/03/2018].
- Fillmore, Charles, J. & Atkins, B. T. Sue (1992). Toward a Frame-based Lexicon: The Semantics of RISK and its Neighbors. In Adrienne Lehrer, Eva Feder Kittay (eds.) *Frames, Fields and Contrast. New Essays in Semantic and Lexical Organization*. Hillsdale & London: Erlbaum, pp. 75-102.
- Fuertes-Olivera, Pedro A. (2013). e-lexicography: The Continuing Challenge of Applying New Technology to Dictionary-Making. In Howard Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 323-340.
- Keibel, Holger & Belica, Cyril (2007). CCDB. A Corpus-Linguistic Research and Development Workbench. In *Proceedings of the 4th Corpus Linguistics Conference, CL 2007, 27-30 July 2007*. University of Birmingham, UK.

- Kövecses, Zoltán & Csábi, Szilvia (2014). Lexicography and cognitive linguistics. In *Revista Española de Lingüística Aplicada* 27(1), pp.118-139.
- Kwary, Deny Arnos (2012). Adaptive hypermedia and user-oriented data for online dictionaries: A case study on an English dictionary of finance for Indonesian students. In *International Journal of Lexicography* 25(1), pp. 30-49.
- Marková, Věra (2012). *Synonyme unter dem Mikroskop. Eine korpuslinguistische Studie*. Tübingen: Narr.
- Müller, Wolfgang (1973). *Leicht verwechselbare Wörter*. Duden Taschenwörterbücher Bd. 17. Mannheim: Bibliographisches Institut.
- Müller-Spitzer, Carolin (ed.) (2014). *Using Online Dictionaries*. Berlin/Boston: de Gruyter.
- Murphy, Lynne (2013). *What we talk about when we talk about synonyms (and what it can tell us about thesauruses)*. In *International Journal of Lexicography* 26(3), pp. 279-304.
- Rundell, Michael (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In Ruth Vatvedt Fjeld, Julie Matilde Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress in Oslo/Norway, 7.-11.August 2012*, pp. 47-92.
- OWID<sup>plus</sup>: Accessed at: <http://www.owid.de/plus/> [30/03/2018].
- Paronyme – Dynamisch im Kontrast: coming soon at <http://www.owid.de/plus/> [14/05/2018].
- Pollmann, Christoph & Wolk, Ulrike (2010). *Wörterbuch der verwechselten Wörter. 1000 Zweifelsfälle verständlich erklärt*. Stuttgart: Pons.
- Sketch Engine: Accessed at: <https://www.sketchengine.co.uk/>. [30/03/2018].
- Schnörch, Ulrich (2015). Wie viele Paronympaare gibt es eigentlich? Das Zusammenspiel aus korpuslinguistischen und redaktionellen Verfahren zur Ermittlung einer Paronymstichwortliste. In *Sprachreport* 2015(4), pp.16-26.
- Storjohann, Petra (2017a). Cognitive features in a corpus-based dictionary of commonly confused words. In Iztok Kosem, Caroline Tiberius, Milos Jakubíček, Jelena Kallas, Siomon Krek, Vít, Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of the 5th eLex 2017 conference in Leiden* (19.09.-21.09.2017). Brno: Lexical Computing CZ s.r.o, pp. 138-154.
- Storjohann, Petra (2017b). Cognitive descriptions in a corpus-based dictionary of German paronyms. In Anatol Stefanowitsch/Stefan Hartmann (eds.) *Yearbook of the German Cognitive Linguistics Association* 5, Boston: de Gruyter, pp. 107-118.
- Storjohann, Petra (2015). Was ist der Unterschied zwischen sensitiv und sensibel? In *Zeitschrift für Angewandte Linguistik* 62(1), pp. 99-122.
- Storjohann, Petra & Schnörch, Ulrich (2014). Empirical approaches to German Paronyms. In Andrea Abel, Chiara Vettori, Natascia Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15.- 19. July 2014, Bolzano/Bozen, pp. 463-476.
- Storrer, Angelika (2013). Representing dictionaries in hypertextual form. In Rufus. H. Gouws, Wolfgang Schweickard, Herbert Ernst Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography*. Suppl. Vol.: Recent Developments with Focus on Electronic and Computational Lexicography. Boston/Berlin: de Gruyter, pp. 1244-1253.
- Tarp, Seven (2009). Reflections on Lexicographical User Research. In *Lexikos* 19, pp. 275-296.
- Teichmann, Mareike (forthcoming). SOM und CNS als korpuslinguistische Methoden zur Analyse von Paronymen am Beispiel technisch/technologisch. In Petra Storjohann (ed.) (forthcoming): *Paronymie im deutschen Sprachgebrauch*. Sonderheft Deutsche Sprache 1/2019, Berlin: Erich Schmidt.
- Vachková, Marie & Belica, Cyril (2009). Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography. In *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, 13/2, pp. 223-260.



# Slovenian Lexicographers at Work<sup>1</sup>

*Alenka Vrbinc<sup>1</sup>, Donna M. T. Cr. Farina<sup>2</sup>, Marjeta Vrbinc<sup>1</sup>*

<sup>1</sup>University of Ljubljana, <sup>2</sup>New Jersey City University

E-mail: [alenka.vrbinc@ef.uni-lj.si](mailto:alenka.vrbinc@ef.uni-lj.si), [dfarina@njcu.edu](mailto:dfarina@njcu.edu), [marjeta.vrbinc@ff.uni-lj.si](mailto:marjeta.vrbinc@ff.uni-lj.si)

## Abstract

This paper reports part 1 of findings from a grant project between Slovenia and the U.S. that set out to understand the context and content of modern Slovenian lexicography. Interviews were conducted with six Slovenian lexicographers and one terminographer working on different projects within several institutions (the Slovenian Academy of Sciences and Arts; the University of Ljubljana; and Trojina, Institute for Applied Slovenian Studies). The grant's aim was to discern the philosophical underpinnings, most noteworthy accomplishments, and main projects of Slovenian dictionary work. The focus was on those aspects of lexicographic work that have the greatest significance for the general educated public rather than areas (such as dialectology, etymology, etc.) that might attract primarily language specialists.

The interview script consisted of thirteen narrative questions, designed to allow interviewees to reflect on their daily practice and their underlying vision of what lexicography or terminography is. This paper focuses on a single interview question that captured the interviewees' views on drudgery in lexicography, and on the social/ethical role of the lexicographer.

**Keywords:** harmless drudge, drudgery, interview, lexicographer, objectivity

## 1 Introduction

While dictionaries are created and delivered in similar ways internationally, to the best of our knowledge there have been no in-depth studies of any country's or culture's lexicographic philosophy or practices based upon an analysis of interviews with lexicographers. The present work thus aims to provide a glimpse of the world views of Slovenian lexicographers; it reports the first set of findings from a grant project between Slovenia and the U.S.

Although the Slovenian language has a relatively small number of speakers, there is a significant Slovenian lexicographic tradition; this history, like that of many other traditions (cf. Béjoint 2016; Farina and Durman 2009; Fontenelle 2016) began with needs arising from contact between languages and cultures. Since the 1970s but particularly in the new century, there has been a stream of scholarly work putting forward the underlying philosophy of what general Slovenian lexicography should be (Gantar 2015; Gliha Komac et al. 2015; Gorjanc et al. 2015; Gorjanc et al. 2017; Ledinek et al. 2015; Snoj 2004; Srebnik 2015; Žagar Karer 2011). We build upon this work here via interviews with prominent practitioners.

Around the globe, there are some lexicographers who are familiar with others' work through conference attendance and scholarly publications. On the other hand, many dictionary makers labor alone without a deep awareness of what others in the field are doing, even when similar dictionaries are being created in other countries. Working on a dictionary is by its nature solitary: Despite the availability of 21st-century technological tools, to some extent not so much has changed since 1755, when

1 An expanded version of this article is forthcoming in *Lexikos* 28 (2018) under the title: "Objectivity, Prescription, Harmlessness, and Drudgery: Reflections of Lexicographers in Slovenia".



Samuel Johnson defined the word *lexicographer* as a “harmless drudge”. The present paper strives to broaden the knowledge base of world lexicography by discussing views obtained through interviews with seven distinguished Slovenian lexicographers. A later article will concentrate on Slovenian lexicographic practice as addressed by our seven interviewees.

## 2 Aims of the Study

Through intensive interviews with lexicographic and terminographic specialists, this study set out to address the following research questions:

- A. What is the philosophical and intellectual framework governing the work of Slovenian lexicographers?
- B. What are the main areas of concern and common significant problems that inform the work of Slovenian lexicographers?
- C. What do the lexicographers consider both the main strengths and weaknesses of their current efforts in dictionary creation? What would they most like to change about their practice?
- D. What are the differences among our interviewees in their conception of what lexicography is all about?

The present article addresses mostly A, with some elements of D: What do the lexicographers think about before they even sit down to work; what are their reflections on the most important underlying ideas that drive how they perform their duties. (Later publications are planned to address B and C.)

## 3 The Selection of Interview Subjects

In order to select whom to invite for interviews, we first considered how lexicographic work is organized in Slovenia. Within the Research Center of the Slovenian Academy of Sciences and Arts there is the Fran Ramovš Institute of the Slovenian Language, which specializes in the following areas: lexicology, etymology, onomastics, dialectology, terminology, and historical dictionaries. In addition to work within the Academy of Sciences, there are ongoing lexicographic projects in a variety of units at the University of Ljubljana and the University of Maribor, as well projects led by Trojina, Institute for Applied Slovenian Studies, usually in cooperation with other units.

Since we were concentrating on aspects of lexicography that affect the general educated public rather than linguists/scholars, we sought interviewees who work on synchronic topics, and who concentrate more on the standard language and terminology (rather than on areas such as dialectology, etymology, etc.). Only seven persons could be interviewed due to constraints of time. Therefore, this study should be considered a sampling of views prevailing within the modern Slovenian lexicographic tradition.

## 4 Our interview subjects

Our seven interviewees were not anonymous participants. Due to their positions and influence in the field, their reflections are quoted and cited here so that these ideas might advance lexicography worldwide. The interviewees had the option at all times to provide information “off the record”. What follows is a brief introduction to the interviewees and their areas of expertise:

Apolonija Gantar is a researcher at the University of Ljubljana. She works on collocations, a new grammar of Slovenian, and non-standard Internet Slovenian.

Nataša Jakop, of the Fran Ramovš Institute, is in charge of phraseology for the third edition of *The Dictionary of Standard Slovenian*.

Iztok Kosem, affiliated with Trojina, the Institute for Applied Slovenian Studies and the University of Ljubljana, has worked on projects including a Hungarian–Slovenian dictionary, collocations, and a new grammar of Slovenian.

Nina Ledinek, the Head of the Lexicological Section of the Fran Ramovš Institute, coordinates work on *The Dictionary of Standard Slovenian* and also worked on the FRAN online dictionary portal.

Jerica Snoj worked on the first edition of *The Dictionary of Standard Slovenian* (*Slovar slovenskega knjižnega jezika* 1970–1991), and today works on the third edition. Past projects include *Slovenian Orthography* (Toporišič et al. 2001) and the *Dictionary of Slovenian Synonyms* (Snoj et al. 2016).

Anita Srebnik teaches Dutch at the University of Ljubljana and authored the *Slovenian–Dutch European Dictionary* (2006) and the *Dutch–Slovenian Dictionary* (2007) intended for Slovenian learners of Dutch.

Mojca Žagar Karer is the Head of the Terminological Section of the Fran Ramovš Institute. Her projects include the *Dictionary of Theater Terms* (Sušec Michieli et al. 2007), the *Dictionary of Automated Control Systems and Robotics* (Karba et al 2014), and an ongoing dictionary of legal terminology.

## 5 The Interview Script

The interview script (see Appendix) consisted of thirteen narrative questions, designed to allow the interviewees to reflect on their daily practice as well as their underlying vision of what lexicography or terminography is.

The first two interview questions as well as Script Questions 7 through 9 provided us with personal background information as well as information about the lexicographers' daily work, projects, and accomplishments; Script Questions 4 through 6 treated the philosophical and theoretical underpinnings to their work. Since the study was conducted within the framework of a bilateral project between Slovenia and the U.S. (see Acknowledgements), we asked directly in Script Question 6 about any U.S. sources, theories, or practices that may have influenced the Slovenian lexicographers' work. Script Questions 10 through 12 dealt with the problems and constraints the lexicographers commonly face as they strive to deliver high-quality products to dictionary users. Finally, Script Question 13 asked them to help us by recommending different ways in which international cooperation could take place, and how it might improve lexicographic practice everywhere.

Script Questions 3a and 3b proved to be the most important in advancing our understanding of the interviewees' underlying ideas about lexicography. The interviewee responses to these two questions are our main focus here:

3. The famous English lexicographer, Samuel Johnson, defined the word *lexicographer* as follows, in 1755: “a writer of dictionaries; a harmless drudge, that busies himself in tracing the original, and detailing the signification of words”.
  - a. We would like to know, first: What elements of your own work do you consider “drudgery”, hard, menial, or monotonous work?
  - b. Second, do you think the lexicographer is “harmless”? Does he or she play an invisible, unnoticed social role, or the opposite? How are lexicographers significant to the society of which they are a part?

## 6 Script Question 3a: Lexicography as Drudgery?

Our interview question (3a) on drudgery was intended to encourage interviewees to speak about the more unpleasant or undesirable aspects of their work. We assumed the interviewees would still, in this era of advanced technologies, consider some aspects of lexicographic work to be drudgery, but that they would expand upon both the positive and negative aspects of their work. Among our seven interviewees, we received one “no” and four “yes” responses to the notion that modern lexicographic work has drudgery in it. Two interviewees gave a qualified (“yes, but ...”) answer that focused less on the existence of drudgery and more on ways of mitigating the amount of drudgery in lexicographic work.

The sole terminographer among our interviewees was the only person to answer an unequivocal “no” to the idea of lexicographic drudgery. This is not surprising, given that the work approach of terminography is radically distinct from that of other realms of lexicography. The terminographer’s work, in the words of Mojca Žagar Karer, is much more “dynamic” and is highly interactive. She does not find any of her tasks to be monotonous because she is constantly engaged with experts from different fields. It is the experts who labor over the definitions which have to be precise from the perspective of their field; Dr. Žagar Karer and other terminographers then edit them. Terminographers do not work alone, in “peace and quiet”, but instead are constantly coordinating terminological work or checking fine points in the definitions completed by others. If the terminological work at hand is bilingual or multilingual, Dr. Žagar Karer would most likely need to consult with several different experts to reach a general consensus about the most appropriate way for Slovenian to convey a concept from the terminology of another language.

Among those four who provided an emphatic “yes” to our drudgery question were Nataša Jakop and Jerica Snoj. They said that *all* lexicographic work and *all* phases of dictionary making are drudgery. Nina Ledinek and Anita Srebnik used the word “monotonous” to describe many aspects of such work. Dr. Snoj noted the repetitive nature of the work; each task must be performed thousands of times, for as many words as are being investigated; Dr. Jakop pointed out that monotony can lead to waning concentration, a single moment of which can lead to an error. Dr. Ledinek emphasized how difficult it is to analyze a word with numerous concordance lines in a corpus and multiple meanings. She noted how extremely difficult it is to be consistent, systematic, and coherent when treating grammatical patterns and collocates. Such answers appear to support Ladislav Zgusta’s prediction of more than four decades ago: “The lexicographer has been called a harmless drudge by Dr. Johnson, and he will not advance to a harmless electrician” (1971: 357).

According to Dr. Ledinek, it is also challenging to describe what is the standard language and what is the norm, or to try to describe similar things in a unified way. Finally, Dr. Srebnik, who, of these four interviewees is the only one who compiled her dictionary independently (i.e. not under the auspices of an institute), contributed one not-strictly-lexicographic aspect of her work as additional drudgery: fundraising. Dr. Srebnik stressed that Slovenia needs much better financial support for bilingual lexicographic work.

Our two qualified (“yes, but ...”) answers came from lexicographers who acknowledge that many aspects of lexicographic work are drudgery, but whose remarks focused more on how to lessen the amount of this. Apolonija Gantar works on semantic description and discrimination of senses; she acknowledges that this is challenging but not menial work—what is monotonous is the transfer of such work into a database. She noted that the dictionary is no longer a book; users now expect much more than they did from the print dictionaries of the past. Web-based dictionaries can include lengthy semantic descriptions, grammar, examples, exercises, etymology, phraseology, and other types of information. This is logical: The space limitations of print dictionaries did not allow for all of these

possibilities. Dr. Gantar is interested in the roles that automatization and crowd-sourcing play now and can play in the future in reducing the amount of drudgery in lexicography.

Over the past five years Iztok Kosem has also had as his focus how to get drudgery out of lexicographic work. He works on identifying the menial and routine tasks of lexicography in order to reduce them. He mentioned GDEX, “Good Dictionary Examples” (Sketch Engine | GDEX n.d.), an electronic tool that takes all available corpus examples and ranks their suitability for a specific meaning or sense according to predetermined criteria, thus significantly reducing drudgery and saving time. Dr. Kosem considers that the advent of GDEX is a big step forward in lexicographic work; as corpora have grown to a billion or more words, the problem of too many examples has become ever greater.

While our subjects had diverse views on exactly how much drudgery is involved in dictionary work, there was consensus that they still find their work extremely rewarding. Jerica Snoj commented that in the course of his or her work the lexicographer reaches insights into the language that no one else has, and it is these that help one to endure in such tasks. As Dr. Snoj stated: “It is a gift for all your suffering, but you must be serious in your work to get this satisfaction; otherwise, you can’t reach this stage of insight and there will be only suffering! You must invest a lot to reach this satisfaction.”

## 7 Script Question 3b: Harmless or Harmful?

Question 3b was intended to address the public anonymity of the lexicographer and their potential to do harm; our interviewees gave extensive thought to whether the lexicographer has the potential to be *harmful*, and were very concerned with what for them was the essential nature of their role in society. Their focus on lexicographic ethics was one of the most interesting findings in our study.

Immediately we discovered a variety of opinions concerning the relationship of objectivity in lexicography to harm. In Apolonija Gantar’s previous employment at the Fran Ramovš Institute, she worked in its consulting service for the public. Even when Dr. Gantar was not fully satisfied with an answer she provided, the users believed her due to their perception of her status. While Dr. Gantar considers that “people have to take responsibility for their own language and take part in the [lexicographic] decisions,” most “people don’t want gray areas: they want a straightforward answer” as to whether something is “correct” or “incorrect”.

Interviewee Nina Ledinek noted that people often consult the dictionary to see what is “right”. While the users want a dictionary that guides them, lexicographers cannot move away from objective description. Moving toward prescription risks failing to depict how most people actually talk and write, which would result in dictionaries of no use and with no credibility or authority.

Iztok Kosem saw lexicographers not as harmless but as individuals whose responsibility to the user can be abused. The lexicographer is a mediator between the complexity of language and the final explanation in the dictionary. This mediating role can be quite influential: If a word does not appear in the dictionary, users might believe that it does not exist or might be suspicious of it. They might also be suspicious of the dictionary if it omits a word they like. From Dr. Kosem’s perspective, lexicographers have a duty *not* to be prescriptive. It is the description that really matters, finding the relevant information (evidence) for the users and delivering it quickly.

Nataša Jakop was also an advocate for a descriptive approach. As an individual, the lexicographer is invisible and harmless, but in order to avoid becoming harmful, the lexicographer must consider the linguistic material as objectively as possible. If lexicographers cannot do this and insert their own beliefs or [prescriptive] views, especially without looking at the linguistic material, then they would become harmful.



Apolonija Gantar noted that while there is no single objective interpretation of what a language is, nevertheless the lexicographer must still strive toward objectivity. A well-developed initial plan and conceptualization of the dictionary to be compiled can contribute to the objectivity of the final work. On the other hand, a too-rigid adherence to an initial plan could be harmful, if some specific set of objective data indicates later that you need to do things differently. Dr. Gantar's comment shows that the goals of objectivity and descriptive accuracy, despite the lexicographers' best intentions, can be quite elusive.

While Nina Ledinek considered that lexicographers are not visible, she emphasized that they must be socially responsible and sensitive to the different groups in society. However, Dr. Ledinek maintained that the *Dictionary of Standard Slovenian* does and should have a normative value; she notes that the language has connected Slovenians throughout their history. Dr. Ledinek's comments bring home the descriptive challenge posed by a language like Slovenian, with only about two million speakers; while objectivity is still very much in the focus of Slovenian lexicographers, they also must consider the role of their language very differently than would any lexicographer of English.

Anita Srebnik noted that other languages bring the outside world to Slovenia and allow Slovenians to communicate when they cross any border. Slovenia is small and thus it cannot live without exchange, and an asset of its people is the ability to learn other languages well. Her comments bring to light the important relationship of Slovenian to other languages, as depicted in its bilingual dictionaries. In the case of a (relatively) small language such as Slovenian, bilingual lexicography takes on a special significance: It encompasses not just equivalence in two languages, but also differences of connotation and culture. Dr. Srebnik stated that it is deplorable that the public regards only some dictionaries as conveyers of the norm, the authorities on the language. For the Slovenian media, this authority only accrues to the work of the Academy of Sciences. In Dr. Srebnik's opinion, it is thus the media that causes harm because it limits the focus to a small number of lexicographers and lexicographic projects; in particular, she faults the lack of status and authority for bilingual lexicography.

The terminographer Mojca Žagar Karer had a very different take on the whole notion of objectivity. For Dr. Žagar Karer, it is clear: Lexicography is more subjective and therefore might not be harmless. Because lexicographers write definitions and analyze meaning themselves, they are subjective. Terminographers, in her view, must be objective because they must be credible for the subject field and society. They are trying to create quality language resources which are useful for translators, language editors, and others. It is interesting that Dr. Žagar Karer saw the processes of analyzing meaning and defining as subjective endeavors. While her perception is understandable, we must point out that the field of terminography cannot escape the danger of subjectivity. In terminography, definitions are written by field specialists with the terminographer playing the role of arbitrator and consensus builder. Field specialists, just like general lexicographers, must guard against a lack of objectivity as they try to define words. It is possible that, given their lack of lexicographic experience, some field specialists do inadvertently bring their personal beliefs, perceptions, and prescriptive ideas to definition-writing, what for them is a relatively new endeavor. If two field specialists were to disagree about which of two terms is the best to designate a concept, then certainly we would have two persons striving toward objectivity of description who come up with different results. It is then that the terminographer/editor can ensure a more objective final consensus.

Jerica Snoj stressed that lexicographers are very important for society. Their dictionaries bring the description of language to users, thereby helping them express their thoughts in an appropriate way. When a new dictionary appears, a new insight into the language is opened up. Dr. Snoj considered that a dictionary has a very important role in exploring the possibilities of a language, while Nataša Jakop cited the significant role such works in the preservation of cultural heritage. Dr. Jakop's point is of special significance for the lexicography of any language with a relatively small number of speakers: Preservation is crucial for such languages.



Whether visible or invisible, whether harmless, and whether a drudge, the lexicographer is *the* source of insight into a given language. The responsibility to provide these insights to users in the most ethical way possible is something that all of our interviewees agree on.

## 8 Conclusions

The insights of our interviewees are significant for the development of lexicographic theory broadly construed. Our interviewees accepted some implications of Samuel Johnson's humorous "harmless drudge" designation, while they categorically rejected others. They certainly acknowledged that some aspects of their work can be tedious. While their strong commitment allowed them to accept drudgery as part of the picture, the interviewees were aware that repetitive work can cause the lexicographer's attention to wane and mistakes to be introduced. Because of the potential deleterious effects of monotony, some of the interviewees are actively working toward the development of new technologies to replace the hard, repetitive, and routine lexicographic work that is still done by people.

However, rather than focusing on the tedium or anonymity of their work, our interviewees were more concerned about the lack of public understanding of their job. This lack of understanding can contribute to an overestimation of the lexicographer's authority, which in turn could lead to the disengagement of the public from interest in the Slovenian language. If it is only lexicographers who know the language, then there is nothing left for the educated language user to think about or do except follow the "advice" of the dictionary. Conversely, as the bilingual lexicographer in the group of interviewees pointed out, a lack of public awareness can undermine the valuing of dictionary work by the media or society at large, to the detriment of the production of sorely-needed bilingual and monolingual dictionaries.

The Slovenian interviewees examined in this work would reject outright the idea that the dictionary writer is *a priori* "harmless"; they perceived many possibilities for harm and were thus motivated to avoid it. It is the ethical responsibility of the lexicographer to the dictionary user that is the most important factor in preventing such harm. If a lexicographer were to ignore or misrepresent language facts as found in a corpus or other lexicographic source and veer away from linguistic description, this imposition of personal bias would be socially harmful.

The serious discussion engaged in during this study by the seven Slovenian specialists should not leave the reader with the impression that for them lexicography is a grim and onerous business, quite the contrary. Certainly, as one interviewee put it, lexicography requires a tremendous persistence because, despite constantly improving facilities and research tools, there is still a lot of menial work. Most certainly, media portrayals and the society's general misapprehensions about what lexicography is complicate the already-challenging work of linguistic description. Nevertheless, the six Slovenian lexicographers and one terminographer spoke frequently about "satisfaction": The satisfaction of gaining real insight into the language, the satisfaction of meeting the language needs of users, and the satisfaction of helping them to engage more fully with a language that is such an important part of Slovenian identity. A future article will delve further into Slovenian lexicographic practices as addressed by our seven interviewees.

## References

- Béjoint, Henri. 2016. Dictionaries for General Users: History and Development; Current Issues. Durkin, Philip (Ed.). 2016: 7-24.
- Cowie, Anthony P. (Ed.) 2009. *The Oxford History of English Lexicography*, Vol. I. Oxford: Oxford University Press.

- Durkin, Philip (Ed.). 2016. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.
- Farina, Donna M. T. Cr. and George Durman. 2009. Bilingual Dictionaries of English and Russian in the Eighteenth to the Twentieth Centuries. Cowie, Anthony P. (Ed.). 2009: 105-126.
- Fontenelle, Thierry. 2016. Bilingual Dictionaries: History and Development; Current Issues. Durkin, Philip (Ed.). 2016: 44-61.
- Gantar, Polona. 2015. *Leksikografski opis slovenščine v digitalnem okolju* [Lexicographic description of Slovenian in a digital environment]. (Zbirka Sporazumevanje). 1<sup>st</sup> ed., e-publication. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gliha Komac, Nataša, Nataša Jakop, Janoš Ježovnik, Simona Klemenčič, Domen Krvina, Nina Ledinek, Tanja Mirtič, Andrej Perdih, Špela Petric, Marko Snoj and Andreja Žele. (Eds.). 2015. *Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika* [Preliminary conceptualization of a new explanatory dictionary of standard Slovenian]. Različica 1.1. Ljubljana: Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU.
- Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek. (Eds.). 2015. *Slovar sodobne slovenščine: problemi in rešitve* [Dictionary of modern Slovene: problems and solutions]. (Zbirka Prevodoslovje in uporabno jezikoslovje). 1st ed. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek. (Eds.). 2017. *Dictionary of modern Slovene: problems and solutions*. (Book series Prevodoslovje in uporabno jezikoslovje). 1<sup>st</sup> ed., e-ed. Ljubljana: Ljubljana University Press, Faculty of Arts. [This is the translation of Gorjanc et al., 2015.]
- Karba, Rihard, Gorazd Karer, Juš Kocijan, Tadej Bajd, Mojca Žagar Karer and Tanja Fajfar (Eds.). 2014. *Terminološki slovar avtomatike* [Dictionary of Automated Control Systems and Robotics]. Zbirka Slovarji. Ljubljana: Založba ZRC.
- Ledinek, Nina, Kozma Ahačič and Andrej Perdih. 2015. *Fran: slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, Vodnik* [A Guide to *Fran*: the dictionaries from the Fran Ramovš Institute of the Slovenian Language, in the Research Center of the Slovenian Academy of Sciences and Arts]. (Zbirka Fran). Različica 1.0. Ljubljana: Založba ZRC.
- Sketch Engine | GDEX. n.d. <https://www.sketchengine.co.uk/user-guide/user-manual/concordance-introduction/gdex>.
- Slovar slovenskega knjižnega jezika*. 1970–1991. [Dictionary of Standard Slovenian]. Ljubljana: Državna založba Slovenije.
- Snoj, Jerica. 2004. *Tipologija slovarske večpomenskosti slovenskih samostalnikov* [The lexicographic treatment of polysemous nouns in Slovenian]. (Zbirka Linguistica et philologica). Ljubljana: Založba ZRC, ZRC SAZU.
- Snoj, Jerica, Martin Ahlin, Branka Lazar and Zvonka Praznik. 2016. *Sinonimni slovar slovenskega jezika* [Dictionary of Slovenian Synonyms]. 1<sup>st</sup> ed. Ljubljana: Založba ZRC.
- Srebnik, Anita. 2006. *Slovensko-nizozemski evropski slovar* [Slovenian–Dutch European Dictionary]. (Zbirka Evropski slovarji). Ljubljana: Cankarjeva založba.
- Srebnik, Anita. 2007. *Nizozemsko slovenski slovar = Nederlands Sloveens woordenboek* [Dutch–Slovenian Dictionary]. (Slovarji DZS). 1<sup>st</sup> ed. Ljubljana: DZS.
- Srebnik, Anita. 2015. *Jezikovnotehnološki postopek obračanja dvojezičnih slovarjev* [The technology and linguistics behind the process of reversing bilingual dictionaries]. Praha: Verbum.
- Sušec Michieli, Barbara, Marjeta Humar, Katarina Podbevšek, Slavka Lokar, Edi Majaron, Viktor Molka, Janko Moder, Miran Herzog, Ana Kocjančič, Mojca Žagar Karer and Marjeta Humar (Eds.). 2007. *Gledališki terminološki slovar* [Dictionary of Theater Terms]. Zbirka Slovarji. Ljubljana: Založba ZRC.
- Toporišič, Jože, Franc Jakopin, Janko Moder, Janez Dular, Stane Suhadolnik, Janez Menart, Breda Pogorelec, Kajetan Gantar, Martin Ahlin and Milena Hajnšek Holz (Eds.). 2001. *Slovenski pravopis* [Slovenian Orthography]. Ljubljana: Založba ZRC.
- Žagar Karer, Mojca. 2011. *Terminologija med slovarjem in besedilom: analiza elektrotehniške terminologije* [Terminology from the text to the dictionary: analysis of electro-technical terminology]. (Zbirka Linguistica et philologica, 26). Ljubljana: Založba ZRC, ZRC SAZU.
- Zgusta, Ladislav. 1971. *Manual of Lexicography*. Prague: Academia.

## Acknowledgements

This study could not have taken place without the cooperation of our Slovenian interviewees. Many thanks are thus due to: Apolonija Gantar, Nataša Jakop, Iztok Kosem, Nina Ledinek, Jerica Snoj, Anita Srebnik, and Mojca Žagar Karer. We would also like to thank Marko Snoj, the director of the Fran Ramovš Institute of the Slovenian Academy of Sciences, for welcoming us there. The authors acknowledge the project, *Lexicographic exchange as a way of building bridges between Slovenian and American lexicographic philosophy, governing principles, goals, and work tools*, No. BI-US/16-17-053, which was financially supported by the Slovenian Research Agency. They also acknowledge the approval of the New Jersey City University (NJCU) Institutional Review Board for the Protection of Human Participants in Research. D. Farina thanks NJCU for travel support to Ljubljana, Slovenia and the time released from its Separately-Budgeted Research program. The authors thank NJCU, in particular Tamera Cunningham, Assistant Vice President for Global Initiatives, for providing housing and hospitality during the research visit of A. Vrbinc and M. Vrbinc to the United States.

## Appendix

### Interview script

#### *Beginning of interview*

We want to thank you very kindly for agreeing to work with us on this project. Our working title is: “Slovenian Lexicographers at Work”. Our goal is to add to the worldwide understanding of what lexicographic work is by focusing on work in this country. We consider that the practices in Slovenia should be known and will prove relevant to lexicographers everywhere.

As indicated by the statement you signed, your remarks are not anonymous; we would like to mention you by name and highlight your ideas in any resulting publications. But, on the other hand, if any specific remark you make is not one that you want attributed to you by name, just tell us that it is “off the record”. In that case, we would quote you or cite you generally, using language such as: “Some of our interviewees considered that ....”

#### *Questions*

1. First of all, can you tell us a little bit about yourself? Why were you attracted to the field of lexicography? How did you end up doing what you do today?
2. Can you describe your daily work as a lexicographer? What are the main activities that you do on a daily, weekly, or monthly basis? What aspects of your work do you like best?
3. The famous English lexicographer, Samuel Johnson, defined the word *lexicographer* as follows in 1755: “a writer of dictionaries; a harmless drudge, that busies himself in tracing the original, and detailing the signification of words”.
  - a. We would like to know, first: What elements of your own work do you consider “drudgery”: hard, menial, or monotonous work?
  - b. Second, do you think the lexicographer is “harmless”? Does he or she play an invisible, unnoticed social role, or the opposite? How are lexicographers significant to the society of which they are a part?
4. What is the philosophical and theoretical framework that governs your work? In other words, what is the “umbrella” of ideas under which you do everything that you do?

(Follow-up to Question 4, if needed: What are the “big” ideas that influence how you go about your habitual work as a lexicographer?)

5. Can you explain what are the two or three driving principles that govern your work as a lexicographer? How do you think about these principles as you engage in the minute tasks which lexicographers of necessity must perform?
6. The two previous questions tried to better understand the theoretical and philosophical basis for your lexicographic work. Now we wish to ask: Can you name any theories or practices used in other countries, including the U.S., that inform your own lexicographic work? Or, perhaps when you formulated the principles of your work you incorporated some ideas from abroad?
7. Related to the previous question, have you joined any lexicographic organizations such as the Dictionary Society of North America or EURALEX? Do your memberships of this type affect your work? How?
8. Can you describe two or three of the current projects that you are involved with? We are looking to describe, as completely as possible, what is going on today in Slovenian lexicography. We are also very interested in any future projects that are in the planning stages.
9. In recent years, what are the most noteworthy accomplishments in the work of you and your immediate colleagues?
10. It goes without saying that lexicographic work takes place in the real world and is subject to the usual constraints and challenges of any practical work. In particular, there are always budgetary constraints, but not only budgetary. We would like to know: How is your work challenged by a variety of circumstances; what are the challenges and constraints?
11. Can you name the major strengths of your work situation? What is a best practice for you and your colleagues (e.g., access to different information/sources, user-friendly dictionary-making software, cooperation with IT specialists and/or corpus linguists and/or experts from other fields, etc.)? What affects most positively the compilation of your dictionaries?
12. If you could change one thing about the circumstances of your lexicographic work, what would it be? If you could change one feature of the lexicographic philosophy / theory that underpins your work, what would it be?
13. Could you offer us some suggestions? How do you think the cooperation and exchange of ideas between Slovenian and American lexicographers can be encouraged? Do you consider that more cooperation would improve lexicographic work in Slovenia, the U.S., and beyond?

# Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish

**Piotr Żmigrodzki**

*The Institute of the Polish Language at the Polish Academy of Sciences*

*E-mail: piotr.zmigrodzki@ijp.pan.pl*

## Abstract

Polish Academy of Sciences Great Dictionary of Polish (pol. *Wielki słownik języka polskiego PAN*) is being created by a team of linguists, lexicographers and other specialists in The Institute of the Polish Language PAN. This process started in Kraków in 2006. The dictionary is published exclusively in the online format, free of charge. Before August 2018, 70,000 entries, which include the most frequent words, idiomatic expressions and proper names of the Polish language, will have been finished. The paper provides a general description of the dictionary and focuses especially on following issues of its compilation: a) workflow and methods of organizing work, b) lexicographical system of the dictionary and some of its innovations, which are useful during the process of entry creation and the process of overseeing the entries by lexicographers, c) a new, introduced in 2015, graphical user interface and its features. The final part is devoted to a short presentation of the plans the team has for the future.

**Keywords:** Polish language, electronic lexicography, general dictionary of Polish, online dictionary

## 1 Introduction

Polish Academy of Sciences Great Dictionary of Polish (pol. *Wielki słownik języka polskiego PAN*, WSJP PAN, WSJP) has been created by a team of linguists, lexicographers and other specialists in The Institute of the Polish Language PAN in Kraków (including people from other scientific centers) since 2006. It is being published exclusively in the on-line format, free of charge, under the following URL: <http://wsjp.pl>). By August 2018, the second stage of the work on the dictionary will have come to an end (and at the same time, its second period of financing). After that stage, about 70,000 entries, including the most frequently used Polish words, as well as idiomatic expressions and proper names, will have been completed. The total size of the dictionary (not including inflectional paradigms, which are automatically imported, and which for verbs include several dozen forms for each word) corresponds to about 20 volumes of a printed traditional dictionary. The goal of this article is to concisely present the most important achievements of the lexicographical team, accomplished until this point during their work, as well as to point out the general plans for the future.

## 2 The General Characteristics of the Dictionary

When it comes to typology, WSJP PAN should no doubt be classified as a general dictionary. As mentioned above, it is an electronic dictionary “born digitally” (without a paper version). Apart from that, we could classify it in the following ways:

- A documentary dictionary, i.e. based on an authentic material base, texts in the Polish language. The sources will be discussed in more detail later.



- A dictionary of the contemporary Polish language in the broad meaning of the term – it has been decided, that it will include lexical items (meanings) recorded since 1945. This was substantiated by the fact that the previous great dictionary of the Polish language (Doroszewski, ed. 1958-1969) did include the materials from the post-World War II period, but in a very selective way. At the same time, the choice of sources as well as their description, especially the semantic one, were highly influenced by the socio-political circumstances in the country ruled by communists.
- A descriptive dictionary, not a normative one – the authors do not exclude linguistic realizations considered to be incorrect, or – for any reason – unworthy to be included in the dictionary. They only go as far as to include the information about the normative unacceptability of the listed facts, based on *Wielki słownik poprawnej polszczyzny PWN* (Markowski, ed., 2003) and stylistic qualification of the substandard units.
- An academic dictionary – in the sense that the authors aim to maximize the inclusion of the achievements of the 20<sup>th</sup> century linguistics, especially in the field of semantic, inflectional and syntactic description of the lexical items. At the same time, however, the description in the dictionary is intended to reach as many users of the Polish language as possible (more on this below).

### 3 Source Database, Resource of Entries, and the Scope of Information in the Dictionary

#### 3.1 Source Database

During the planning stage of the dictionary, the National Corpus of Polish (Przepiórkowski et.al. ,2012) was intended to be its main (or even the sole) source database. The creation of this corpus was only slightly more advanced than the work on our dictionary. Therefore other corpora were also used: Polish Scientific Publishers PWN and Institute of Computer Science PAS corpus. An additional minor corpus was also created, solely for the purpose of complementing the dictionary. It included materials which could not make it into National Corpus of Polish and be made public for legal reasons. Unfortunately, the National Corpus of Polish project was concluded in 2011, and since that time the corpus has not been updated, which made it obligatory to search for other sources, especially ones that document the newest language phenomena. As a result, internet resources became more important. These were browsed with the use of the Monco tool (<http://monco.frazeo.pl>), as well as directly through an internet browser. At the same time, words collected in The Institute of the Polish Language PAN were used, both published and unpublished. The editors of the entries can also include materials which they find in printed publications they come across on their own, but this is only done in exceptional cases.

#### 3.2 Resource of Entries

The dictionary entries are not created alphabetically. The initial idea was to describe the most frequently used words in the Polish language. For this reason, a frequency list was taken from Institute of Computer Science PAS corpus. Fifteen thousand of the most frequently used words were chosen from the list to be described. The initial sorting according to thematic fields was conducted, and different groups of words were given to different editors to work on (at the same time, idioms and abbreviations connected to these words also were worked on). This task was finished by the end of 2012, after which the second stage of the work started. This involved describing synonymous or cognate lexemes to the aforementioned ones, as well as some of the newest words, which entered the language after 2000. This will also be the direction of further extensions of the dictionary in the future, because according to the initial plans of the authors it should include "pretty much all words used in the Polish

language”. Among the currently completed entries, the biggest group (about 78%) are those describing single lexemes, which belong to traditionally perceived lexical categories such as nouns, verbs, adjectives, adverbs and numerals. About 18% of the entries are idioms (multi-word lexical units), and few percent are articles describing functional units (prepositions, conjunctions, etc.), proper names (mostly geographical), as well as abbreviations. It was also a norm in general Polish dictionaries of the 20<sup>th</sup> century, to include entries describing morphemes with high derivational productivity, as well as affixes such as *auto-*, *hiper-*. In WSJP we have decided to abandon this approach, because these are not lexical units of language, and only such elements should be included in a dictionary.

### 3.3 The Scope of Lexicographical Data in the Dictionary

A typical entry in WSJP includes the following elements:

- headword form (with variants)
- information about the pronunciation (only for the words with unpredictable pronunciation, especially fresh borrowings)
- chronology
- etymology
- description of meaning
- thematic classification
- hypernyms, synonyms and antonyms of the entry word in the specific meaning
- inflexion (especially the full paradigm of the words inflexion, its affiliation to a part of speech)
- syntactic requirements (especially for verbs)
- collocations and full sentence quotations
- abbreviations (if there are any)
- normative information (pertaining to some incorrect uses of the word)
- notes on usage (any other information pertaining to the usage of the word in texts).

These entries also include idioms, which use the particular word – the idioms themselves are described in separate entry articles with a different structure, though. The scope of lexicographical information in WSJP therefore largely overlaps with the scope of such information in other big Polish and European dictionaries (even though the syntactic requirements of different units or the semantic relations in the Polish lexicographical tradition occur less frequently). In the structure of the dictionary there are also elements of a less typical nature, taken from pedagogical (like the thematic classification of the meanings of each word or idiom) or scientific lexicography (chronology). Thematic classification is done according to the rules and pattern of the meaning division, created by Barbara Batko-Tokarz (2008). It is important not because of the description of a single lexeme, but because it allows the user of the dictionary, thanks to the usage of digital tools, to search for words connected to a given category. Etymological information, originally planned only for fresh borrowings to the Polish language, was extended to all one-word entries after receiving feedback from the readers of the dictionary. The specialists inserting the information are using the findings previously established in etymological dictionaries of the Polish language, but are also very often forced to conduct their own research and to create original etymologies (for details cf. Dębowski et al., 2017). Chronology is typical for scientific dictionaries in Poland. Its goal is to provide the possibly earliest appearance of a given word in Polish texts. In contrast to, for example, the OED, it does not rely on quotes listed in the dictionary. Quotes are listed based on their illustrative usefulness, and also, as we mentioned earlier, WSJP quotes texts created after 1945. The aforementioned dates are therefore established, first and foremost, on the basis of texts which quote the old material, and also other collections of dated texts, including Google Books, which is slowly becoming a good source for establishing a chronology of function words since the beginning of the

19<sup>th</sup> century, up until the first half of 20<sup>th</sup>. Unfortunately, the chronology of words in WSJP does not include the division into meanings. Technical, financial and organizational limitations did not allow for this to happen, especially during the first stage of the project.

## 4 The Lexicographical Team and the Organization of Work

### 4.1 The Lexicographical Team

So far, over 80 people have been involved in the work on WSJP at some point. The amount of their input is of course different, and the team has been modified significantly during the 12 years in which the work on the project has been conducted. The core of the team consists of people employed indefinitely at The Institute of the Polish Language PAS, the lexicographical workshop, located in Kraków. Currently there are 17 such people. Previously, in the first period of the project, separate teams were organized in Warsaw (University of Warsaw), Katowice (University of Silesia) and Toruń (Nicolaus Copernicus University). There were also additional people working on fixed-term contracts (and these were the people whose jobs were affected the most by changes during the project's runtime). This was possible due to the dictionary being edited online, with the use of the Internet. Also, all the source materials are available either as corpora, or as digitalized texts, or are simply widely available in the web. The communication among the members of the project and the teams also took place mostly through the Internet, and only two times during the year were more general meetings of the teams organized. With time, the main work on the dictionary has moved to the team in Kraków, connected to the IJP PAN workshop. The teams in Katowice and Warsaw stopped working as separate parts of the undertaking. The only other group that is left is the team from Toruń, singled out due to the specific material they are tasked to work on (function words, i.e. prepositions, conjunctions, etc.). A group of computer specialists are cooperating with the team. They are responsible for creation and maintenance of the lexicographical system, creation and service of the user interface, processing of the inflectional data, and similar tasks. Finally, the administrative personnel of The Institute of the Polish Language PAS are responsible for financial backing of the project.

### 4.2 Organization of the Work

The work on WSJP is the very definition of a team effort. The rules governing the creation of the dictionary were created during long discussions held in the Workshop of the Great Polish Language Dictionary in The Institute of the Polish Language PAS, and every member of the team has also had an influence on the current state of these rules. It would however be hard to single out and credit any particular part to any particular team member. Due to the way the entries are created, there is never a single person responsible for any one of them. The hierarchy of the authors of the dictionary and the way of creating entry articles have already been described in more detail, e.g. in Żmigrodzki (2011, 2014), cf. also Fig.1. At this point, it should be remembered that apart from the editor, who prepares the initial version of the entry by inserting the data into the form of the dictionary database, the so-called specialists are also active. These are filling just one field in a given group of entries. This pertains to etymology, thematic classification and chronology. Each entry is also controlled by other editors, called supereditors, and before publication the entries are also read by a scientific editor of the dictionary, or another person authorized by one. Inflectional paradigms, prepared by the authors of *Grammatical Dictionary of Polish* (Saloni et al., 2007), are added to the entries. Select people can also fill other information in the entry article. Because of this, the final version of the entry can be a result of a combined effort of more than ten people. WSJP entries are therefore not credited to a specific person, which actually is not that peculiar, considering how such work is done in many dictionaries published around the world.

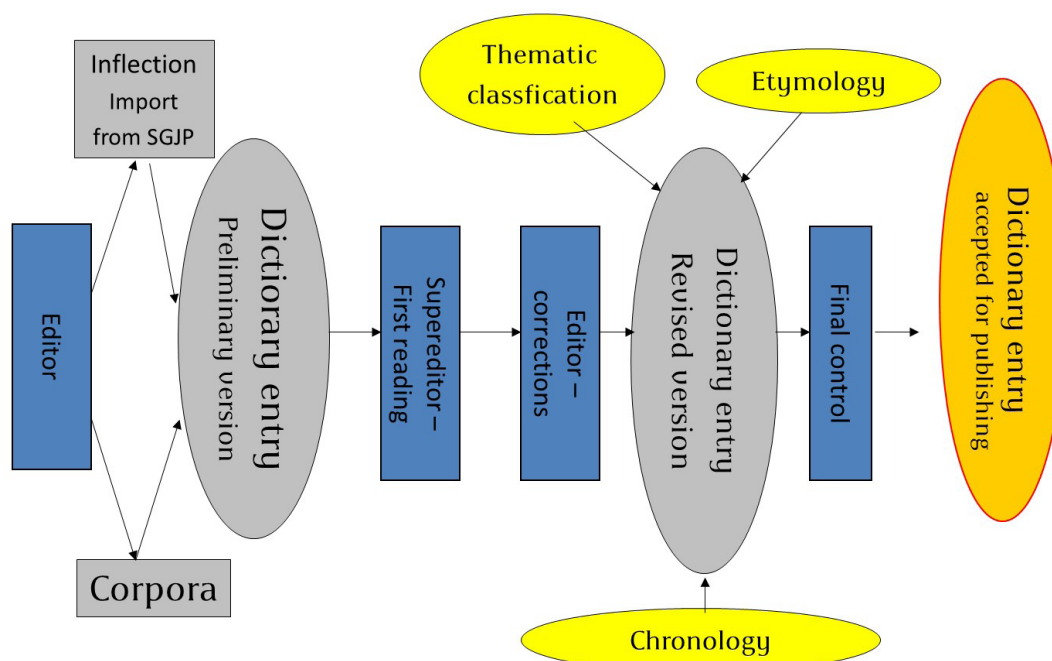


Figure 1: Lexicographical workflow in the WSJP team

## 5 Select Theoretical Aspects and Methodological Studies of the Dictionary

Since the 1980's, we have had increasing criticism of dictionaries in the Polish linguistic community. It has been claimed that their authors are not trying to include the newest achievements in the fields of semantics and syntax, which were heavily researched by the Polish scholars of the time. Because of that, the creators of WSJP decided to make it their goal to include these as much as possible. The fact is that the dictionary is based on the achievements of Polish linguistics of 1980's and 1990's, mostly in the structural paradigm. It develops the theoretical ideas contained in the academic *Grammar of the Contemporary Polish Language* (Grzegorzczkowska et al., ed. 1984) and the works of such Polish scientists as Maciej Grochowski (esp. Grochowski 1993) and Andrzej Bogusławski (1997) (semantics), Mirosław Bańko and Zygmunt Saloni (inflection and syntax), Wiesław Boryś (2005), Andrzej Bańkowski (2001) (etymology). The idea of the WSJP is to create a description of language phenomena based on the modern methodological foundations, which would at the same time be accessible to a wider audience. For this reason, theoretical achievements in linguistics cannot have been introduced directly, but have had to be simplified when it comes to the metalanguage and formal descriptive conventions.

## 6 The IT Aspects of the Work on the Dictionary

As mentioned above, the dictionary is being created and published exclusively in the digital format. Such was the aim of the team since the very beginning, even though in 2005, and even a couple years later, there were some doubts and suggestions to publish a paper version at the same time.<sup>1</sup>

<sup>1</sup> In 2013 when the authors of the dictionary attempted to get a patronage of the Senate of Poland, doubts were also voiced by some senators, who were asking what would become of the dictionary in an event of a lack of electricity or a general downtime of the whole Internet.

However, the structure of the lexicographical description in WSJP was planned in such a way that converting it into a paper version would be impossible, even if financing was to be found. In this article we will not be able to present an in-depth outlook at the technical side of the dictionary, because most members of the team are lexicographers who possess only the basic IT skills. The architecture of the system was created by programmers cooperating with The Institute of Polish Language PAS during the first decade of the 21<sup>st</sup> century, especially Mateusz Żółtak (currently University of Natural Resources and Life Sciences, Vienna)<sup>2</sup> Tomasz Żółtak and Paweł Fronczak. It is systematically modified and upgraded, currently with the last one as an overseer. The system is based on the idea of a database and PostgreSQL system. It involves three basic components: 1) the database of the dictionary, 2) the editing panel, i.e. internet form, used for introducing data into the database, and 3) the presentation panel, which is the representation of the lexicographical data available to the end-user as a website.

### 6.1 WSJP Editing Panel

The editing panel is the basic element of the IT system, from the point of view of its creators. It allows creation of new entries, filling them with data, introducing corrections and remarks directed at editors, by proof-readers or the leader of the project. Apart from that, it contains various functionalities that allow effective management of the entries and the whole dictionary. These include:

- A search engine for the editing panel, which includes such criteria as searching by author, by status (degree of completion), by filled or unfilled fields, by type of entry (connected to the type of the object it describes, e.g. single lexemes, idioms, abbreviations etc.) – with the possibility of combining the criteria and sorting the results based on alphabetical order, editors, degree of completion, date of last change. The list of results includes direct links to specific entries, which allows going directly into editing mode.
- A system of managing editors (available to the main editor and select other individuals), which allows the creation of editor accounts, giving access levels, and setting permissions to modify different groups of entries or the extent of this modification.
- A system of reporting about the work of specific editors, which shows how many entries (and which particular ones) a given editor created (or modified to a satisfactory extent) during a particular time period (which can be changed freely) – this serves to monitor the tasks assigned to particular people, and also for the purposes of calculating wages for the members of the project, as well as for statistical purposes (Fig. 2). It is possible to choose (by ticking the appropriate box) one or multiple editors, whose initials are listed in the upper part of the screen, or even all of them, as well as the type of the report (entry creation, control, thematic classification, etymology, chronology).
- A possibility to track changes in each entry, which makes it possible to establish which editor modified a particular entry and to what extent, including the possibility of displaying the old version of the entry.
- An advanced search engine for the editing panel, released in 2018. Thus allows free searching of any string of characters in any field of the database and any entry, searching for entries labeled with a specific tag, entries coming from a specific language, searching for entries with empty fields of a specific type, etc. The aim of implementing this kind of mechanic is to simplify the final corrections of the dictionary, which requires unification of the descriptions between entries, deletion of inconsistencies in the descriptions and other errors, which were not spotted before (even though each entry is controlled twice).
- A tool to import inflectional paradigms and to transform them.

2 He is the author of the IT groundwork of another dictionary being created in IJP PAN, namely, the Electronic Dictionary of 17<sup>th</sup> and 18<sup>th</sup> Century Polish (<https://sxvii.pl>).



Figure 2: Starting screen of the report about the work of the editors page.

- A preview of the current state of the entry in the presentation panel (see below).
- Dictionary autocorrect tools. These allow users to search for errors and (formal) inconsistencies in the entries, as well as to simplify correcting these errors without having to open particular entries. These include:
  - correction of the semantic relations – this tool produces a list of pairs of entries, where in entry A entry B is listed as its synonym or antonym, but in entry B there is no reference back to entry A (as we know, synonymy and antonymy are reciprocal relations), and allows users to correct this without the need to open the entries;
  - correction of the aspectual oppositions – working similarly to the above;
  - searching for entries without imported inflectional paradigms or with imported paradigms, which are not included in the description of a given entry article;
  - searching for entries describing idioms, which are not linked to proper entries about the single lexemes, with the possibility of automatically collecting them;
  - searching for sequences, inserted by mistake as collocations and idioms in the same entry, with the possibility of easily removing them;
  - correction of the spelling, conversion of the “computer”-style inverted commas, which are not normally used in the Polish language, reduction of excess spaces, etc.

## 6.2 The Presentation Panel of the Dictionary and Its Website

The presentation panel is how we refer to the user interface of the dictionary. Its main functions have already been discussed in Żmigrodzki (2015). The changes which have been introduced since then mainly pertain to the graphical interface of the website, designed from scratch by professional computer graphic designers, and to some optimization of the advanced search engine (though its efficiency will require some improvements as the number of entries grows). Currently, there are the following possible ways of searching in the dictionary:

- Alphabetically – after choosing a letter from the ones listed at the bottom of the screen (Fig. 3) the entries will be displayed in the form of a scrollable list.
- Simple search, typing the form in the window situated in the middle of the initial screen (as a result, all the entries containing the form are shown).

- Advanced search (according to criteria such as: part of speech, qualifier, etymology, thematic category); also searching for specific words, forms and clusters of letters in the specific parts of the entry article.

The default layout of the entry is structuralized, with tabs (Fig. 4), in which the entry article is shown in a format similar to a traditional paper catalogue, and particular parts of the entry are shown or hidden by choosing the appropriate tabs. In the case of entries with more meanings, finding the proper one is made easier by a menu. This consists of semantic tags, which guide the user towards the meaning they are interested in. The second option is a consolidated layout (Fig. 5), which displays the whole content of the entry article in linear fashion. This format allows users to easily print or save the article as a PDF file, if need be.

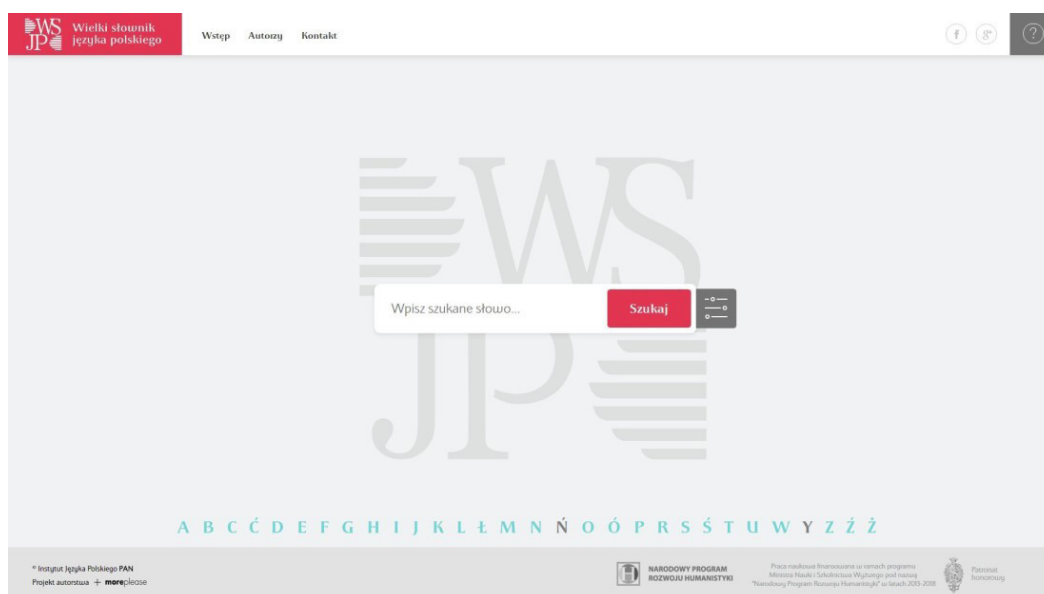


Figure 3: The homepage of the dictionary with the search box.

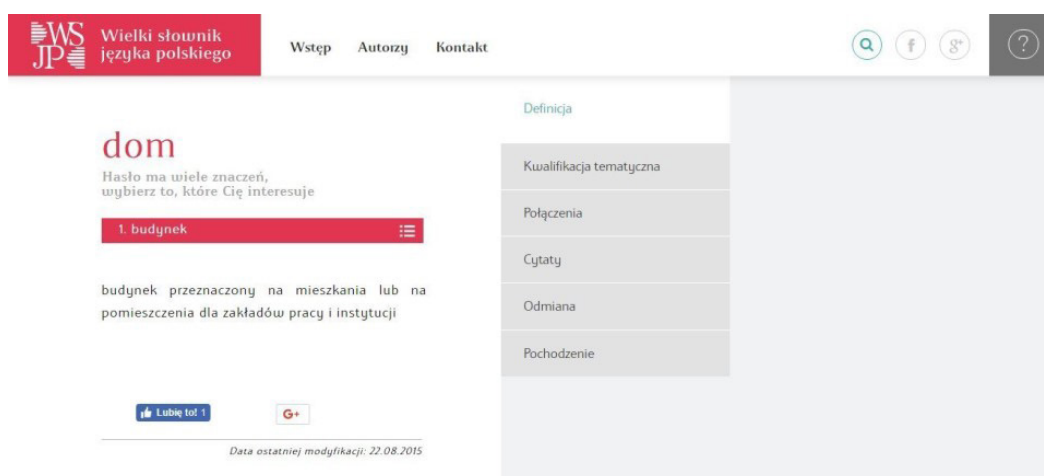


Figure 4: Entry *dom* (house) in the structuralized layout (1<sup>st</sup> subentry 'building').

**dom** DRUKUJ

**Chronologizacja**

SStp  
SPXVI  
SKN  
SJXVII  
STR  
SL  
SWil  
SJPWVar  
SJPDor  
SJPSz  
SJPDun  
ISJP  
PSWP  
USJP

**Odmiana**

część mowy: *rzeczownik*  
rodzaj gramatyczny: *m3*

<i>liczba pojedyncza</i>	<i>liczba mnoga</i>
M: <b>dom</b>	M: <b>domy</b>
D: <b>domu</b>	D: <b>domów</b>
C: <b>domowi</b>	C: <b>domom</b>
B: <b>dom</b>	B: <b>domy</b>
N: <b>domem</b>	N: <b>domami</b>
Ms: <b>domu</b>	Ms: <b>domach</b>
W: <b>domu</b>	W: <b>domy</b>

**Pochodzenie**

psł. \*domъ

**1. budynek**

Figure 5: Entry *dom* (house) in the consolidated layout (fragment).

Apart from the entries, the reader of the dictionary is also able to view the materials, which in classic lexicography would be categorized as *front matters*. So we have a tab called “Introduction”, the contents of which are akin to the introductions of multiple-volume paper dictionaries. It contains a bibliography of the works about the dictionary together with hyperlinks to their electronic versions, if they are available, as well as a subpage with short information about the authors of the dictionary and a contact form for the readers.

## 7 The Internet Life of the Dictionary

The issues of presence and availability of WSJP on the Internet were discussed in a few papers delivered at scientific conferences both in Poland and other European countries, as well as in some published texts, like the most recent Kozioł-Chrzanowska (2017). Not to reiterate the information available in these sources, I will just state that the interest in the dictionary is steadily rising, together with the increase of available entries. The overall number of sessions (according to Google Analytics) in 2017 amounted to 1,066,693, out of which 44.88% came from smartphones<sup>3</sup>, 52.22% from a personal computer or laptop, and 2.91% from tablets (cf. Biesaga 2018). The majority of the connections were tracked to an Internet search engine (about 85%), about 11% to manually inputting the dictionary’s address in the address bar of the browser, and about 3% to links from other sources, which includes social network sites, which constitute about 1% on their own. Apart from the webpage of the dictionary, there is also an official profile on Facebook, as well as Google Plus. It should be noted though that

<sup>3</sup> The dictionary webpage is responsive, meaning it adjusts the display to the characteristics of the device which is used to browse the dictionary.

the makers of the dictionary are currently focused mostly on the lexicographical work, while neglecting promotional activities. So adverts for the dictionary on other sites were not purchased, efforts at positioning the dictionary in search engines were also not made. There is definitely much to be done in this regard in the future. Because the dictionary is of rather a scientific nature, and not social one, we are not aiming at an intensive communication with the readers, even though we do share a contact form with them. The interest in this form of communication is not very great, and thus the form is not used by many people, only a few per month on average. The feedback we get from this is, however, mostly valuable, usually pertaining to errors found in the published entries, or to suggestions about including currently missing entries. From time to time we also get requests for clarification on the proper usage of some words and expressions, but the WSJP team is not providing this kind of advice. In such cases we direct these queries to appropriate people or institutions, which can provide this kind of expertise.

## 8 The Future of the Dictionary

A couple words should be said about the dictionary team's plans for the future. The most important thing is further extension of the number of entries. At the end of 2017, it was assumed that within the next five years (until 2023) it would be possible to have around 150,000 entries, which would already be quite a serious number (the aforementioned dictionary by Doroszewski had about 125,000 entries in 11 printed volumes), while the biggest general dictionary of Polish, the so-called Warsaw dictionary (Karłowicz et al. 1900-1927), has almost 280,000., but with a very poor microstructure and scope of information.

Some technological improvements are also planned:

- Automation of the allocation of entries and opening them. So far allocations were done manually. The editor had to create them in the editing panel, download the inflectional paradigm, and start describing them, after having received the list of entries. It was established that an automatic method of creating the initial versions of entries, including inflectional information, readily available to the user, which the editor could start working on at a convenient time, would be desirable (they would be unavailable to the end-user until published).
- Adjusting the dictionary to the new receiving devices, which might be popularized in the future, which possibly includes a mobile application (that would, however, require the involvement of a specialized IT company).
- Integration with other sources of lexicographical information. The main idea here was to add links to SłowoSieć (the Polish version of WordNet) to the dictionary entries, as well as other dictionaries created and published in The Institute of the Polish Language PAN.
- Extension of the internet analytics and promotion of the dictionary, to make it the basic source of knowledge of the Polish language to as many users as possible, not only to the scientific community.

Unfortunately, the sudden change of the method of financing Polish scientific projects, which was introduced at the beginning of 2018 (and further reforms are planned) created a situation in which the main goal is currently to find funds to sustain the dictionary team, so that the lexicographical work can be continued, even in the most basic format. Because of that, the aforementioned new ideas have had to be put on hold. We can only hope that the determination of the scientific editor and the team working on the dictionary, as well as maybe the voice of its users and the Polish linguistic community, will sway the authorities towards providing a proper financial base for the further development of WSJP PAN.

## References

- Bańkowski, A. (2000): *Etymologiczny słownik języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.
- Batko-Tokarz, B. (2008). Tematyczny podział słownictwa w Wielkim słowniku języka polskiego. In P. Żmigrodzki, R. Przybylska (eds.) *Nowe studia leksykograficzne*, vol. 2, Kraków: Lexis pp. 31–48.
- Biesaga, M. (2018). Wielki słownik języka polskiego PAN w Internecie. In P. Żmigrodzki, M. Bańko, B. Batko-Tokarz, M. Biesaga, J. Bobrowski, A. Czelakowska, M. Grochowski, R. Przybylska, J. Waniakowa, K. Węgrzynek (eds.) *Wielki słownik języka polskiego PAN. Geneza, koncepcja, zasady opracowania*. Kraków: Instytut Języka Polskiego PAN (in print).
- Bogusławski, A. (1988). *Język w słowniku. Desiderata semantyczne do wielkiego słownika polszczyzny*. Wrocław: Zakład Narodowy im. Ossolińskich.
- Boryś, W. (2005). *Słownik etymologiczny języka polskiego*. Kraków: Wydawnictwo Literackie.
- Dębowiak, P., Ostrowski, B., Waniakowa, J. (2017). Etymology in the Polish Academy of Sciences Great Dictionary of Polish. In *Lexikos*, 27, pp. 597–608.
- Doroszewski, W., (ed.) (1958-1969). *Słownik języka polskiego PAN*, vol. 1-11. Warszawa: Państwowe Wydawnictwo Naukowe.
- Grochowski, M. (1993): *Konwencje semantyczne a definiowanie wyrażen językowych*. Warszawa: Zakład Semiotyki i Logicznej Uniwersytetu Warszawskiego, Polskie Towarzystwo Semiotyczne.
- Grzegorzczakowa, R., et. al., ed. (1984). *Gramatyka współczesnego języka polskiego. Morfologia*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Karłowicz, J., Krynski, A., Niedźwiedzki, W. (eds.) (1900-1927). *Słownik języka polskiego*, vol. 1-8. Warszawa: Kasa im. J. Mianowskiego.
- Kozioł-Chrzanowska, E. (2017). What Do Users of General Electronic Monolingual Dictionaries Search for? The Most Popular Entries in the Polish Academy of Sciences Great Dictionary of Polish. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubiček, V. Baisa (eds.) *Electronic lexicography in the 21st century: Lexicography from scratch*, Proceedings of eLex 2017 Conference. Brno: Lexical Computing CZ s.r.o., pp. 202–220.
- Markowski, A. (ed.) (2003). *Wielki słownik poprawnej polszczyzny PWN*. Warszawa: Wydawnictwo Naukowe PWN.
- Przepiórkowski A., Bańko, M., Górski, R., Lewandowska-Tomaszczyk, B. (eds.) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.
- Saloni, Z., Gruszczyński, W., Woliński R., Wołosz R. (2007). Grammatical Dictionary of Polish. Presentation by the Authors. In: *Studies in Polish Linguistics*, 4, pp. 5-26.
- Żmigrodzki, P. (2011). Polish Academy of Sciences Great Dictionary of Polish. History, presence, prospects In *Studies in Polish Linguistics*, 6, pp. 7-26.
- Żmigrodzki, P. (2014). Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. in: *Slovenščina 2.0*, 2 (2), pp. 37–52.

## Acknowledgements

This publication was financed under the program of the Ministry of Science and Higher Education entitled “National Program for the Development of the Humanities” in the years 2013-2018, Project No.: 11H 12 014581.

Publikacja finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Narodowy Program Rozwoju Humanistyki” w latach 2013-2018, nr projektu: 11H 12 014581.





# **Lexicographical Projects and Phraseology**



# Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement

**Michal Boleslav Měchura**

Masaryk University, Brno

E-mail: [michmech@mail.muni.cz](mailto:michmech@mail.muni.cz)

## Abstract

This paper introduces a new way of dealing with phraseology in dictionaries. A classical question in lexicography is whether multiword items such as *third time lucky* should be listed under *third*, *time* or *lucky*. The ideal answer is ‘under all of them’ but, until now, the only way to do that in conventional tree-structured dictionaries has been to keep multiple copies (of what conceptually is one and the same item) in several places throughout the dictionary. We present a way to achieve the same goal without copying. The multiword item becomes a semi-independent subentry which exists in only one copy but appears simultaneously in several places in the dictionary. The structure of the dictionary remains a tree but the lexicographer is empowered to occasionally ‘break out’ of the tree in order to avoid duplication. This paper explains the reasoning behind the concept of shareable subentries, and shows how this new functionality has been implemented in the dictionary writing system Lexonomy.

**Keywords:** subentries, phraseology, Lexonomy

## 1 The Problem of Multiword Item Placement

A perennial problem in lexicography is deciding on the placement of multi-word items (Bogaards 1990): should a phraseological unit such as *third time lucky* be located inside the entry for *third*, *time* or *lucky*? In many such cases the best imaginable answer is ‘under all of them’. But such a suggestion is difficult to accommodate in the classical model of dictionary entries as a tree structure. The only way to include a phraseological unit in more than one entry is to duplicate it, but this is an inelegant solution. Most importantly, it opens up the potential for inconsistency: if a lexicographer makes a change to the subentry *third time lucky* under *third*, there is no automatic way to propagate the change to the other copies under *time* and *lucky*.

A popular method to deal with this in born-digital dictionaries is to treat multi-word phrasemes as independent entries, in effect promoting them to the same level as single-word headwords. This approach ‘solves’ the problem of multiword item placement by deciding not to place them anywhere, and that is also its drawback: it strips the lexicographer of the ability to include a multiword item like *third time lucky* in a specific sense of a single-word entry, for example in a specific sense of *time*. Instead, it delegates the placement question to the search algorithm, hoping that *third time lucky* will indeed appear somewhere on the user’s screen when the user has looked up *time*. This is far from ideal: the job which an item like *third time lucky* does in a dictionary is not just that of a phraseme which users might look up independently. It is (or can be) simultaneously an illustrative example of specific senses of the words it is composed of. This means that the desire to include it in a specific location inside one or more specific entries is lexicographically well-motivated and the ‘treat-multi-words-as-headwords’ method is only a workaround. What is needed is a method for including a single multiword item in several locations inside several entries, but without having to keep multiple copies of them in multiple locations.

## 2 From Trees to Graphs (and Then Back a Little)

The almost total computerization of lexicography in recent decades has not solved the multiword item placement problem. Dictionary entries are usually encoded as XML documents, a formalism which, while making the structure of an entry explicit, offers no innovative departures from the classical tree-structured model.<sup>1</sup> In XML, if one wishes to share an XML fragment between several XML documents, one has to resort to extensions such as XLink which require additional processing and lack implementation in existing dictionary writing systems and other XML tools.

This limitation of the classical tree-based model has inspired some authors to re-imagine dictionaries as graphs rather than trees. In a graph, an element (for example, a phraseological subentry) can be connected to any number of other elements (for example, to several senses of several headwords). Compare this to tree-structured XML, where every element can only be contained inside one other element (for example, a phraseological subentry can only be contained inside one sense of one headword).

Curiously, graph-based dictionaries have not become the norm in (human-oriented) lexicography. While graph-structured datasets are common in machine-oriented lexicons (e.g. various wordnets), human-oriented lexicography has been reluctant to adopt the graph formalism. Existing implementations are rare (e.g. Polguère 2004), and the graph often serves only as an export format for what was originally a tree-structured dictionary: this is the case for most human-oriented lexicons on the Semantic Web (e.g. Aguado-de-Cea et al. 2016, Klimek & Brümmer 2015) where the dictionaries are exported from proprietary tree-structured XML into RDF graphs. Most of mainstream lexicography has remained firmly committed to the tree paradigm. One consequence is that the problem of multi-word item placement remains unsolved.

The disadvantage of graphs (such as RDF graphs on the Semantic Web) is that they are not as easily human-readable as XML trees (and trees in general), not to mention human-writable. Trees can be visualized neatly as two-dimensional objects, while graphs often cannot. Trees are easy for humans to grasp mentally, while graphs are more difficult to ‘take in’. For this reason, it is unlikely that lexicographers will switch to authoring graph-based dictionaries directly any time soon.

The problem then is that, while graphs are the more adequate structure for dictionaries, trees are more ‘lexicographer-friendly’. What we need is a compromise: a set-up which keeps dictionaries in a tree-like structure as much as possible, but which also allows them to ‘break out’ of the tree when necessary: for example to allow the sharing of phraseological subentries between entries (Figure 1). Importantly, we also need a dictionary writing system which allows lexicographers to work with dictionary entries in the familiar tree format as much as possible, while only forcing them to ‘think outside the tree’ when necessary.

In this paper we present how the dictionary writing system Lexonomy<sup>2</sup> (Měchura 2017) offers such a compromise by introducing the concept of **shareable subentries**. A dictionary administrator, while defining the entry schema, can designate certain sections of the XML tree, for example phraseological subentries, to be ‘shareable’, turning them into snippets of XML which are allowed to appear in several entries simultaneously. The structure of the dictionary remains a tree but the mechanism of shareable fragments empowers the lexicographers to ‘break out of the tree’ in order to avoid duplication. This structure is a compromise between trees and graphs, and could perhaps be described as ‘graph-augmented trees’ (Měchura 2016).

1 It would be tempting to assume that the modelling of dictionary entries as tree structures was an innovation introduced into lexicography by the arrival of XML. That is not true. The ‘imagining’ of dictionary entries as theoretical tree structures predates even the invention of XML. A summary of this thinking from pre-computerization times can be found in Wiegand (1989). I am grateful to David Lindemann for this insight.

2 <https://www.lexonomy.eu/>



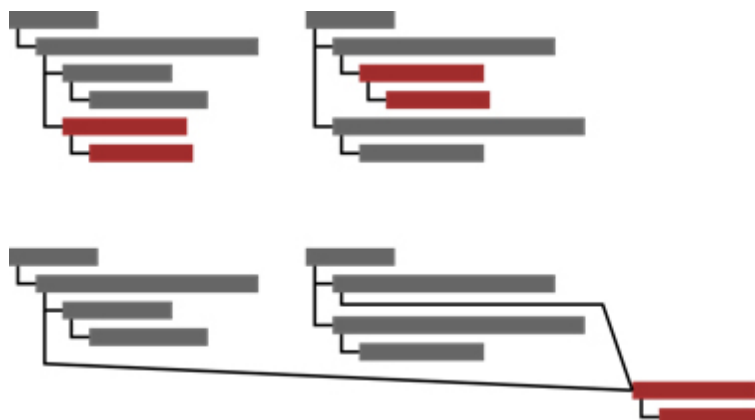


Figure 1: In a classical tree-structured dictionary (above) each element can have only one parent, leading to unnecessary duplication of data across entries. In the graph-augmented tree structure proposed in this paper (below) an element can have multiple parents, allowing reuse of the same element in multiple entries.

### 3 Working with Shareable Subentries in Lexonomy

Setting up a dictionary with subentries in Lexonomy begins just like setting up any dictionary in Lexonomy: you need to create the entry structure first (Figure 2). This is where you decide which XML elements your entries will be made up of, what their names will be, how they will stack up inside each other and so on. In this example we have a very simple bilingual dictionary where each entry has a headword, an optional part-of-speech label and one or more senses. Each sense can have translations and something called *phrasemes*: these will be our multiword items such *third time lucky*.

The next step is to tell Lexonomy that we want phrasemes to be shareable subentries. We do this in the *Subentries* section of the dictionary's configuration screen (Figure 3). All the elements we list here will be treated as subentries by Lexonomy, and we will be able to share them among several entries.

The screenshot shows the Lexonomy web interface for configuring a dictionary. The browser address bar shows 'localhost/mytest/config/xema/'. The interface has a top navigation bar with 'My Test Dictionary', 'Edit', 'Configure', 'Entry structure', 'Download', and 'Upload'. Below this is a toolbar with 'Save', 'Cancel', 'Use your own schema...', and 'Source code'. The main area is divided into two panels. The left panel shows a tree structure of XML elements: <entry>, <headword>, <partOfSpeech>, <sense>, <definition>, <translation>, <phraseme>, <phrasemeText>, and <phrasemeTranslation>. The right panel is for configuring the 'entry' element. It has fields for 'Element' (set to 'entry'), 'Attributes' (with an 'Add...' button), and 'Content'. Under 'Content', there are radio buttons for 'Child elements', 'Text', 'Text with markup', 'Value from list', and 'Empty'. The 'Child elements' section lists three elements: <headword> (min 1, max 1), <partOfSpeech> (min, max 1), and <sense> (min 1, max). Each element has up/down arrows and a delete button.

Figure 2: Setting up the entry structure of a new dictionary.

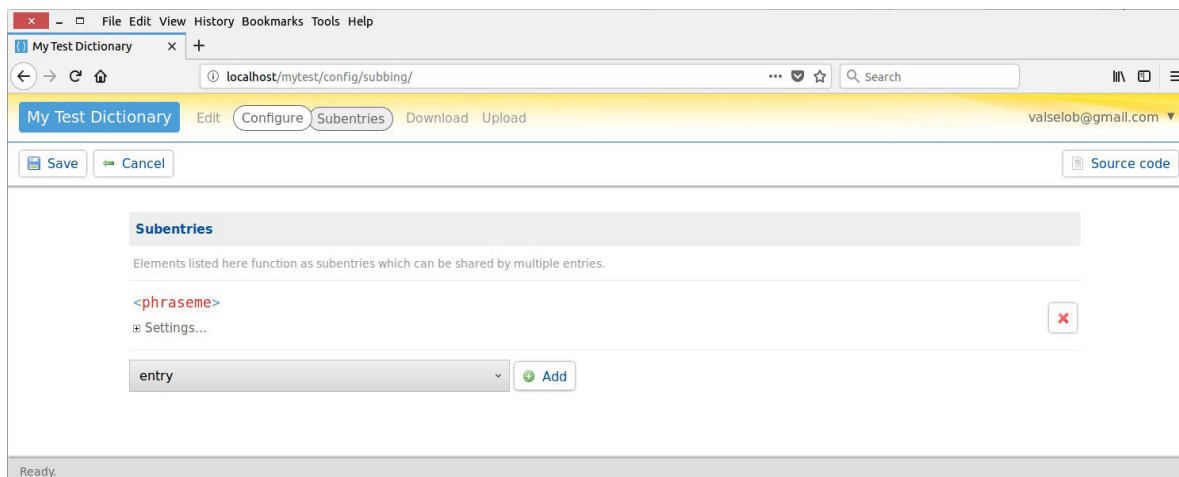


Figure 3: Telling Lexonomy which XML elements should be treated as shareable subentries.

Now that the dictionary has been configured, we can have a look at how lexicographers work with entries and subentries in the editing interface. Figure 4 shows the entry for *lucky* being edited. You will probably agree that this looks almost exactly like entries usually look in Lexonomy, with the XML source visible and open for editing. The only difference is that the `<phraseme>` element has a shaded background: this is Lexonomy's way of telling us that this element (with all its children) is a shareable subentry, and may be shared with other entries. The button near the element tells us how many other entries, besides this one, share this subentry: in this case, two. You can click the button to see which entries those are (Figure 5).

This is the point at which lexicographers need to mentally 'break out' of the tree and realize that any changes made to the content of `<phraseme>` here will automatically be propagated into the other two entries that share this phraseme. In effect, you are editing not just the entry you are looking at now, but the other ones too.

Now let's have a look at how subentries are added into entries. As you probably know, every XML element in Lexonomy's editing interface has a menu which you can open by clicking the element's name. To add a new `<phraseme>` to a `<sense>` click the `<sense>` and a menu will appear. Because Lexonomy knows that `<phraseme>` elements are shareable, one of the options it will offer you is an option to find `<phraseme>` subentries that already exist elsewhere in your dictionary (Figure 6). Clicking it will bring up a window where you can search all the `<phraseme>` elements that already exist anywhere in your dictionary (Figure 7). You can tick the ones you want added to the sense and click the *Insert* button. If you have not found a suitable phraseme that exists already, create a new one by clicking the *New* button: this will insert empty XML markup into your entry and you can fill it in (Figure 8). Once you have saved the entry this newly created `<phraseme>` subentry will join the ranks of other subentries in your dictionary and will be available for sharing with other entries.

As you populate your dictionary with entries you will gradually accumulate a collection of shareable subentries inside the dictionary. This collection of subentries is almost like a separate sub-database in your dictionary and, if you want, you can look at it in isolation. Notice that, in the top left corner of the screen, Lexonomy gives you the option to choose which type of entries you want to work with: the main entries whose top-level element is `<entry>` or the subentries whose top-level element is `<phraseme>` (Figure 9). It goes without saying that any changes you make to a subentry here will be automatically propagated to all the entries that contain it.

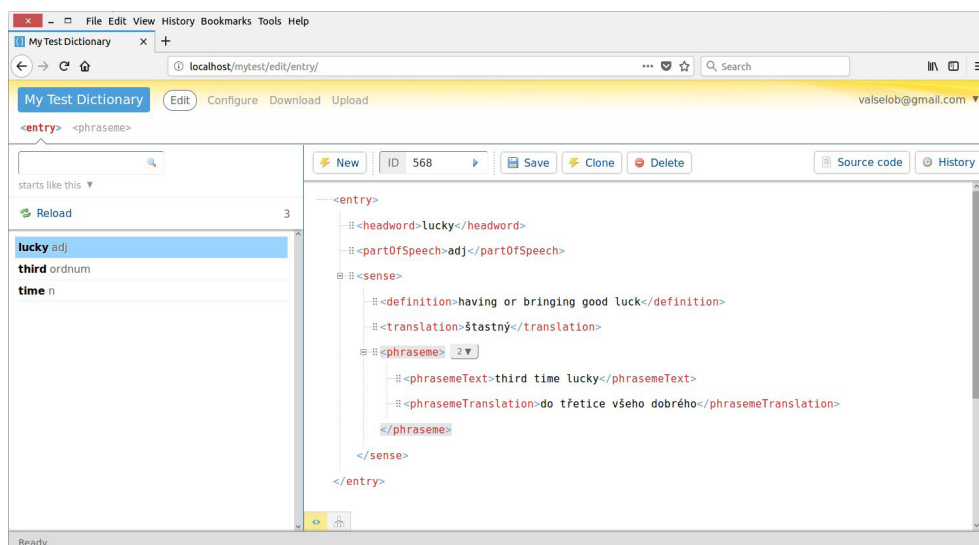


Figure 4: Editing an entry which contains a subentry.

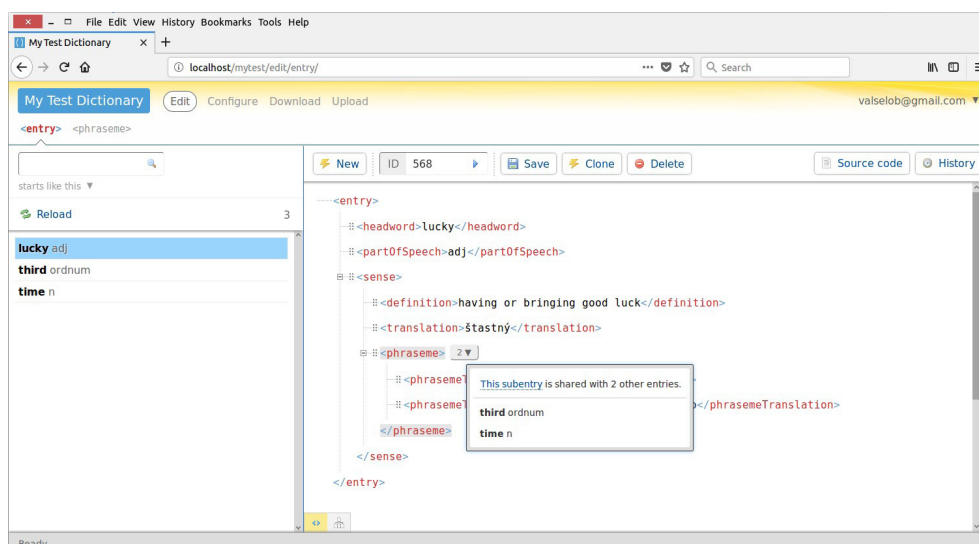


Figure 5: Lexonomy tells us which other entries have the same subentry.

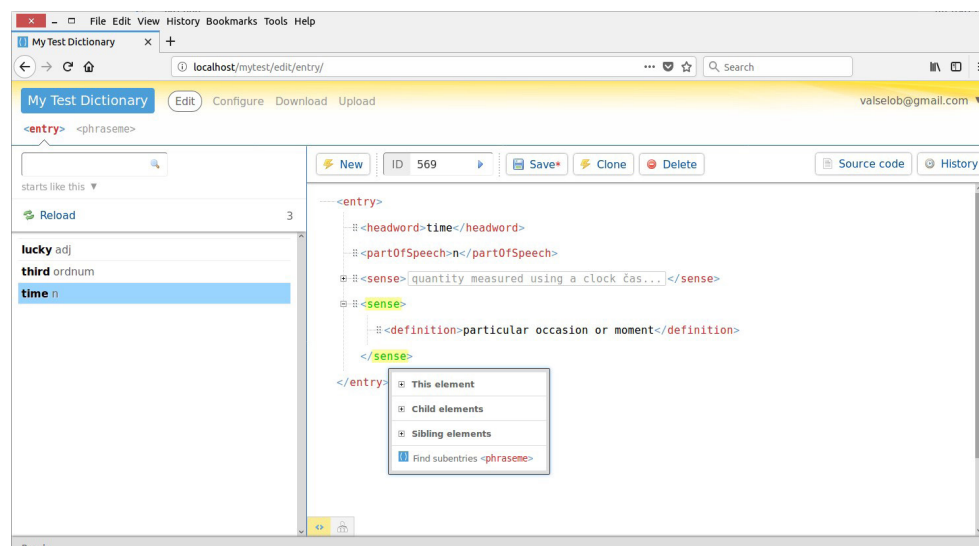


Figure 6: This menu contains an item for finding existing subentries.

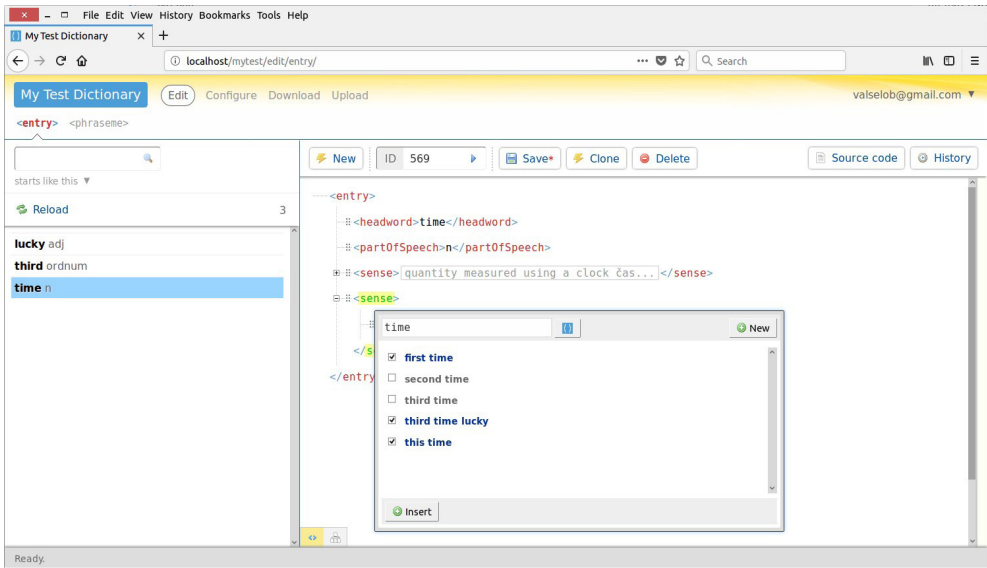


Figure 7: Adding subentries into an entry.

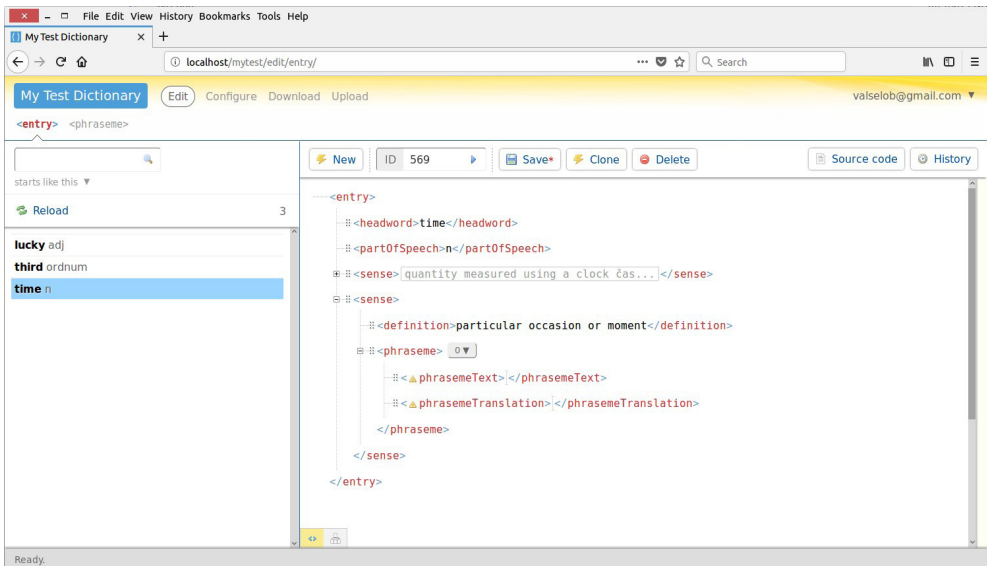


Figure 8: Creating a blank new subentry is the same as inserting a new XML element.

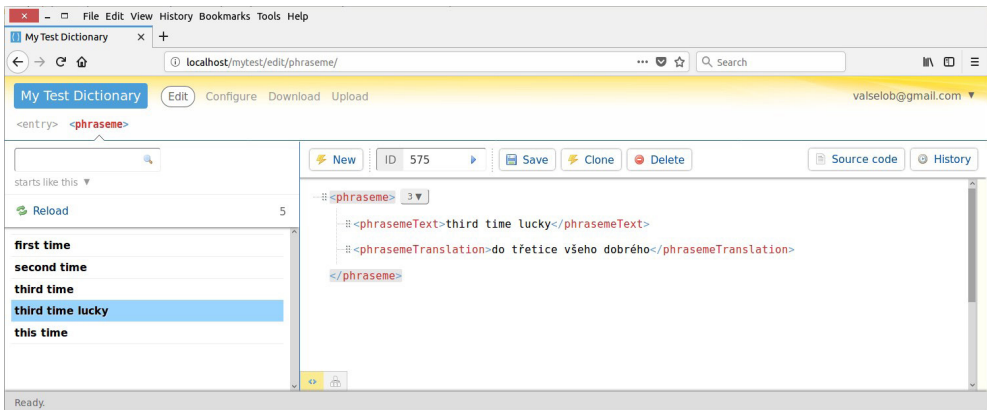


Figure 9: Looking at subentries in isolation.

## 4 Taking the Idea Further: What Can Be A Subentry?

The original motivation for shareable subentries in Lexonomy was multiword phraseological units. But once such a feature exists, the next obvious question to ask is, what other things can it be used for? And it turns out that it can be used for quite a few other things. Everywhere we want to avoid having to duplicate the same information in several places, shareable subentries should be considered.

One obvious candidate is example sentences. A sentence like *That's just great!* would work equally well as examples under *great* and *just*. So this would be an argument in favor of turning the examples into shareable subentries in your dictionary in Lexonomy – especially if the examples carry some other data than just the wording itself, such as pragmatic labels or translations into another language. Then each example is almost like a mini-entry in its own right and might look something like this:

```
<example>
  <exampleText>That's just great!</exampleText>
  <label>sarcasm</label>
  <exampleTranslation>Einfach großartig!</exampleTranslation>
</example>
```

It would be unwise to duplicate this entire fragment in two separate entries (one copy under *great* and one under *just*), because that would be an invitation to inconsistency. If a lexicographer changes one copy (correcting the translation, let's say, or changing the pragmatic label) there is no guarantee that they will remember to change the other copy as well. A wiser approach is to turn the `<example>` element into a shareable subentry. It makes no difference that most such subentries will in fact not be shared (and thus will only occur in one entry): what is important is that the few that will, will always remain synchronized and consistent. Another advantage is that you can treat your example sentences as a separate sub-database inside the dictionary. You can even prepare a large dataset of example sentences beforehand, upload them into Lexonomy (using the *Upload* feature) and then compose all your dictionary entries from them as building blocks.

At this point an inquisitive reader might ask whether it is possible to have subentries inside subentries in Lexonomy. The answer is yes. It is possible (and in fact sensible) to have, let's say, both phrasemes and examples configured as shareable subentries. Some examples will occur at sense level (outside any phraseme) and some will be inside phrasemes. When a phraseme is shared among multiple entries, it will take all its examples along everywhere it goes and they will be shared too.

Another candidate for 'shareability' is translation equivalents in bilingual dictionaries. Let's assume you are working on a bilingual encoding dictionary, that is, a dictionary which is meant to help its users produce texts in a language which is not their mother tongue. In such a dictionary the translations of the headwords are likely to contain a lot of data in addition to the word itself, such as grammatical labels, pronunciation transcriptions and inflected forms. Once again, we find ourselves in a situation where a translation is almost its own mini-entry and could look like this:

```
<translation>
  <translationText>Wohnsitz</translationText>
  <pronunciation>'vo:nzits</pronunciation>
  <partOfSpeech>noun</partOfSpeech>
  <gender>masc</gender>
  <plural>Wohnsitze</plural>
</translation>
```



A translation like the German *Wohnsitz* is likely to be found under many different English headwords: *residence, domicile, abode...* So it makes sense to turn the <translation> element into a shareable subentry and avoid unnecessary duplication. Once a lexicographer has created the <translation> element for *Wohnsitz* in one entry, it can be reused in other entries without having to retype all the information, and without risking inconsistency. And again, it might make sense to treat your translation equivalents as a separate sub-database, prepare a long list of them beforehand, upload them into Lexonomy, and then compose your dictionaries from them as building blocks.

For a final and most extreme example, let's consider the idea of turning headwords themselves into shareable subentries. In a decoding (as opposed to encoding) dictionary, it is the headwords (as opposed to the translations) which carry a lot of grammatical and other annotations. Now, headwords are normally not shared among entries (except when homonyms are given separate entries), but it is possible to imagine unconventional dictionaries where they are: for example, a valency dictionary where each valency pattern is treated as a separate entry. Then we would have, for each headword, many entries which share the headword. In that case it would make sense to turn headwords (along with all their grammatical and other annotations) into shareable subentries.

The conclusion is that the mechanism of shareable subentries is suitable for every situation where we are facing the risk of duplication, which is not limited to multiword phraseological subentries.

## 5 How Subentries are Implemented in Lexonomy

Lexonomy is open-source software where everybody can see and even adapt the source code. For those who might be interested in doing that, a short explanation is in order of how Lexonomy handles subentries internally. Every XML element which is a shareable subentry has an attribute called `lxnm:subentryID` (`lxnm` refers to the namespace `http://www.lexonomy.eu/`) which contains Lexonomy's internal ID of the subentry:

```
<entry xmlns:lxnm="http://www.lexonomy.eu/">
...
<phaseme lxnm:subentryID="54387">
...
</phaseme>
...
</entry>
```

The `lxnm:subentryID` attribute (like all attributes and elements in the `lxnm` namespace) is hidden and never shown to the user in Lexonomy's XML editor, but it is always there. Based on these IDs Lexonomy replaces the element's content with the subentry's content (where by 'content' we mean all of its attributes, text nodes and child elements except the `lxnm:subentryID` attribute itself).

This formalism is functionally equivalent to *Simple Links* with *Replace On Load* behavior, as defined by the XLink standard. The example above could be rewritten in XLink as (hypothetical example, not actually used in Lexonomy):

```
<entry xlink:xlink="http://www.w3.org/1999/xlink">
...
<phaseme xlink:href="54387" xlink:show="replace" xlink:actuate="onLoad">
...
</phaseme>
...
</entry>
```

Additionally, Lexonomy's formalism for subentries is somewhat equivalent to a less well-known feature of Lexical Markup Framework (LMF) where multi-word entries can be independent entries which can then be linked to from specific senses of other entries via their ID.<sup>3</sup> Lexonomy's formalism is more general, however, because it allows the linking of any XML elements.

The process of inserting subentries into entries happens in Lexonomy at the time each entry is saved. For example, when a user saves an entry in which he or she has made a change to a subentry, the changed copy of the subentry is immediately propagated to all other entries where the subentry occurs, so that each entry always has the most recent copy of all its subentries. This means that Lexonomy does in fact keep several copies of the same subentry in several places, but the copies are synchronized with each other behind the scenes. To the user, it seems like there is only one central copy of the subentry in existence. The reason for this apparently unnecessary redundancy is to make sure that the entire contents of each entry is always easily<sup>4</sup> accessible for various searching and indexing operations, for example for XPath queries that 'look inside' subentries.

Because subentries can be recursive (there can be subentries inside subentries), the process of inserting subentries into entries is recursive too. In fact, there is no difference internally between entries and subentries, they are all stored as objects of the same kind and can contain other objects of that kind. Only the dictionary schema decides which of these objects are treated as entries and which as subentries: an entry's top-level element is also the schema's top-level element, typically called `<entry>`, while a subentry's top-level element is something else deeper down in the schema, such as `<phrase>` or `<example>`. This even means that something can simultaneously be an entry and a subentry of other entries, if the dictionary is so configured.

## 6 Conclusion

The problem of multiword item placement has a solution which, in retrospect, seems obvious: to allow subentries to appear in more than one place in the dictionary. The only question is a technical one: how to make this possible.

The classical tree paradigm, as formalized in XML, is not expressive enough for this purpose, because it does not allow elements to have more than one parent. On the other hand, the more generic graph paradigm, as formalized in Semantic Web RDF, for example, is hugely over-expressive and is an overkill for this purpose because it departs too far from the familiar two-dimensional nature of trees. This paper has thus proposed a compromise, a tree-like structure which differs from XML in just one aspect: it allows elements to have more than one parent. This new paradigm – together with the new editing features in Lexonomy – allows lexicographers to continue working with trees most of the time, while occasionally 'breaking out' of the tree to share subentries among two or more entries.

It is hoped that this will be one of the innovations that empower lexicography to depart from a situation where it is merely imitating on computer screens what it was previously doing on paper, and move to a smarter position where it can fully avail itself of the possibilities offered by the digital medium.

<sup>3</sup> I am grateful to Adam Rambousek for this observation.

<sup>4</sup> By 'easily' we mean 'without additional processing'.

## References

- Aguado-de-Cea, G., Montiel-Ponsoda, E., Kernerman, I., Ordan, N. (2016) 'From dictionaries to cross-lingual lexical resources' in: *Kernerman Dictionary News*, 24, pp. 25-31.
- Bogaards, P. (1990) 'Où cherche-t-on dans le dictionnaire ?' in: *International Journal of Lexicography*, 3(2), pp. 79-102.
- Klimek, B., Brümmer, M. (2015) 'Enhancing lexicography with semantic language databases' in: *Kernerman Dictionary News*, 23, pp. 5-10.
- LMF: ISO 24613:2008 Language resource management – Lexical markup framework, [http://www.iso.org/iso/home/store/catalogue\\_tc/catalogue\\_detail.htm?csnumber=37327](http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=37327)
- Měchura, M. B. (2017) 'Introducing Lexonomy: an open-source dictionary writing and publishing system' in: I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, Leiden, pp. 662–679, <http://www.lexonomy.eu/docs/elex2017.pdf>
- Měchura, M. B. (2016) 'Data Structures in Lexicography: from Trees to Graphs' in: Horák, A., Rychlý, P., Rambousek, A. (eds.) *Recent Advances in Slavonic Natural Language Processing*, <http://www.lexiconista.com/raslan2016.pdf>
- Polguère, A. (2004) 'From Writing Dictionaries to Weaving Lexical Networks' in: *International Journal of Lexicography*, 24(7), pp. 396-418.
- Wiegand, H. E. (1989) 'Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven' in: Hausmann, F. J.; Reichmann, O.; Wiegand, H. E.; Zgusta, L. *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, Berlin: de Gruyter, pp. 409-462.
- XLink: XML Linking Language Version 1.1, The World Wide Web Consortium, <https://www.w3.org/TR/xlink11/>

# A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined

**Lut Colman, Carole Tiberius**

*Dutch Language Institute, Leiden*

*E-mail: lut.colman@ivdnt.org, carole.tiberius@ivdnt.org*

## Abstract

*Woordcombinaties* (*Word Combinations*) is to be a new online lexicographic resource in which a Dutch collocation and idiom dictionary will be combined with a pattern dictionary. We believe that the combination of these dictionary types will be of great value to language learners and teachers. In this paper we present the three-year pilot in which we design the project and start with the description of the combinatorics of a selection of verbs for advanced learners of Dutch as a second language. We will merge a pattern dictionary of Dutch verbs, following the example of the *Pattern Dictionary of English Verbs* (PDEV),<sup>1</sup> with a collocation application, following the example of *Sketch Engine for Language Learning* (SkeLL).<sup>2</sup> In a follow-up to this pilot, more verbs and the combinatorics of nouns and adjectives will be dealt with. The long-term purpose of *Woordcombinaties* is a fully-fledged phraseological resource for Dutch.

**Keywords:** word combinations, collocations, idioms, proverbs, conversational routines, patterns, e-dictionary for learners of Dutch as a second language, Corpus Pattern Analysis (CPA)

## 1 Introduction

The importance of phraseology in second language learning and teaching is generally acknowledged (Howarth 1998; Cowie 1981, 2008; Bahns & Eldaw 1993; Wray 2000; Jesen 2006; Granger & Meunier 2008; Peters 2013). Granger and Meunier (2008) rightly propose that phraseological information should be available to learners and teachers and it should be rapidly and easily accessible. Online dictionaries are an obvious means to this end. However, up to now, Dutch dictionaries, in print and online, have many shortcomings as phraseological repositories. On the one hand, they deal with phraseology fragmentedly and rather unsystematically (Fenoulhet 1991; de Kleijn 1999, 2003; Hilgsmann 2005). On the other hand, a fully-fledged phraseological lexicographic resource for Dutch with quick and easy access to the information is still lacking. The focus in general language dictionaries, translation dictionaries and pedagogical dictionaries is mainly on idioms and proverbs, whereas the description of collocations has been given little attention (de Kleijn 1999).

We can illustrate the fragmentation with a few examples. There is a print idiom dictionary *Idioomwoordenboek* (Van Dale 1999) and a print and online collocation dictionary *Combinatiewoordenboek* (de Kleijn 2003).<sup>3</sup> The first one by definition focusses on idioms like *de strijdbijl begraven* (*bury the hatchet*), the latter is a useful collocation dictionary, but idioms are explicitly excluded and it is restricted to collocations of nouns with verbs. The learners' dictionary *Van Dale pocketwoordenboek: Nederlands als tweede taal (NT2)* (Verburg et al. 2017) includes verb valency patterns, like [*ie-mand schetst iets*] (*someone sketches something*) and includes frequently used idioms, but it does not

1 <http://pdev.org.uk>

2 <https://skell.sketchengine.co.uk/>

3 <https://combinatiewoordenboek.nl>

include lists of collocates to fill the slots in the patterns. A systematized pattern description of about 500 verbs is provided in the Dutch – English – French *Contrastive Verb Valency Dictionary (CVVD)*,<sup>4</sup> but this dictionary by definition is restricted to valency and does not include lists of collocates either. The *Algemeen Nederlands Woordenboek (ANW)*,<sup>5</sup> the online general language dictionary of modern Dutch, covers collocations and idioms, but there is no systematized pattern description for verbs yet, and easy and quick access to the phraseological information is still an issue.<sup>6</sup>

As a result of this fragmented phraseological landscape, learners of Dutch are dependent on many different resources to meet various phraseological needs, which is an impractical and undesirable learning environment. A project like *Woordcombinaties* can answer various types of phraseological queries by combining access to collocations, idioms and valency patterns in one tool. This tool can be used by second language learners in computer-assisted language learning (CALL) and data-driven language learning (DDL) for comprehension and production and by teachers as a resource to find data for combinatorics in vocabulary and grammar lessons and sentence-building exercises.

We use the term *woordcombinaties* (*word combinations*) for any meaningful type of combination of words with spaces. This includes free combinations and multiword expressions, like collocations, fixed expressions and idioms, but also more abstract semantically motivated valency patterns. Compounds and phrasal verbs in Dutch are written as one word. The *ANW* encodes these quite comprehensively. As both projects will be linked in the future, multiword expressions of words without spaces are accounted for. We will refer to compounds in *Woordcombinaties* only when they are synonymous with combinations with spaces, for example *slaapwel* (*good night, sweet dreams*) as synonym of the combination *slaap lekker*.

## 2 Background and Related Work

Our own and others' lexicographic experience and usage-based linguistic approaches, like lexico-grammar and construction grammar, made us acknowledge that there is no clear-cut division between lexicon and grammar, and that words get their meaning when used in context, thus in combination with other words (Firth 1957; Fillmore et al. 1988; Halliday & Matthiessen 2014; Sinclair 1991; Goldberg 1995, Gries 2013; Hanks 2013). Hence, in *Woordcombinaties* meanings will be associated with combinations of words rather than with words in isolation.

During the planning stage of the pilot we performed research on several phraseological projects to get a clear picture of the features we wanted to include. Special attention was paid to collocation applications, online valency dictionaries and online pattern dictionaries from a more semantic point of view. An exhaustive account of this research is beyond the scope of this paper, but we will briefly go into the ones that influenced our project most: *Sketch Engine for Language Learning (SkELL)*, the *Pattern Dictionary of English Verbs (PDEV)*, the German valency dictionary *E-VALBU* and *StringNet Navigator*.

### 2.1 Sketch Engine for Language Learning, SkELL<sup>7</sup>

*Sketch Engine for Language Learning (SkELL)* is a fully automated web interface for learners and teachers of English to search for words and phrases in corpora and find example sentences,

4 <http://www.cvvd.ugent.be>

5 <http://anw.inl.nl>

6 In the future, *Woordcombinaties* will be accessible both as a stand-alone dictionary and as a plug-in resource for other applications, for example *ANW*.

7 <https://skell.sketchengine.co.uk>



frequent collocates and words with similar behavior.<sup>8</sup> Example sentences are retrieved using GDEX. GDEX stands for “Good Dictionary EXamples” and is a technology that evaluates sentences with respect to their suitability to serve as good dictionary examples. GDEX is the abbreviation for *good dictionary examples*: short and intelligible, but informative sentences elucidating the definition and exhibiting typical patterns of usage (Kilgariff et al. 2008). Access to multiple examples of usage is useful to language learners. Frankenberg (2012 and 2014), for example, found a beneficial effect on language comprehension and production of data-driven learning through exposure to multiple good examples in experiments with Portuguese learners of English as a second language.

Word sketches in *SkELL* list collocates, which are defined as words which frequently co-occur with the searched word. The collocates are grouped according to syntactic function or another grammatical relation, for example *verbs with x as subject*, *verbs with x as object*, *modifiers of x*, etc., and can be part of a collocation or another type of multiword expression, like an idiom. The word sketch of the verb *bury*, for example, contains among others *dead*, *body* and *hatchet* as direct objects. The first two being part of a collocation, the latter being part of the idiom *bury the hatchet*.

The ‘similar words’ function shows words used in similar contexts. They are listed and visualized with a word cloud.

The main difference between *SkELL* and the *Sketch Engine* as a tool is that the first is intended as a reference and learning tool for language learners, while the latter is a language corpus management and query system for research and lexicography. *SkELL* draws data from a specially built and cleaned corpus and offers only the collocations for a selection of important grammatical relations in language learning, such as subjects, objects and modifiers.

In *Woordcombinaties* we will offer a *SkELL*-like function for Dutch with GDEX examples and word sketches. We expect it to provide a good first and overall impression of the different senses and usage patterns of the searched word and quick access to target collocates for language production. In contrast with the original *SkELL*, we will post-edit the automatically retrieved word sketches to eliminate noise. We will also add more complement types, for example, prepositional objects and clausal collocates.

## 2.2 Pattern Dictionary of English Verbs, PDEV<sup>9</sup>

The *Pattern Dictionary of English Verbs (PDEV)* is a corpus-driven inventory of verb patterns and their implicatures.<sup>10</sup> The implicature is a definition anchored to the arguments in the pattern. Each pattern-implicature pair is illustrated with an example from the British National Corpus (BNC) (Figure 1) and access to more data is provided by links to annotated concordances. Patterns are also linked to FrameNet<sup>11</sup> semantic frames.

*PDEV* is the first dictionary in which meanings are associated with usage patterns of words instead of with words in isolation (Hanks 2008). Patterns include valency structures, but they are more than that. Other syntagmatic features can also determine a pattern and its implicature (Hanks 2004).<sup>12</sup>

8 *SkELL* is also available for other languages, for example, *ruSkELL* for Russian (<http://ruskell.sketchengine.co.uk>) and *csSkELL* for Czech (<https://cskell.sketchengine.co.uk>).

9 <http://pdev.org.uk>

10 Hanks borrowed the term *implicature* from H.P. Grice (1968), “to denote the act of intentionally implying a meaning that can be inferred from an utterance in context, but it is neither explicitly expressed nor logically entailed by the statement itself.” (Hanks 2013: 74).

11 <https://framenet.icsi.berkeley.edu/fndrupal>

12 Hanks (2004) illustrates this with the different meanings of *take place* vs. *take his place*, due to the absence or presence of the determiner.

Patterns are semantically motivated: lexical sets in argument slots are linked to semantic types from a shallow ontology.<sup>13</sup> For example, [[Human]], [[Institution]] and [[Beverage]] are the semantic types that label the lexical sets {he, woman, ...}, {school, firm, ...} and {coffee, tea, beer, ...}. The lexical sets, however, are not made explicit in the patterns. The linking of patterns with their typical lexical sets is a new feature which will be encoded in *Woordcombinaties*.

PDEV uses a specific lexicographic technique to identify the usage patterns which is called Corpus Pattern Analysis (CPA) (Hanks 2004). For each verb a sample of 250 (or more) concordance lines is analyzed and annotated with pattern numbers. The annotated concordances are grouped automatically and the lexicographer associates them with a meaning.

Both the dictionary and CPA technique are developed by Patrick Hanks within the scope of his Theory of Norms and Exploitations (TNE) (Hanks 2008, 2013). This theory distinguishes between normal or prototypical uses of words and exploitations of these, like patterns with anomalous collocates or unconventional metaphors. Corpus pattern analysis will reveal the most normal usage patterns of the verbs, which can top the pattern list in the dictionary, whereas exploitations of prototypical patterns can either be left out or put at the bottom of the list.

In *Woordcombinaties* we will describe verb patterns in a similar way, but we will tailor it to the requirements of our user group, advanced learners of Dutch. The German project *E-VALBU* inspired us for this purpose.

Pattern: **Human 1** or **Institution 1** or **Eventuality** encourages **Human 2** or **Institution 2**  
 Implicature: **Human 1** or **Institution 1** or **Eventuality** has the effect of causing **Human 2** or **Institution 2** to feel more confident or positive  
 Example: Do all that you can to encourage other people in your class who are struggling with certain subjects and activities.

Figure 1: PDEV pattern with semantic types.

### 2.3 E-VALBU<sup>14</sup>

*E-VALBU* is a semantically motivated valency dictionary for German which is developed at the Mannheim Institute for German Language. It is the electronic implementation of *VALBU*, the largest printed dictionary on German verb valency which was completed in 2004. The entry list of 638 verbs in the printed edition meets the requirements for the certificate “German as a Foreign Language” at the federal Goethe-Institut (Schneider 2008) and is integrated in the electronic version.

Verbs in *E-VALBU* are described in a maximum valency frame, which means that all elements essential to explain the meaning of the verb are encoded as complements. The verb *mieten* (rent), for example, needs five complements<sup>15</sup> to distinguish the meaning of the verb from that of *kaufen* (buy) (Kubczak 2014). For each verb-meaning pair *E-VALBU* codes obligatory and optional complements within the sentence structure (called *Satzbauplan*). Optional complements are put between round brackets. In *PDEV* complements in the patterns are embedded as semantic types, whereas in *E-VALBU* they are embedded as dummies, like *jemand*, *etwas*, *irgendwieviel* in *jemand mietet etwas/jemanden für irgendwieviel von jemandem irgendwielange*, so that semantic roles are more or less implicitly recognizable. At mouseover on the dummies, semantic types become visible. In the examples section (*Beispiele*) a few corpus examples are given and in the section *Belegungsregeln* the formal grammatical categories of the complement types are listed. Each dummy in a pattern is assigned a color, which re-occurs in the complement type names in the *Satzbauplan*, the formal categories in

<sup>13</sup> <http://pdev.org.uk/#onto>

<sup>14</sup> <http://hypermedia.ids-mannheim.de/evalbu>

<sup>15</sup> With roles such as renter, tenant, rented object, rent and term of lease.

the *Belegungsregeln* and in the lexical items in the examples. This way, language learners can easily recognize the complements, even if the word order in the example differs from the pattern.

In *Woordcombinaties* we will adopt the dummy practice of *E-VALBU*. A direct embedding of semantic types in patterns may hamper readability, especially when patterns contain many semantic types. We will not use the mouseover, however, because this is impractical in phone apps. Instead, we will show semantic types and the sets of collocates on mouse click. Sentence structures which code obligatory and optional complements will be adopted from *E-VALBU* as well.

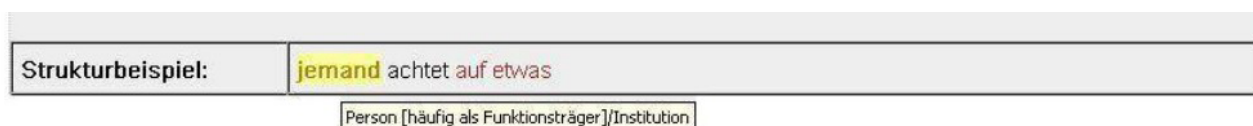


Figure 2: *Achten auf* in *E-VALBU* with dummies and semantic types on mouseover.

## 2.4 StringNet Navigator<sup>16</sup>

*StringNet 4.0* is a corpus-derived online inventory of 1.6 billion English hybrid n-grams (Wible & Tsao 2011). *StringNet Navigator* is the user interface to navigate the resource. Hybrid n-grams, unlike traditional linear n-grams, can be any co-occurrence of POS tags, lexemes, and word forms. For example, *leave [pers pn] in no doubt [conj]* or a construction like *leave me in no doubt that*, which has word forms in the POS tag slots. This way, hybrid n-grams are not just linear strings, but they form a structured net in which one can also search for parent-child relations of constructions. An interesting application of such a structured net is that one can investigate degrees of frozenness and variation in multiword expressions. Wible and Tsao (2011) illustrate this with the string *keep a close eye on*, which, when navigating upward, reveals that *eye* in this string can be replaced by *watch*, and that *close* can be replaced by *careful*, but also that, in the construction *keep a [Adj][N] on*, the verb *keep* is the unsubstitutable lexical anchor to the expression. Tools to search hybrid n-grams will no doubt be of great relevance to constructionists and lexicographers dealing with phraseology in the future.

Another innovative aspect of the *StringNet Navigator* is that it has a collocation search in which collocations are linked to the patterns that contain that collocation. The idea to link patterns and collocations is interesting. However, navigating the tool is still a daunting task, especially for language learners. Suppose one wants to search for possible objects of the verb *bury*. If the learner would search in the *Sketch Engine for Language Learning*, he would get results in a list of possible objects, like *treasure, body, dead, hatchet*. In the collocation tool of *StringNet*, however, one can only search for formal grammatical categories, i.e. nouns, verbs, adjectives, etc., before or after the searched word. These categories are not assigned any complement type or semantic type, which means that the search for nouns before or after *bury* yields results like *churchyard, cemetery, grave, sand, face*, etc. These are indeed nominal collocates that frequently co-occur with *bury*, but they are not the target object collocates the language learner was looking for. To find the target collocates, he/she could start from the patterns, but then, he/she must already have an idea of all the possible phrase structures and/or word forms the object can have, like *bury the [noun]*, *bury their [noun]* or *bury the [noun pl]*, etc. In other words, the restriction to POS tags, lexemes and word forms in the hybrid n-grams has the disadvantage that one cannot search for phrasal collocates with a syntactic function linked to a semantic type or role and lexical sets, for example *bury NP* in which NP = direct object / *[[physical\_object]]* or *[[Body]]* / *treasure, hatchet, dead, body*,....

<sup>16</sup> <http://nav4.stringnet.org>

In *Woordcombinaties* we will implement a functionality to link collocations and patterns, but the collocate sets will be assigned semantic types and will be linked to a syntactic function in the pattern.

No	Pattern	Freq	Relations			
1	bury the [noun]	17	↑	↓	↔	↔
2	bury the [noun sg]	12	↑	↓	↔	↔
3	<b>bury</b> the [noun]	129	↑	↓	↔	↔
4	<b>bury</b> the [noun sg]	90	↑	↓	↔	↔
5	<b>bury</b> the [noun pl]	33	↑	↓	↔	↔
6	burying the [noun]	28	↑	↓	↔	↔
7	buried the [noun]	25	↑	↓	↔	↔
8	buried the [noun]	18	↑	↓	↔	↔

noun		
Words that can appear as the [noun] in: <u>bury the [noun]</u>		
Sort by: spelling, frequency, lemma spelling or lemma frequency		
No	Word (96)	Frequency (129)
1	hatchet	13
2	bodies	7
3	myth	4
4	ashes	3

Figure 3: Patterns and collocates in StringNet 4.0 Navigator.

### 3 Woordcombinaties (Word Combinations)

We will now turn to the description of our own project in which we want to adopt the strengths of the projects described in the previous paragraphs while also including some innovative features which could address their weaknesses. The macrostructure of the pilot will consist of a selection of mid-frequency lexical verbs. These are selected from a vocabulary list of a remedial teaching application for academic Dutch<sup>17</sup> which contains a module for advanced learners of Dutch as a second language. Advanced learners of Dutch as a second language are the user group aimed at in the pilot, but in the long term we can tailor the application to any level of learners, provided that level-stratified corpora are available. As we aim at quick and easy access to phraseological information, it is necessary to deviate from a traditional layered microstructure in which this information is dispersed over a number of senses and subsenses. We offer immediate access to usage patterns in a toolbar instead (Figure 4). Demo screenshots of the combinatorics of the test verb *aanmoedigen* (encourage) will offer a preview of the web application.

#### 3.1 Application Features and Search Options

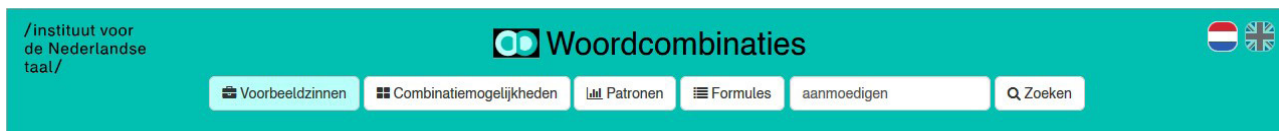


Figure 4: Toolbar of *Woordcombinaties*.

All toolbar buttons can be accessed directly in random order, but the search options are also arranged according to their increasing degree of complexity, ranging from a simple search of the verb in example sentences (*voorbeeldzinnen*), over word sketches with collocates (*combinatiemogelijkheden*), to

<sup>17</sup> <http://www.hogeschooltaal.nl>



pattern-meaning pairs (*patronen*) and conversational routines (*formules*), which are prepatterned situation-bound or speech act-bound utterances with a pragmatic rather than a referential meaning, e.g. apologies, greetings, asking for information. In addition to word searches, we will also offer thematic searches for idioms, proverbs and conversational routines. This option has not been implemented in the first demo screenshots, but we will include the feature in the course of the pilot.

### 3.1.1 Example Sentences (*Voorbeeldzinnen*)

Multiple example sentences provide a bird's eye view of the usage patterns and meaning potentials of the verb. As we already mentioned in discussing the *SkELL* (2.1), these can be used in data-driven learning by learners to work out the different verb senses and usage patterns. Teachers can select good examples for grammar and vocabulary lessons.

The best examples will be retrieved automatically using the GDEX functionality in the Sketch Engine. The GDEX configuration for Dutch will be set such that short and intelligible, but informative sentences elucidating the definition and exhibiting typical patterns of usage are selected. We will post-edit the examples to eliminate mistakes. Post-editing will be restricted to the correction of spelling mistakes and obvious grammar mistakes in order to maintain authenticity. The original *SkELL* provides 40 examples per verb. We will perform tests on which number of examples is sufficient and necessary for an optimal overview of a verb's behavior in Dutch. Above the examples the lemma form of the verb is given, as well as a link to a pop-up window with its conjugation forms.

aanmoedigen werkwoord VORMEN Beeld:

- 1 EVV voelde dat het een kans had en aangemoedigd door een fanatiek publiek ging het team steeds beter spelen. (Kranten, Nederland, 1994)
- 2 In het verleden is lang een beleid gevoerd dat vervroegde uittreding uit de arbeidsmarkt aanmoedigde.
- 3 Zij wil het kabinet aanmoedigen de voorstellen binnen enkele jaren in beleid om te zetten.
- 4 Telewerken werd oorspronkelijk vooral aangemoedigd vanwege de veronderstelde gunstige uitwerking op de dagelijkse files.
- 5 De artiesten worden aangemoedigd door een talrijk publiek!
- 6 Omdat de overheid het sparen voor de oude dag wil aanmoedigen, krijgen pensioenspaarders een korting op hun belastingfactuur.
- 7 Milieuvriendelijke investeringen in de veeteelt worden financieel aangemoedigd via het Landbouwinvesteringsfonds.

Figure 5: GDEX examples.

### 3.1.2 Word Sketches (*Combinatiemogelijkheden*)

Word sketches with collocates can be accessed through the second option on the toolbar: *combinatiemogelijkheden* (combination possibilities). We prefer this term to *word sketch*, because it is self-explanatory: the option shows which words and phrases combine with the searched verb. The functionality can support language learners in finding the right collocations to build sentences. Teachers can use the functionality to decide on which collocations to teach. We will include the following grammatical relations: subjects, direct objects, indirect objects, prepositional objects, subject complements, object complements, adverbials, co-ordination and clausal complementation. To help learners who may not be familiar with syntax terminology, the complement type names are paraphrased by means of questions, such as *who or what encourages?* or *who or what is encouraged?* A notification also mentions that the logical subject is expressed in a *by*-adverbial (*door*-bepaling) in passive sentences.

*SkELL* lists up to 15 collocates for each grammatical relation. We will perform tests on the number of collocates sufficient and necessary for production tasks. We will also examine various rankings



– by score, frequency or alphabetically – and the possibility for users to switch between them. The collocate lists will be post-edited. Mouse clicks on collocates result in example sentences of the collocation. It will not be possible, however, to post-edit all examples of all collocations as well. To give learners something to hold on to, we could check every first example of the collocations.

**subjecten** - Wie/wat moedigt aan? In passieve zinnen kan het logische subject uitgedrukt zijn in een door-bepaling.

publiek supporter ouder toeschouwer succes overheid vriend familie reactie vader moeder klasgenoot fan enthousiasme schare fans

**objecten** - Wie/wat wordt aangemoedigd? In passieve zinnen wordt het logische object het grammaticale subject.

mens kind team deelnemer ander jongere bedrijf vrouw leerling-leerlinge gedrag ondernemerschap man speler medewerker jongen student-studente

**prepositionele objecten** - Waartoe wordt aangemoedigd?

tot ...

**adverbiale bepalingen** - Hoe, hoelang, hoe vaak, etc. moedigt men aan?

luid luidkeels enthousiast fanatiek financieel vaak op die manier sterk fiscaal extra hartstochtelijk zelfs juist verder

**bijzinnen als complement**

(om) te + inf.

**en/of**

ondersteunen steunen helpen stimuleren belonen motiveren begeleiden coachen inspireren uitdagen bevorderen versterken toejuichen feliciteren bevestigen

Figure 6: *Combinatiemogelijkheden* (word sketch) of *aanmoedigen* (encourage).

### 3.1.3 Patterns with Definitions (*Patronen met Definities*)

Examples and word sketches provide a good first impression of usage patterns and meaning. This can be very helpful for advanced learners trying to find target collocates or seeking confirmation of their intuitions regarding a collocation. However, both options have one major disadvantage: the examples and collocations are not explicitly linked to their meaning in context. As patterns and meanings often have a preference for particular sets of collocates, this information is essential and has to be encoded. The pattern functionality will enable learners to build constructions longer and more complex than the binary combinations provided by the word sketches. Patterns with their associated meanings and collocate sets will be accessible through the option *patronen* (patterns)<sup>18</sup>. The slots in the patterns are filled with dummies for the sake of readability. The first pattern in Figure 7, for example, is *iemand moedigt iemand aan* (*someone encourages someone*). The dummies form a limited set which will be established in the course of the pilot. Dummies are assigned colors to distinguish them from similar ones in other syntactic functions in the same pattern. The colored dummies re-occur as anchors in the definition (*Betekenis* in Figure 7) and the colors also re-occur in the corresponding lexical items in the example sentence (*Voorbeeld* in Figure 7). Clicking on the dummies will reveal collocate sets. For example, clicking on *iemand* in the subject position of pattern 1 would reveal the lexical set {*publiek* (*public*), *supporter*, *toeschouwer* (*spectator/audience*)} (Figure 8). The lexical sets will be grouped in semantic types in the database. We will use the same ontology as *PDEV* along with a Dutch version linked to it. In the example of pattern 1 the semantic types are [[Human]] / [[Mens]] and [[Human\_Group]] / [[Mens\_Groep]]. Dependent on the preferences of the users, we can show or hide the semantic types in the application. Links on the right of the screen give access to more examples and to more information categories, for example, the sentence structure (*Bouw* in Figure 9) with obligatory or optional complements and with information on passivisation (*Passief* in Figure 9).

Idioms and proverbs are special types of patterns with very limited lexical preferences and/or specific phrase structures in some or all of the slots. Hence, an idiom like *de strijdbijl begraven* (*bury the*

<sup>18</sup> From the first screenshots it is not clear yet that patterns are assigned meaning. In the final version of the application the toolbar icon ‘*patronen*’ will be replaced by the icon ‘*patronen met definities*’ (patterns with definitions).

*hatchet*) can be encoded as *iemand begraaft de strijdbijl* (*someone buries the hatchet*), but only the subject position is lexically variable whereas the object must include *strijdbijl* (*hatchet*) and the definite article *de* (*the*). Idioms and proverbs will be listed in the ‘patterns with definitions’ section and will be labelled. They will also be accessible through a thematic search option, for example, idioms or proverbs with animal names or body parts, idioms and proverbs for weather conditions, emotions, etc.

Patroon 1	<b>iemand moedigt iemand aan</b>	(meer gegevens)
Betekenis:	iemand juicht iemand toe of supportert voor iemand zodat iemand doorzet of beter presteert.	
Voorbeeld:	Wie zondag de <b>renners</b> op de Kwaremont gaat <b>aanmoedigen</b> , komt helaas een week te laat.	(meer voorbeelden)
Patroon 2	<b>iemand   iets moedigt iemand aan tot iets of (om) te inf.</b>	(meer gegevens)
Betekenis:	iemand   iets stimuleert iemand om te beginnen of door te gaan met een handeling, activiteit of gedraging.	
Voorbeeld:	Onderwijs in eigen taal moet daarom <b>allochtone kinderen aanmoedigen tot zelfstandig denken</b> . Ambitie ontstaat als ouders hun kinderen <b>aanmoedigen om te studeren en te werken</b> .	(meer voorbeelden)
Patroon 3	<b>iemand   iets moedigt iets aan</b>	(meer gegevens)
Betekenis:	iemand   iets bevordert de ontwikkeling of toename van iets	
Voorbeeld:	Zo willen ze <b>culturele investeringen aanmoedigen</b> en voorkomen dat artistiek talent naar het buitenland vlucht.	(meer voorbeelden)

Figure 7: Triples pattern – definition – example.

Patroon 1	<b>iemand moedigt iemand aan</b>	(meer gegevens)
	<div> <div>publiek</div> <div>supporter</div> </div> juicht iemand toe of supportert voor iemand zodat iemand doorzet of beter presteert. dag de <b>renners</b> op de Kwaremont gaat <b>aanmoedigen</b> , komt helaas een week te laat.	(meer voorbeelden)

Figure 8: Lexical set on mouse click.

Patroon 1	<b>iemand moedigt iemand aan</b>	(minder gegevens)
Betekenis:	iemand juicht iemand toe of supportert voor iemand zodat iemand doorzet of beter presteert.	
Voorbeeld:	Wie zondag de <b>renners</b> op de Kwaremont gaat <b>aanmoedigen</b> , komt helaas een week te laat.	(meer voorbeelden)
Bouw:	subject, direct object.	(verberg bouw)
Passief:	mogelijk, met worden en zijn.	(verberg passief)
Voorbeeld:	<ul style="list-style-type: none"> <li>Ik word elke wedstrijd <b>aangemoedigd</b> door vriendin Anke.</li> <li>Ze heeft het heel goed gedaan en is <b>aangemoedigd</b> door haar familie, vrienden en klasgenoten met een mooi spandoek.</li> </ul>	
Commentaar:	vaak passief.	(verberg commentaar)

Figure 9: More pattern information.

### 3.1.4 Conversational routines (formules)

Conversational routines, the fourth option *formules* in the toolbar, deserve special attention in a learners’ application. Like idioms and proverbs, conversational routines are a special type of pattern. Coulmas defines them as “highly conventionalized prepatterned expressions whose occurrence is tied to more or less standardized communication situations” (1981, 1-3). Aijmer (2014: 2) distinguishes three major classes: formulaic speech acts, such as apologizing, thanking and greeting, for example *I’m sorry, but ...* or *Thank you!*, discourse markers, such as *As I say ...* and attitudinal routines expressing the speaker’s attitude or emotion, such as *Go to hell!* Routines come natural to native speakers, but they are difficult to master for non-native language learners.

As the pragmatic function of the routines predominates over the referential meaning (Aijmer 2014: 11), it is not practicable to list them with the semantically motivated pattern-meaning pairs. Aijmer acknowledges that “the referential meaning does not completely disappear, however, but it is ‘overlaid’ with a pragmatic function which may be more or less dominant” (2014: 11). Moreover, verb patterns can serve as templates in “free” sentence building, but conversational routines are conventionalized and fixed to the extent that they cannot be produced by language learners by simply using the patterns as templates. One just has to know the conversational routine. Therefore, it is better to provide access to these formulae in a different way. A separate icon on the toolbar can be used for word searches, but as formulae are bound to specific communication functions and situations, it is only logical to offer a thematic search option as well. Hence, the conversational routines will also be accessible through predefined lists of speech acts and/or communication situations. Common speech acts and situations

will be collected from textbooks and applications for second language learners. For example, ‘asking information’ (speech act/function) in the theme or situation ‘public transport’ would yield a formula like *Hoe laat vertrekt/gaat de trein naar x?* (*What time does the train for x depart?*). The functionality is still being developed at the time of writing this paper, so there is no screenshot to illustrate it. Possibly, thematic search options will be provided in a vertical toolbar on the left.

### 3.2 Corpus, Tools and Methodology

Parts of the project will be automated and parts will be manually produced. For the pilot we compiled a test corpus of about 300 million tokens which consists of newspaper material, spoken material, domain specific texts and fiction. The corpus contains material from the Netherlands and Belgium to reflect language variety in Dutch. The corpus is loaded in the Sketch Engine.<sup>19</sup> It is lemmatized and part-of-speech tagged and has a word sketch grammar to retrieve word sketches. We will also run tests with a parsed corpus. A parsed corpus may (or may not) open better possibilities of automatically preprocessing word sketches and patterns.

Example sentences will be automatically generated with the GDEX functionality, but they will be checked manually in the example search option (option 1 in the toolbar). Word sketches (option 2) will also be automatically retrieved in the Sketch Engine, but noise in the collocate lists will be eliminated manually. Collocates are clustered in complement types, such as subject, direct object, prepositional object, etc., which will make them easily and quickly accessible in the application for sentence building tasks.

Patterns (option 3) will be manually annotated using the CPA-technique in corpus samples of 250 concordances, or 500 or more for more polysemous verbs. The Sketch Engine supports CPA-annotation with pattern numbers in both concordance lines and word sketches. The possibility to annotate in the word sketches as well and the technique of TickBox Lexicography (TBL)<sup>20</sup>, makes it practicable to assign relevant collocates to slots in patterns of the pattern-meaning pairs. This task is best performed manually, as collocates may occur in more than one pattern. The CPA-tool can cluster the annotated patterns automatically, which makes it possible to rank them according to frequency in the dictionary application.

For the lexicographic description of the patterns a tailored pattern editor will be developed, which will be connected with the Sketch Engine. Collocate sets will be grouped in semantic types in the database, using the *PDEV*-ontology and a linked Dutch version of it. Semantic type annotation may be useful for semantic parsing in NLP (El Maarouf et al. 2014). However, as some semantic types may be too abstract for language learners with little or no linguistic or semantic background, it is advisable to display them in the language learners’ application only on demand. In order to visualize semantic differences between syntactically identical patterns, one or two superscript lexical items can represent semantic types and serve as sense markers instead. For example, patterns like *someone checks something*<sup>e-mail</sup> and *someone checks something*<sup>oil level, tyre</sup> will immediately draw the learner’s attention to the target pattern-meaning pair in a straightforward and simple way. The default pattern view option will display elementary information: pattern, definition and example. More information categories, such as information on sentence structure, optionality of complements and passivization, will be made accessible through fold-out links. More examples will be made accessible through links to the corpus.

CPA is a corpus-driven approach which is a workable method to discover socially salient (frequent) patterns of use. However, many idioms and proverbs are cognitively salient, which means they are

<sup>19</sup> <http://sketchengine.co.uk>

<sup>20</sup> <https://www.sketchengine.co.uk/user-guide/user-manual/tickbox-lexicography>

salient, not so much in terms of frequency, but in terms of ease of recall (memorability) (Hanks 2013: 5, 344). In a strictly corpus-driven approach in which only samples of the corpus are analyzed, we may miss out on cognitively salient idioms and proverbs which are not at all that infrequent or rare. Dictionaries and lexical databases have already encoded many of them, so we advocate combining corpus-driven CPA with a corpus-based approach in which we check the currency of already encoded idioms and proverbs in our up-to-date corpus.

Conversational routines (option 4 in the toolbar) pose yet another challenge. They are mainly used in spoken language and in specific social contexts, but so far spoken language is underrepresented in Dutch corpora. Acquisition of a larger corpus of spoken Dutch or language coming close to spoken Dutch, such as subtitles of television programs, will be aimed at, but is beyond the scope of this pilot. Textbooks for language learning and online teaching materials will also be used to make an inventory of the routines used in specific situations.

## 4 Discussion

In the previous paragraphs we expounded upon projects which inspired us to develop *Woordcombinaties* and we offered a preview of our project and the methodology applied. One has to be aware of the fact that the pilot is an experimental lexicographic resource which leaves room for improvement.

We tried to make the application as self-explanatory as possible: on the home page all search options are clearly defined in short instruction sentences. Still, any application for computer-assisted and/or data-driven language learning requires some instruction or brief training (Boulton 2012). Supported by a user-friendly tool, learners and teachers can develop the motivation to learn and teach phraseology, and they can develop better and faster search strategies. More motivation and better search strategies can generate more fluency and proficiency in the target language.

However promising and useful we may find our project, we are well aware of a few issues which will have to be dealt with as it develops. In the first place, there is the issue of the Dutch corpora and their suitability as bases for language learning and teaching tools. A substantial portion of the present corpora is written language from newspapers, domain corpora and the web. It will be suitable enough for the target user group of advanced learners in the pilot, especially if GDEX configurations are fine-tuned to retrieve the best possible examples. However, if we want to develop more level-stratified applications for more user groups in the follow-ups of the pilot, learner corpora for Dutch will have to be structurally developed. More spoken standard language in conversations is required to disclose conversational routines and other situation-bound utterances. To disclose school language, which younger language learners have to master in education, and the combinatorics of words used in education in various topics, subcorpora of textbooks, simplified literature and easy non-fiction texts are required. A textbook corpus should contain books for language learning and textbooks for other topics in education, like history, geography, physics, and so on. Level-stratified corpora are needed to differentiate our application between levels in language learning. Many teachers and institutions provide free access to online teaching materials. In the course of the project we will try to acquire some of these, but a more structural acquisition policy is required in the long term.

Secondly, we are aware of the fact that our project is an eclectic synergy of the projects which inspired us, but we do not see this as a disadvantage. On the contrary, we took the best of all worlds by adopting the positive aspects of these projects and by providing solutions for the missing links in them. Post-editing automatically retrieved information is only one improvement. The most innovative but also the most challenging aspect of the Dutch project is the insertion of the lexical preferences and semantic types in the patterns, and the possibility to switch from collocation dictionary to



pattern dictionary and vice versa. Language learners will not only have access to patterns or collocate lists separately, but will be able to see which collocates fill the slots in a pattern. The collocate lists in the word sketches provide a quick overview of target collocations, but often one also wants to know which collocates are preferred in or restricted to specific pattern-meaning pairs. Therefore, easy switching between the search options will be a useful functionality. Another important innovation is that we will pay more attention to conversational routines and thematic search options to make them easily accessible.

As for the work-flow of the project, we expect the production of the examples-section and the word sketches to be relatively easy and fast, as they will be semi-automatically produced. Pattern description will be a much more complicated and time-consuming task. Another more daunting task in due time could be a clustering of collocations according to some common Mel'čukian lexical functions as is suggested by Atkins and Rundell (2008: 151) and/or a clustering of collocations by generic semantic categories. Promising experiments in distributional semantics have been conducted with regard to this (Wanner, Ferraro, & Moreno, 2017), so we will follow developments in this field attentively.

Last but not least, with a view to reusability of existing resources, we have the intention to assimilate other phraseological resources, such as the phraseological information from *ANW* and multiword expressions in computational lexicons, such as *DuELME*<sup>21</sup>, *RBN*<sup>22</sup> and *Cornetto*<sup>23</sup>. Information in the latter resources is not presented in a learner-friendly environment, but the contents are useful. They can be checked for currency in our corpus and adapted to the format of our application.

## 5 Conclusion

The pilot of *Woordcombinaties* is only the first move towards a fully-fledged phraseological resource for Dutch. We realize that such a resource is an ambitious long-term goal, but at the same time we have to acknowledge that the good match of collocations, idioms and patterns is necessary in order to support language learners and users of Dutch better in the future. Follow-ups of the pilot will be essential to prevent the project from becoming another fragment in the fragmented Dutch phraseological landscape. A first sequel to this pilot will consist of the description of more verbs and the combinatorics of nouns and adjectives. Special attention will be paid to core vocabulary, level-stratified spin-offs of the application and the optimization of Dutch corpora for this purpose.

## References

- Aijmer, K. (2014). *Conversational Routines in English: Convention and Creativity*. Routledge.
- Algemeen Nederlands Woordenboek (ANW)*. Accessed on: <http://anw.ivdnt.org> [22/03/2018].
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bahns, J., Eldaw, M. (1993). Should we teach EFL students collocations? In *System*, 21 (3), pp. 101-114.
- Boulton, A. (2012). What Data for Data-Driven Learning? European Association for Computer-Assisted Language Learning (EUROCALL).
- Contrastive Verb Valency Dictionary (CVVD)*. Accessed on: <http://www.cvv.d.ugent.be> [21/03/2018].
- Cornetto Demo. Combinatorial and Relational Network as Toolkit for Dutch Language Technology*. Accessed at: <http://cornetto.clarin.inl.nl/index.html> [23/03/2018].

21 <http://duelme.inl.nl>

22 <http://tst.inl.nl/producten/rbn>

23 <http://cornetto.clarin.inl.nl/index.html>



- Coulmas, F. (Ed.). (1981). *Conversational Routine: Explorations in Standardized Communication Situations and Prepatterned Speech*. Mouton.
- Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. In *Applied Linguistics*, 2 (3), pp. 223–235.
- Cowie, A. P. (2008). Phraseology. In *Practical Lexicography*, pp. 163–167.
- DuELME. *Dutch Electronic Lexicon of Multiword Expressions*. Accessed at: <http://duelme.inl.nl> [23/03/2018].
- El Maarouf, I., Baisa, V., Bradbury, J., & Hanks, P. (2014). Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing. In *Proceedings of LREC*, pp. 1001–1006.
- E-VALBU. *Das Elektronische Valenzwörterbuch Deutscher Verben*. Accessed at: <http://hypermedia.ids-mannheim.de/evalbu/index.html> [22/03/2018].
- Kleijn, P. de (1999). Nederlandse woordenboeken als basis voor een woordenboek van vaste verbindingen? In *Neerlandica Extra Muros*, 37, pp. 14–22.
- Kleijn, P. de (2003). *Combinatiewoordenboek. Nederlandse substantieven met hun vaste verba*. Rozenberg Publishers, Amsterdam. Online version accessed at: <https://combinatiewoordenboek.nl> [21/03/2018].
- Fenoulhet, J. (1991). Fraseologie en lexicografie. In *Handelingen Elfde Colloquium Neerlandicum Utrecht 1991*, Woubrugge, IVN, pp. 107–120.
- Fillmore, C.J., Kay, P. & O'Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. In *Language* 64 (3), pp. 501–538.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*.
- Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. In *International Journal of Lexicography*, 25(3), pp. 273–296.
- Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. In *ReCALL*, 26(2), pp. 128–146.
- Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.
- Granger, S., and Meunier, F. (2008). Phraseology in language learning and teaching: Where to from here? In *Phraseology in Foreign Language Learning and Teaching*, (Amsterdam: John Benjamins Publishing Company), pp. 247–252.
- Gries, S. T. (2013). 50-something years of work on collocations. In *International Journal of Corpus Linguistics*, 18 (1), pp. 137–166.
- Groot, H. (1999). *Van Dale Idiomwoordenboek*. Utrecht: Van Dale Lexicografie.
- Halliday, M., Matthiessen, C. M. & Matthiessen, C. (2014). *An Introduction to Functional Grammar*. Routledge.
- Hanks, P. (2004). Corpus pattern analysis. In *Euralex Proceedings*, 1, pp. 87–98.
- Hanks, P. (2008). Mapping meaning onto use: a Pattern Dictionary of English Verbs. *Proceedings of the AACL*.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.
- Hilgsmann, Ph. (2005). Enkele recente woordenboeken Nederlands onder de NVT-loep. In *Neerlandica extra Muros*, 43, pp. 27–38.
- Hogeschooltaal. Accessed at: <https://www.hogeschooltaal.nl> [23/03/2018].
- Howarth, P. (1998). Phraseology and second language proficiency. In *Applied Linguistics*, 19 (1), pp. 24–44.
- Jesen, V. (2006). Phraseologie und Fremdsprachenlernen. Zur Problematik einer angemessenen phraseodidaktischen Umsetzung. In *Linguistik Online*, 27 (2), pp. 137–147.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008, July). GDEX: Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.
- Kubczak, J. (2014). Das Versteckspiel der Komplemente - wie obligatorisch sind obligatorische Komplemente und wie geht man damit in den VALBUS um. Accessed at: <http://nbn-resolving.de/urn:nbn:de:bsz:mh39-29323> [22/03/2018].
- Pattern Dictionary of English Verbs (PDEV). Accessed at: <http://pdev.org.uk> [21/03/2018].
- Peters, E. (2013). Collocaties leren in een vreemde taal. In *Handelingen der Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis*, 56, pp. 177–192.
- Referentiebestand Nederlands Online (RBN). Accessed at: <http://tst.inl.nl/producten/rbn> [22/03/2018].
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sketch Engine for Language Learning (SkELL). Accessed at: <https://skell.sketchengine.co.uk/run.cgi/skell> [21/03/2018].
- StringNet Navigator 4.0. Accessed at: <http://nav4.stringnet.org> [22/03/2018].

- Verburg, M. E., Stumpel, R. J. T., & de Groot, H. (Eds.). (2017). *Van Dale pocketwoordenboek Nederlands als tweede taal (NT2)*. Utrecht: Van Dale Lexicografie.
- Wanner, L., Ferraro, G., & Moreno, P. (2017). Towards distributional semantics-based classification of collocations for collocation dictionaries. In *International Journal of Lexicography*, 30(2), pp. 167-186.
- Wible, D., Tsao, N. L. (2011). The StringNet lexico-grammatical knowledgebase and its applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 128-130. Association for Computational Linguistics.
- Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. In *Applied Linguistics*, 21 (4), pp. 463-489.

# ColloCaid: A Real-time Tool to Help Academic Writers with English Collocations

**Robert Lew<sup>1</sup>, Ana Frankenberg-Garcia<sup>2</sup>, Geraint Paul Rees<sup>2</sup>, Jonathan C. Roberts<sup>3</sup>, Nirwan Sharma<sup>3</sup>**

<sup>1</sup>Faculty of English, Adam Mickiewicz University in Poznań, <sup>2</sup>School of Literature and Languages, University of Surrey, <sup>3</sup>School of Computer Science, Bangor University

E-mail: rlew@amu.edu.pl, a.frankenberg-garcia@surrey.ac.uk, g.rees@surrey.ac.uk, j.c.roberts@bangor.ac.uk, n.sharma@bangor.ac.uk

## Abstract

Writing is a cognitively challenging activity that can benefit from lexicographic support. Academic writing in English presents a particular challenge, given the extent of use of English for this purpose. The ColloCaid tool, currently under development, responds to this challenge. It is intended to assist academic English writers by providing collocation suggestions, as well as alerting writers to unconventional collocational choices as they write. The underlying collocational data are based on a carefully curated set of about 500 collocational bases (nouns, verbs, and adjectives) characteristic of academic English, and their collocates with illustrative examples. These data have been derived from state-of-the-art corpora of academic English and academic vocabulary lists. The manual curation by expert lexicographers and reliance on specifically Academic English textual resources are what distinguishes ColloCaid from existing collocational resources. A further characteristic of ColloCaid is its strong emphasis on usability. The tool draws on dictionary-user research, findings in information visualization, as well as usability testing specific to ColloCaid in order to find an optimal amount of collocation prompts, and the best way to present them to the user.

**Keywords:** writing assistant, collocation, academic writing, English for academic purposes

## 1 Background and Rationale

As lexicography moves forward into the digital age (Lew & De Schryver 2014), stand-alone dictionaries are gradually giving way to sophisticated and specialized lexicographic devices integrated in digital tools which may be optimized for specific tasks. One task that requires extensive lexicographic assistance is writing. The present contribution introduces the ColloCaid tool, which will be able to suggest collocational choices in real time during the process of writing, with a focus on academic English. ColloCaid recognizes that there are no native users of academic language (Frankenberg-Garcia 2017; Hyland 2006; Kosem 2010), and is therefore foreseen to be of value to both native and non-native writers who do not have sufficient command of academic English collocations.

Existing automated collocation-extraction tools tend to adopt a one-size-fits-all strategy. This is true of the domains they address; for example, in addition to other functions, Grammarly, Read & Write Gold, and Write Away provide collocation suggestions for general English. It is also true of the type of collocations they deal with; Wanner, Verlinde and Alonso Ramos (2013) argue that the assumption that all collocation errors can be corrected in the same way is mistaken. They claim instead that tools should focus on collocations comprising the parts of speech which pose writers most problems. Those few existing tools that do deal with specific domains and genres are undoubtedly useful for the writer, however they address a limited range of collocation errors. For example, although Cambridge's Write and Improve provides non-native writers with feedback on set writing tasks, as far as

collocations are concerned this feedback is limited to highlighting missing or incorrect prepositions. In Spanish ArText provides feedback on texts from the domains of Public Administration, Medicine and Tourism. Its feedback on collocations centers on the over or underuse of connectors such as *por lo tanto* (therefore) and *sin embargo* (however).

In contrast, the emphasis of the present project is on providing carefully curated content based on relevant and extensive resources focusing on general academic English. Starting from the generally accepted notion (Hausmann 2004; Martin 2008) of a collocate comprising a base (sometimes called a node) and collocate (sometimes called a collocator), up-to-date academic vocabulary lists are first referenced to identify the relevant sets of collocational bases, then a number of state-of-the-art corpora are explored to identify the salient collocates of these collocational bases.

## 2 Curated Collocational Data

### 2.1 Master Word List

For noun bases, we plan to include their typical pre-modifiers, verbs that take those noun bases as subjects and objects, as well as any characteristic prepositions. For verb bases, adverbial modifiers and prepositions would be added. Finally, adjective bases would be supplied with their salient pre-modifying adverbs. To supplement the ‘positive evidence’, the tool should be able to identify inappropriate collocational choices attested in learner corpora and other sources.

The rationale underlying the decision to concentrate on these types of bases and collocates is that writers are more likely to start with a noun in mind and then look up a verb collocate than start with a verb and then search for a noun collocate. For example (see Figure 1), a writer might wish to comment on a certain *measure*, provoking the questions (and potential collocates): ‘What preposition should I use?’ (*a measure of/for*), ‘How do I characterize the measure?’ (*a reliable/objective/quality measure*), ‘How do I say that this measure was used?’ (*we adopted/introduced/developed a measure*), ‘What does the measure do?’ (*a measure captures/indicates/represents something*). Conversely, it is unlikely that a writer would think of the verb *develop* then wonder ‘What to develop?’ (*a theory, a measure, a system*). Nonetheless, it is possible that he or she might wonder how to qualify the verb in an idiomatic way. For example, the idea that *CO<sub>2</sub> emissions contribute to global warming* might prompt the questions: ‘To what degree?’ (*significantly, substantially*), or when a model is found to *account for patterns in data*, one might wonder what adverb to use to qualify the degree of fit of the model (*fully/largely/partially account*). Similarly, adjectives might also provoke collocational doubts during the writing process, for example, when two groups or conditions turn out to be *different*, typically a question arises: ‘How different?’ (*substantially, significantly*).

Even with this restriction on the parts of speech to be considered for inclusion in the master word list, the number of potential bases would be impractically large to be wholly relevant to the user or to permit any thoroughgoing lexicographic treatment. To address this problem, the results of three widely recognized studies of EAP lexis were applied to draw out those node words which would likely be relevant to EAP writers. The first, the Academic Vocabulary List (AVL, Gardner & Davies 2014), comprises 3,000 core lemmas that occur across a range of academic disciplines in the 120-million-word academic sub-corpus of the Corpus of Contemporary American English (COCA, Davies 2008-). Some 2,700 of these AVL lemmas fell within the part-of-speech categories specified above. Durrant (2016) found that only 427 AVL items were found frequently in over 90% of disciplines in university student writing as represented by the BAWE corpus (Alsop & Nesi 2009). Of these 174 were nouns, 136 verbs and 79 adjectives. Applying this AVL-BAWE filter gave a workable number of potential node words. Further validation is provided by the Academic Keyword List (AKL, Paquot

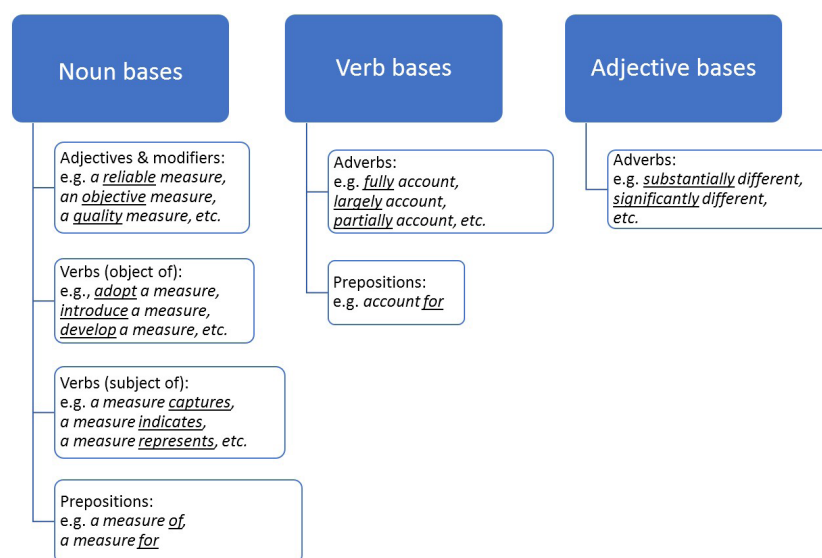


Figure 1: Types of collocational nodes included in ColloCaid, with examples.

2010). Cross-referencing the 353 nouns, 233 verbs and 180 adjectives contained in the AKL with the results of the AVL-BAWE filtered list, provided another means of drawing out potentially useful node words. A final means of filtering relied on the itemized list of 526 noun bases, 96 verb bases and 83 adjective bases of the Academic Collocation List (ACL, Ackermann & Chen 2013) found in the appendix of the *Longman Collocations Dictionary* (Mayor 2013). Table 1 shows the crossover among these three sources.

Table 1: EAP vocabulary considered in ColloCaid.

	AVL-BAWE lemmas	AKL lemmas	ACL lemmas	Total EAP lemmas considered	Lemmas attested in all three lists	Lemmas in at least two lists (ColloCaid)
Nouns	172	353	525	643	125	282
Verbs	129	233	95	283	38	136
Adjectives	86	180	83	231	24	94
Total	387	766	703	1157	187	513

Lemmas attested in at least two of the lists were considered as bases in the master list, with priority given to those 187 lemmas present in all three lists. Ultimately, the decision about the inclusion of the 513 node words in the final ColloCaid tool would depend on their collocational behavior. The following section sets out how this behavior was examined.

## 2.2 Collocates and Examples of Use

Collocational bases (see previous section) were looked up using the Sketch Engine (Kilgariff et al. 2014; Kilgariff et al. 2004).

As has been seen from the discussion of vocabulary lists above, corpora of student writing, namely BAWE and LOCNESS, were used to select collocation bases that novice writers were likely to use. However, corpora of professional academic writing representing ‘expert performances’ (Bazerman 1994: 131) are a more appropriate source of collocation information. Two such corpora in the Sketch Engine, made available with the kind permission of Pearson Longman and Oxford



University Press, were consulted: the Pearson International Corpus of Academic English (PICA, Ackermann et al. 2011) and the Oxford Corpus of Academic English (OCAE). With around 70 million words of expert academic writing, the OCAE is more than double the size of PICA, and was therefore given priority.

The Word Sketch tool was employed to list common collocates by syntactic set, arranged by their logDice scores (see Figure 2), currently the default measure of collocability in the Sketch Engine (Kilgarriff & Kosem 2012). By inspecting word sketches for a random sample of high and low-frequency bases from the different part-of-speech categories, a set of logDice and frequency thresholds corresponding with our intuitive judgements about collocations for EAP writers was found. The thresholds arrived at were a logDice score of  $\geq 5$  for all parts of speech with minimum co-occurrence frequency of 10 for lexical collocates and 100 for prepositions. This stage offers the opportunity to further curate the data. Collocates which are too general to be of relevance to the user e.g. *own* and *good*, in the modifier measure, are filtered out; as are base-collocate pairs which are obviously restricted to a small number of disciplines, for example, *entrepreneurial* found in the modifier relation for *ability*. The collocation *entrepreneurial ability* is likely not of interest to users working outside business studies and related disciplines, while it is highly likely that users working in these disciplines would have mastered this collocation.

important (adjective)		
Oxford Corpus of Academic English (April 2012) freq = 55,647 (659.13 per million)		
ADV ADJ*		12.29
particularly +	912	10.94
is particularly important		
equally +	401	10.40
is equally important		
especially +	430	10.35
is especially important		
increasingly +	398	10.17
an increasingly important		
very +	1,495	10.03
very important		
extremely +	296	9.75
is extremely important		
as +	839	9.52
as important as		
so +	361	9.31
is so important		
critically +	124	9.13
critically important		
vitality +	118	9.11
is vitally important		
potentially +	138	8.79
potentially important		
crucially	78	8.52
is crucially important		
really	75	8.19

Figure 2: Query in the Oxford Corpus of Academic English for the collocational node *important* using Sketch Engine.

The collocates selected as above are systematically entered into a spreadsheet (see Figure 3), one collocate per row. The spreadsheet includes the base form along with its syntactic class (POS), type of collocational relation, the collocate, its raw frequency of co-occurrence with the base in the Oxford Corpus of Academic English, and the corresponding logDice score. Following the finding reported in Frankenberg-Garcia (2014; 2015) that one example alone may not be sufficient to aid language production, corpus citations are used to supply three examples per each collocate-base pair. In addition to the revision of base-collocate pairs in Word Sketch outlined above, the extraction of citations from

KWIC lines offers another opportunity to filter out those collocations which are predominantly used in a restricted set of disciplines. For example, from Word Sketch alone there was nothing about the collocation *unauthorised access* which suggested its usage is restricted to a particular field. However, while collecting citations from KWIC lines it became apparent that all instances of this collocation were related to computer science.

The examples included are based on corpus citations but are rarely verbatim excerpts. Elements not central to the core meaning expressed in the citation, primarily certain prepositional phrases and adjectives, are removed so as not to distract the user. To protect the identity of the source of the citations proper nouns are deleted or replaced with pronouns, e.g. *It is sometimes said that Watson and Crick discovered DNA* becomes *It is sometimes said that they discovered DNA*; numbers and dates are rounded, e.g. *1982* becomes *1980*; numerical references to figures and tables are changed, e.g. *Table 7* becomes *Table 1*; and in-text citations in author-date styles, e.g. *(Surname, 2018)*, are changed to a documentary-note style, e.g. *[1]*.

BASE	POS	RELATION	COLLOCATE	CO-	ASSOCI	EXAMPLE1
equal	j	ADV ADJ*	roughly	108	11.42	the latter two groups had roughly equal rates of break
equal	j	ADV ADJ*	exactly	62	10.85	total costs and our total revenues are exactly equal
equal	j	ADV ADJ*	nearly	57	9.89	three experiments were performed using nearly equal
equal	j	ADV ADJ*	almost	69	9.26	men and women are apparently almost equal now in t
equal	j	ADV ADJ*	formally	10	8.35	this occurs where treatment is formally equal
equal	j	ADV ADJ*	necessarily	14	7.81	attachment does not necessarily equal ownership
equal	j	ADV ADJ*	relatively	20	5.48	all household members have relatively equal access to
important	j	ADV ADJ*	particularly	912	10.94	the sensitivity of a test is particularly important
important	j	ADV ADJ*	equally	401	10.4	the two stages are equally important and interlinked
important	j	ADV ADJ*	especially	430	10.35	especially important were the localization of brain and
important	j	ADV ADJ*	increasingly	398	10.17	public relations is becoming an increasingly important
important	j	ADV ADJ*	very	1 495	10.03	it is very important for an economy to be stable
important	j	ADV ADJ*	extremely	296	9.75	they see work and its consequences as extremely imp
important	j	ADV ADJ*	critically	124	9.13	determine which points of critically important inform
important	j	ADV ADJ*	vitality	118	9.11	the link between the two concepts is therefore vitality

Figure 3: Excerpt from a collocations database underlying ColloCaid.

### 3 ColloCaid as a User-friendly Tool

Writing relies on cognitive processes such as user-attention, working memory and content retrieval from long-term memory in order to utilize different types of knowledges (domain, linguistic, pragmatic and procedural) for text production (Alamargot & Chanquoy 2001). In addition to this, features of the task environment such as the nature of audience, collaborators, already-composed text and the medium of writing add to the cognitive workload of the user. In the context of a digital learning environment this includes prompts and associated information offered through writing assistants, such as ColloCaid, which require further information processing resulting in new decision-making demands while performing the writing task. Therefore, from a learning perspective, this new information from a digital tool should be integrated and displayed to the user in a manner which does not disrupt the primary task of writing. Furthermore, after using this information, the user should be able to resume the writing task. Using a learning-centric approach we aim to prototype interactive tools which integrate lexicographic information into existing forms of text-editors, thus providing collocation information in context of the writing task. We then intend to evaluate these prototypes in order to understand how the new information is appraised by the users for improving their texts as well as developing writing expertise. The insights from these evaluations will help in further improving the design of ColloCaid and similar tools, and potentially offer opportunities to explore novel interaction and information-visualization techniques which may be appropriate for user-learning and improving

writing using text-editors (Roberts et al. 2017).

### 3.1 Example Scenario

Most free-to-use and commercial text editors offer a similar set of features (spellcheck, word-repetition, grammar check, etc.) to support users during the writing task. This provides us with a real-world context for using a task-centered design approach for integrating the collocation information into the existing user-workflow. We illustrate this approach using a task-based scenario which may inform the design of ColloCaid (Lewis & Rieman 1994). The tool would monitor the progress of the user as is standard in most text-editors which prompt the user when an error is encountered and/or a suggestion is recommended. In a similar manner, as soon as the user types one of the nodes, the tool would prompt the user that possible collocations may be available for that node (shown using dashed line under the node *research* in Figure 4). When the user interacts with this highlighted node, the tool would offer collocation suggestions, as in a simulated example in Figure 4, where the writer is given general patterns with the node *research* (in this case, the noun), a word which several studies of EAP lexis have highlighted as important; syntactic disambiguation needs to be dealt with in the occasional cases where there exist identically spelled nodes representing more than one syntactic category. In our case, a pop-up window would appear, indicating (here with pluses) that finer detail is available (this process is called drill-down).

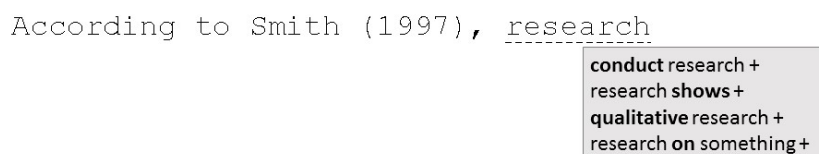


Figure 4: A pop-up general prompt triggered by the collocational node *research*.

To continue our example, let us assume the writer wants to report here on the research so far, therefore she clicks on the ‘research **shows**’ combination; to this, ColloCaid might respond by presenting a more detailed list of collocational choices, as in Figure 5, for example. It is important not to flood the user with too much information, a good general guide being considerations of working memory capacity (Miller 1956).

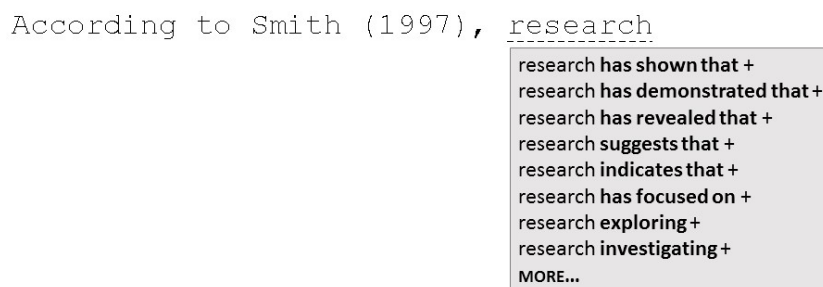


Figure 5: A narrower list of semantically related collocates are presented, following the writer’s selection.

At this point, the choices available at the top of the prompt have been narrowed down to collocates that talk about research indicating something, possibly accompanied by related salient meanings (here towards the bottom). Again, plus symbols indicate that further information is available for each and every row. In this case, these would be the terminal nodes in the form of examples (Figure 6), which further

guide the user's writing. Here the user chose to use *suggest*, and examples are offered that illustrate this particular combination. In line with the recent finding (Frankenberg-Garcia 2015) that three examples are more helpful than a single one in supporting the writing process, three examples are given.

According to Smith (1997), research

research suggests that happiness is likely to be higher if...  
past research suggests that the public tends to...  
although research suggests that volunteering is in general beneficial...

Figure 3. Examples of three alternative sentence combinations

## 4 Conclusion

The present project aims to develop an intuitive lexicographic resource integrated with digital writing environments to help academic English writers write more idiomatically in terms of their collocational choices. This paper has discussed the process of deciding which data the ColloCaid tool should cover, how this data is curated, and how it might be presented on screen in a way that is useful to the end-user. Thus far, the focus has been on 'positive evidence'. Lexicographically, this has involved reference to existing studies of academic lexis and corpora of expert academic writing, while from a visualization perspective it has focused on existing research on the on-screen visualization of text. The next step in the development process involves complementing this evidence. Lexicographically, this means adding information about those collocations which tend to present problems for EAP writers, and, from both a lexicographic and visualization perspective, conducting end-user studies to evaluate the tool. In completing the development process, it is anticipated that ColloCaid will provide useful contributions to the fields of human computer interaction, data visualization and lexicography. More importantly, it is hoped that the tool will make a positive practical difference to EAP writers of many proficiency levels, language backgrounds, and academic career stages, helping them to concentrate on the content of their writing and agonize less over the writing process.

## References

### Writing Tools

- Grammarly. <https://www.grammarly.com>  
Read & Write. <https://www.texthelp.com/en-us/products/read-write/>  
WriteAway. <http://writeaway.nlpweb.org/>

### Other References

- Ackermann, K. & Chen, Y.-H. (2013). Developing the Academic Collocations List (ACL) – A Corpus-driven and Expert-judged Approach. *Journal of English for Academic Purposes* 12. 235-247.  
Ackermann, K., de Jong, J., Kilgariff, A. & Tugwell, D. (2011). The Pearson International Corpus of Academic English (PICA-E).  
Alamargot, D. & Chanquoy, L. (2001). *Through the models of writing. Studies in writing*. Dordrecht, Netherlands ; Boston: Kluwer Academic Publishers.  
Alsop, S. & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4 (1). 71-83.  
Bazerman, C. (1994). *Constructing Experience*. Carbondale: Southern Illinois University Press.  
Davies, M. (2008-). The Corpus of Contemporary American English.

- Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes* 43. 49-61.
- Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL* 26 (2). 128-146.
- Frankenberg-Garcia, A. (2015). Dictionaries and encoding examples to support language production. *International Journal of Lexicography* 28 (4). 490-512.
- Frankenberg-Garcia, A. (2017). Assessing the productive collocation repertoire of writers for the development of dedicated writing assistant tools. *Electronic Lexicography in the 21st Century (eLex 2017)*. Leiden.
- Gardner, D. & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics* 35 (3). 305-327.
- Hausmann, F. J. (2004). Was sind eigentlich Kollokationen. In K. Steyer (ed.), *Wortverbindungen - mehr oder weniger fest*, Berlin: de Gruyter. 309-334.
- Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book*. London/New York: Routledge.
- Kilgariff, A. et al. (2014). The Sketch Engine: Ten years on. *Lexicography* 1. 7-36.
- Kilgariff, A. & Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger, M. Paquot (eds.), *Electronic lexicography*, Oxford: Oxford University Press. 31-55.
- Kilgariff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*, Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 105-116.
- Kosem, I. (2010). Designing a model for a corpus-driven dictionary of Academic English. Ph.D., Aston University.
- Lew, R. & De Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography* 27 (4). 341-359.
- Lewis, C. & Rieman, J. (1994). *Task-Centered User Interface Design: A Practical Introduction*.
- Martin, W. (2008). A unified approach to semantic frames and collocational patterns. In S. Granger, F. Meunier (eds.), *Phraseology: An interdisciplinary perspective*, Amsterdam: John Benjamins. 51-66.
- Mayor, M. (2013). *Longman Collocations Dictionary and Thesaurus*. Harlow: Pearson Education.
- Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (2). 81-97.
- Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.
- Roberts, J. C., Frankenberg-Garcia, A., Lew, R., Rees, G. P. & Pereda, J. (2017). Visualisation and graphical techniques to help writers write more idiomatically. *IEEE Conference on Visualization (VIS)*. Pheonix, Arizona.
- Wanner, L., Verlinde, S. & Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. *eLex 2013, 2013*, 427-487.

## Acknowledgements

This research was supported by the Arts and Humanities Research Council [grant number AH/P003508/1].



# Looking for a Needle in a Haystack: Semi-automatic Creation of a Latvian Multi-word Dictionary from Small Monolingual Corpora

**Inguna Skadiņa**

*University of Latvia, Institute of Mathematics and Computer Science*

*E-mail: inguna.skadina@lumii.lv*

## Abstract

Multiword expressions (MWEs) are an indispensable part of almost any dictionary. However, the identification of missing MWEs that have recently appeared in a language is not a simple task. In this paper we describe automated methods for MWE identification in a rather small Latvian text corpora. We propose starting with the application of statistical measures to identify a wide range of MWEs and then applying linguistically motivated filters to clean the list of initially extracted MWE candidates. We show that for morphologically rich languages, such as Latvian, in cases with a small amount of language data better results can be achieved with lemmatized data. We also demonstrate that in the case of a small general domain (balanced) corpus, automatic methods can be used to find good MWE candidates – terminological units, named entities and some lexicalized phrases. However, finding idiomatic expressions in small, general domain corpora is looking for a needle in a haystack: only a larger or more expressive corpus can help in the identification process.

**Keywords:** multi-word expressions, low resourced languages, collocations, named entities, terminology

## 1 Introduction

Multi-word expressions (MWEs), often defined as “lexical items that (a) can be decomposed in multiple lexemes and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity” (Baldwin and Kim 2010: 269), are indispensable part of almost every dictionary – general or terminological, monolingual or multi-lingual. Most commonly used MWE categories include idioms, phrasal verbs, multi-word conjunctions and prepositions, multi-word terms and named entities. While idiomatic expressions and other MWE categories that are used in a language for many years are usually fixed in printed and electronic dictionaries, idiomatic expressions, verbal constructions and terms (as well as named entities) that have more recently appeared in a language are usually not included, because manual identification (recognition) of such missing lexical items is a difficult task.

MWEs are frequently seen as a “pain in neck” (Sag et al. 2002: 1), because identification and processing of MWEs is a complicated task for many natural language processing applications. Different methods of how MWEs could be identified and extracted have been researched for several decades. These include statistical, linguistic-based and hybrid approaches (e.g., Ramisch 2015, Constant et al. 2017). Some methods are designed for specific MWE categories, e.g., noun compounds or light-verb constructions, while others try to cover different MWE categories.

The role of MWEs in natural language processing, especially parsing, has been addressed in the recent COST action ParseMe - PARSing and Multi-word Expressions (Savary et al. 2015). One of the outcomes of the PARSEME project is a survey of the state of the art techniques for MWE processing (Constant et al. 2017). This survey aims “to shed light on how MWEs are handled in NLP applications” (Constant et al. 2017: 839), in particular, in parsing and MT tasks. The results show that

most research on MWE identification, extraction, annotation and translation addresses widely used languages with large language corpora, while much less work has been done on languages that lack such broad resources.

However, the problem of MWE identification and extraction for the Latvian language is being addressed in a large-scale national research project, “Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian” (Gruzitis et al. 2018). This aims to create multi-layered semantically annotated language resources for Latvian, anchored in widely acknowledged multilingual representations (AMR, PropBank, FrameNet, Universal Dependencies, Grammatical Framework, BabelNet, DBpedia), that are required for the development of natural language understanding and generation applications.

An important role in this set of language resources is assigned to the tools for identification, extraction and annotation of multiword expressions. These tools aim to extract lists of good quality MWE candidates, which, (1) can be delivered as open experimental MWE lexicon for Latvian, and, (2) after manual inspection, will be added to the largest Latvian open lexical database, *tezaurs.lv* (Spektors et al. 2016).

Automated processing of multi-word expressions in Latvian is mostly studied in the context of machine translation. During decades when rule-based machine translation systems were dominant, special MWE dictionaries were created manually or semi-automatically. Such an approach was also chosen by Deksne et al. (2008) for an English-Latvian rule-based machine translation system. The authors proposed using a special, manually created, dictionary of MWEs together with a set of MWE processing rules, and to include additional MWE processing step during parsing. In the era of statistical machine translation (SMT), Pinnis and Skadiņš (2012) investigated a term translation problem for domain specific SMT. Using automated methods, Pinnis (2013) also created a multilingual term dictionary (which includes multi-word terms, too) and demonstrated its importance in statistical machine translation. Finally, Skadiņa (2016) reported improvements in machine translation output when an automatically extracted MWE dictionary is integrated into a domain specific machine translation system.

All these solutions use automatically extracted MWE dictionaries in another natural language processing task, namely machine translation. In the case of statistical machine translation, the dictionary of automatically extracted MWE candidates can contain noise, e.g. parts of MWEs (shorter phrases), widely used phrases that are not terms, or even some frequently used sequences of words. In the case of lexicon building and supplementing, where incorrect phrases create additional work for lexicographers, the quality of extracted MWEs and proportion of incorrect candidates is very important.

In this paper we describe the process and strategies for finding Latvian MWEs using a rather small amount of data. We propose using statistical measures at first and then apply linguistic filters to avoid ungrammatical, but frequent sequences of words. We demonstrate that in the case of a small general domain (balanced) corpus automatic methods can be used to find good MWE candidates – terms, named entities and some lexicalized phrases. However, a rather small balanced corpus is not suitable for the identification of idiomatic expressions.

## 2 Strategies for MWE Identification and Extraction

The Latvian language is often mentioned among morphologically rich under-resourced languages (e.g., Skadiņa et al. 2012). For morphologically rich languages, MWE identification and extraction usually consists of two steps – at first morpho-syntactic patterns are applied to extract the initial list

of MWE candidates, then this list is filtered by means of statistical measures (e.g. Pinnis et al. 2012, Ramisch 2015). The main limitation of this approach is that it converts only MWEs that represent particular linguistic patterns (usually noun phrases, sometimes verb phrases), leaving other MWEs out.

The aim of our work is to support automatic identification of different categories of MWEs that, after manual inspection, could be then added to the Latvian explanatory dictionary *tezaurs.lv*. We thus propose starting with the application of statistical measures. This allows us to identify a wide range of MWE categories, although it could also result in a high amount of ungrammatical constructions. Thus, as the next step, we apply linguistically motivated filters (patterns) to clean the list of initially extracted MWE candidates.

## 2.1 Data

Three different datasets were used in our experiments: the Balanced Corpus of the Modern Latvian language (Levāne-Petrova, 2012), the Latvian-Lithuanian parallel corpus (Utka et al. 2012) and Open Subtitles corpus (Lison and Tiedemann 2016). Depending on the related experiment these corpora were used as the original raw text corpus, lemmatized corpus or morphologically annotated corpus. Table 1 provides general information about these data sets.

Table 1: Corpora used for experiments

Corpus	Size			
	Sentences (thousands)	Tokens (million)	Unique tokens (thousands)	Unique lemmas (thousands)
Balanced Corpus of the Modern Latvian language	148	5,54	408,01	111,59
Latvian-Lithuanian parallel corpus	223	3,24	307,53	87,88
Open Subtitles corpus	454	2,37	117,01	56,44

The Balanced Corpus of the Modern Latvian language only modern, standard Latvian language texts that were written no more than 20 years ago. The corpus was collected using the following balancing criteria: 55% periodicals (27% national newspapers, 22% regional newspapers, 14% internet news, 13% special periodicals and 24% popular periodicals), 20% fiction, 10% scientific publications, 8% legal texts, 5% different other texts and 2% parliamentary transcripts.

The starting point for the Latvian-Lithuanian parallel corpus were texts that are written either in Latvian or Lithuanian and then translated into the other language. However, during the collection process it was discovered that such texts are insufficient to reach the goal of eight million words. Thus, legal texts, usually written in English and then translated into Baltic languages, were also included in the corpus. The resulting parallel corpus contains 19.3% texts that were originally written in Latvian, 39.3% texts that were written in Lithuanian, and 41.4% EU legal texts translated into Baltic languages. Texts originally written in Baltic languages represent the following domains: modern fiction (86%), periodicals (5.9%), popular literature (5.6%). In our experiments only texts that were originally written in Baltic languages were used, as we found noise in the translated legal texts.

## 2.2 Application of Statistical Measures

Different statistical measures are well known means for extraction of MWEs, especially collocations. Among the widely used measures the most popular are the t-score, mi-score, log-likelihood and Dice score (e.g., Manning and Schütze 1999).

In our experiments we applied different combinations of frequency, mi-score and t-score. The mi-score (mutual information score) measures the strength of association (frequency of co-occurrence vs. separate occurrence). A mi-score of three or higher is usually considered to be significant. However, for low frequency words the mi-score could be misleading, as demonstrated in Table 2, where none of the top 10 MWE candidates is an MWE. The t-score measures the confidence of association and can also be applied for low frequency words, with a t-score of two or higher considered to be statistically significant (Hunston 2002), although it also recognizes frequent word combinations (e.g., in Table 2, *kas ir* (which is), *tas ir* (it is)).

Our first experiments were performed on the Balanced Corpus of Modern Latvian. The Collocate tool (Barlow 2004) was used for calculation of statistical measures. At first the mi-score and t-score were used individually and the most frequent MWE candidates were investigated. We then applied the t-score as an initial filter, and afterwards sorted the results by mi-score and frequency. The threshold for the t-score was set at 2.5, while for the mi-score it was set at 3. The higher t-score, as recommended by Hunston (2002), was set to avoid unnecessary noise. Word sequences consisting of two to five words were investigated.

Table 2 summarizes the top 10 word sequences extracted from the Balanced Corpus of Modern Latvian with the mi-score, t-score and combination of both. The word sequences that are bold in the table could be considered as good MWE candidates<sup>1</sup> – some of them (mostly short ones) are already included in existing dictionaries, while others could be added after investigation by a lexicographer.

Table 2: Top 10 word sequences extracted with different statistical measures  
(bolded MWE candidates could be accepted as MWEs).

identified by mi-score		identified by t-score		identified by t-score, filtered by mi-score	
ordered by mi-score	ordered by frequency	ordered by t-score	ordered by frequency	ordered by mi-score	ordered by frequency
“(Caune, Rata, Grigule, Svīklis, Ugaine”	<b>kā arī</b> (also)	kā arī (also)	kā arī (also)	nolikums” (Latvijas Vēstnesis, 168 (3116), 22.10.2004.	<b>stājas spēkā</b> (enter into force)
“Pirts, baseini, vanna, solārijs, sports”	tas ir (it is)	(Ar grozījumiem, kas izdarīti ar (with amendments made by)	tas ir (it is)	nolikums” (Latvijas Vēstnesis, 129 (3705), 10.08.2007.)	<b>kas stājas spēkā</b> (that enters into force)
kurējās uguns vilinot knišļus gaiņājos	<b>stājas spēkā</b> (enter into force)	<b>likuma redakcijā, kas stājas spēkā</b> (the law version that comes into force)	kas ir (it is)	nolikums” (Latvijas Vēstnesis, 76 (3652), 11.05.2007.)	<b>likumu, kas stājas spēkā</b> (the law that enters into force)
aizā kurējās uguns vilinot knišļus	to, ka (the fact that)	ne tikai (not only)	<b>stājas spēkā</b> (enter into force)	nolikums” (Latvijas Vēstnesis, 124 (3072), 06.08.2004.)	likumu, kas stājas (the law that enters)
“(Pranka, Lāce, Trupovniece, et al.”	ar to (with this)	kas ir (it is)	to, ka (the fact that)	nolikums” (Latvijas Vēstnesis, 70 (2835), 13.05.2003.)	<b>Ministru kabineta</b> (Cabinet of Ministers)

<sup>1</sup> In some cases post-processing (removal of delimiters ) is necessary

identified by mi-score		identified by t-score		identified by t-score, filtered by mi-score	
“Lāce, Trupovniece, et al. 2003/”	ne tikai (not only)	tas ir (it is)	ar to (with this)	izdarīti ar 10.06.1998., 25.11.1999., 20.06.2001.,	grozījumiem, kas izdarīti ar (amendments made by)
uguns vilinot knišļus gaiņājot zvērus	par to, (about it)	bet arī (also)	ne tikai (not only)	pensiju shēmas līdzekļu pārvaldītāju reģistrā (register of the pension scheme asset managers)	grozījumiem, kas izdarīti (amendments made)
rullī (2.lasījums. Steidzams) Datums: 09.11.2006.	kas stājas (that enters)	ar to (with this)	par to, (about it)	fondēto pensiju shēmas līdzekļu pārvaldītāja (funded pension scheme asset manager)	kas izdarīti ar (made by)
klusā aizā kurējās uguns vilinot	<b>kas stājas spēkā</b> (that enters into force)	to, ka (the fact that)	<b>kas stājas spēkā</b> (that enters into force)	fondēto pensiju shēmas līdzekļu pārvaldītājs (funded pension scheme asset manager)	(Ar grozījumiem, kas izdarīti ar (with amendments made by)
ugunsgrēks, zibens spēriens, zādzība, vētra	ir ļoti (is very)	tā ir (it is)	ir ļoti (is very)	fondēto pensiju shēmas līdzekļu pārvaldītāju (funded pension scheme asset manager)	(Ar grozījumiem, kas izdarīti (with amendments made)

The table clearly demonstrates the strengths and weaknesses of each approach: when MWE candidates are ordered by statistical significance then longer word sequences are identified (*likuma redakcijā, kas stājas spēkā* – *in the form of law which has effect*), while ordering by frequency identifies short, but stable phrases, e.g., multi-word conjunctions (*kā arī* – *as well as*, *ne tikai* – *not only*).

When MWE candidates were selected and ordered by mi-score the top 10 word sequences were noun phrases or word sequences with a very high (more than 70) mi-score, but none of them was an MWE. In the case of the t-score, both short (*bet arī* – *but also*) and longer (*ar grozījumiem, kas izdarīti ar* – *with amendments that has been made with*) MWE candidates are identified. Although the top 10 MWE candidates identified by t-score include several MWEs, more than half of the identified MWE candidates are frequent word sequences or parts of phrases.

Finally, the t-score was applied as the first filter and then the candidate list was filtered by the mi-score. The Top 10 MWE candidates (ordered by mi-score) include four acceptable MWE candidates (others are typical initial phrases of legal documents). All four MWE candidates are complex noun phrases (terms): three are morphological variants (inflected forms) of the phrase ‘*fondēto pensiju shēmas līdzekļu pārvaldītājs*’ (*manager for funded pension scheme assets*) and the fourth is another term – ‘*pensiju shēmas līdzekļu pārvaldītāju reģistrā*’ (*in a register of funded pension scheme managers*). When this MWE candidate list is sorted by frequency, seven of the top 10 MWE candidates can be accepted as MWEs. These MWEs are either verbal constructions (e.g., *stājas spēkā* – *enter into force*) or nouns followed by a relative clause (e.g., *grozījumiem, kas izdarīti* – *amendments made*).



These three initial experiments demonstrated that in the case of a small corpus the most promising approach uses a combination of t-score and mi-score.

Latvian is a morphologically rich language, and thus application of statistical measures on a small corpus allows us to find only frequent phrases that in many cases are already in dictionaries (e.g. multi-word conjunctions, *kā arī – as well as*). To delve further and obtain not so trivial (although useful) data, we applied a Latvian lemmatizer (Paikens et al. 2013) and repeated the same set of experiments with lemmatized data. This allowed us to find more MWEs – we found many named entities (people's names and their occupations, as well as the related organization names) and terminological units from different domains. Table 3 shows the top 10 MWE candidates that were identified with a t-score and then filtered with the mi-score. Four named entities and five terms are among top 10 MWE candidates in the MWE candidate list that is ordered by mi-score. When the list is ordered by frequency, three complex function words (*kā arī – also*, *kaut kas – something*, *pēc tas – after*) and two frequent MWEs (*pants punkts – article* and already mentioned *stāties spēkā – enter into force*) are included.

Table 3: MWE candidates extracted from the lemmatized corpus (MWE candidates in bold could be accepted as MWEs).

Word sequences with highest mi-score	Most frequent word sequences
Arco Real Estate ‘ ‘ (company name)	kā arī (also)
<b>pārvalde priekšnieks palīdz Linda Zubāne</b> (assistant chief of administration Linda Zubāne)	pants punkts (article)
<b>Černobiļa AES avārija sekas likvidēšana</b> (Chernobyl nuclear plant disaster recovery)	, kas būt (which is)
šķirne ‘ Koričnoje Novoje ‘ (named entity)	tas , ka (the fact that)
<b>jaukt dispersija kovariāt analīze iegūt</b> (mixed variance covariance analysis provides)	<b>kaut kas</b> (something)
<b>ar akūts katarāli strutot endometrits</b> (with acute catarrhal stomach endometritis)	tas , kas (that/what ...)
<b>ar hronisks katarāli strutot endometrits</b> (with chronic catarrhal stomach endometritis)	būt ļoti (to be very)
<b>pārvalde priekšnieks palīdz Ieva Sietniece</b> (assistant chief of administration Ieva Sietniece)	viens no (one of)
Valmiera / Rūjiena / Strenči-1 (list of names)	<b>stāties spēks</b> (enter into force)
līcis piekraste krasts kāpa aizsargjosla (coastal protection zone)	pēc tas (after)

### 2.3 Filtering MWE Candidates

The identified word sequences that are extracted using statistical measures are not always grammatical, as demonstrated in Tables 2 and 3. Moreover, the list of MWE candidates contains word sequences that are not MWEs (e.g., phrases or word sequences that are frequent in a particular corpus), and thus need to be removed from the list. Therefore, after selection of initial MWE candidates, statistical and morpho-syntactic filters are used for the final selection of MWE candidates.

Statistical filters are used to avoid unnecessary noise that is typical for MWEs with a low confidence score. In the case of the mi-score, we found that a high frequency (and mi-score in a range of four to 11) is a better signal that the string could be an MWE than a high mi-score and low frequency (e.g. below 10). In the case of the t-score – high frequency together with a high t-score is a signal of a good MWE candidate. Finally, if the t-score is used as the initial filter and mi-score is used as the second filter, then: (1) most of the MWE candidates will be frequent and have a mi-score value between 10 and 35, or, (2) will have a high mi-score and low frequency.

The simple regular expressions and morpho-syntactic filters allow to filter out word sequences that are ungrammatical. Regular expressions are used to filter out sequences of tokens that start or end with a punctuation mark, include parentheses or numbers. For instance the word sequence ‘*par to ,*

(*about*.) from Table 2 ends with a comma and thus needs to be removed or replaced with ‘*par to*’. Language specific regular expressions include words that in a specific position makes an MWE candidate ungrammatical, e.g., *un* (*and*), *vai* (*or*) as the last word, or, *būt* (*to be*) at the beginning or end of a string consisting of two words (e.g. *būt ļoti* (*to be very*) in Table 2).

Morpho-syntactic filters are used to filter ungrammatical MWE candidates, as well as to extract specific categories of MWEs, e.g., verbal phrases. The most complicated case is an ungrammatical sequence that contains parts of two or more phrases (e.g., in Table 2: *kurējās uguns vilinot knišķus gaiņājos* – *fire burned luring flies fight*) or contains only part of the phrase (e.g., in Table 3: *ir ļoti* – *is very*). In such cases the process of filtering patterns needs to be defined carefully, to avoid situations when good MWEs are removed. For instance, the verbal phrase *stājas spēkā* (*comes into force*) could be mistakenly removed, as it contains a verb followed by noun in the locative form.

Finally, in the case of overlapping MWE candidates (e.g. *stājas spēkā* (*comes into force*), *stājas spēkā ar* (*comes into force from*), or *kas stājas spēkā* (*which comes into force*)) the choice of the most appropriate MWE needs to be made by a lexicographer.

### 3 Application and Results

We evaluated our method on three different Latvian language corpora: the Balanced Corpus of Modern Latvian, Latvian-Lithuanian corpus and Open Subtitles corpus. The choice of these corpora was justified by the aim of this research – to provide good MWE candidates for a Latvian explanatory dictionary. Therefore, we excluded well-known domain specific corpora, such as JRC Acquis or EMEA, because the term extraction problem (as a special category of MWEs) has been researched by Pinnis et al. (2012).

#### 3.1 Balanced Corpus of the Modern Latvian Language

The Balanced Corpus of the Modern Latvian language was the starting point and the main resource of our research. This corpus (and its updated versions that are under construction) is the main resource on which other language resources (such as the universal dependency treebank, FrameNet and PropBank for Latvian) are currently created.

Our initial hypothesis was that this corpus is a good source to identify different types of MWEs that occur in Latvian rather frequently. However, as was demonstrated in the previous section, we found that applying simple statistical measures to this corpus allows us to identify good MWE candidates for the legal domain (e.g., *stājas spēkā* – *comes into force*). The main reason is the rather strict language of legal texts: although legal documents form only 5% of the corpus texts, typical legal domain phrases, that appear again and again, are identified as legal domain terminology entries in our MWE candidate list.

As was demonstrated in the previous section, in the case of a lemmatized corpus our method allows us to find many named entities and terminological units from different domains. Most of the identified MWEs consist of two or three words, and thus in the next experiment we identified strings of words up to three words long – these strings were identified by mi-score or t-score and then filtered by the former. The list of the top 10 MWE candidates is shown in Table 4. When MWEs are identified by mi-score, all the top 10 word sequences are MWEs. However, most of MWE candidates are named entities, the only exception is ‘*Pīrsons hī kvadrāts*’ (*Pearson’s chi-square*). In the case when MWEs are identified by t-score, five strings are named entities, four are terms and one (JP NVO RV) is a string of characters. When the top 20 MWE candidates were analyzed, eight of them were terms.

Table 4: Top 10 lemmatized MWE candidates selected with t-score and mi-score (MWEs in bold are terms, while others are named entities, except JP NVO RV).

mi-score	t-score
Legacy by Angosturs (named entity)	Arco Real Estate (company name)
Eastgate Properties Limited (company name)	Satja SAI Baba (company name)
Nike Riga Run (named entity – event)	<b>Pirsons hī kvadrāts</b> (Pearson's chi-square)
ģenerāldirektors Jespers Koldings (general director Jesper Kolding)	katarāli strutot endometrīts (catarrulous endometritis)
Arco Real Estate (company name)	JP NVO RV
fon den Brinkena (name)	<b>amonijs nitrāts slāpeklis</b> (ammonium nitrate nitrogen)
Satja SAI Baba (company name)	Ge Money Bank (named entity – bank)
Satja Sai Baba(company name)	Parex Asset Management (named entity)
Latvian Art Theory (named entity)	New York Time (named entity)
<b>Pirsons hī kvadrāts</b> (Pearson's chi-square)	jaukt dispersija kovariāt (mixed variance covariance)

This experiment shows that t-score allows better to identify terms that can be included into electronic dictionary. Therefore the threshold for t-score was raised up to 10: 8 terms, one named entity (LPP/LC – name of party) and one sequence of words ( $^{\circ}$  C *temperatūra*) was identified between top 10 candidates (Figure 1).

Frequency	Mi-score	MWE candidate
33	27.988560	ģenētiski modificēt kultūraugi (genetically modified crops)
34	27.120896	ģenētiski modificēt mikroorganismus (genetically modified microorganism)
42	26.717372	noziedzīgs nodarījums izdarīšana (committing criminal offences)
112	26.346762	LPP / LC (name of party)
137	26.146896	ģenētiski modificēt organismus (genetically modified organism)
168	25.801889	fondēta pensija shēma (funded pension schema)
9	25.669791	konkurētspējīga priekšrocība pārvešana (competitive advantage transfer)
8	25.222240	civila aizsardzība aizsargbūve (civil defence protection structure)
37	25.169589	$^{\circ}$ C temperatūra ( $^{\circ}$ C temperature)
31	25.112838	infekcija slimība izraisītājs (infectious disease agent)

Figure 1: Frequency, mi-score for top 10 MWE candidates identified by t-score $\geq$ 10.

As the project is organized around the top 2,000 Latvian verbs, in our next experiment, after the application of statistical measures to the lemmatized corpus, we filtered out only MWE candidates that contain a noun and verb in a person form (Figure 2). From the top 10 MWE candidates, seven could be accepted as MWEs: four of these are included in *tezaurs.lv*, while other three are included in the Latvian-English dictionary (Veisbergs 2005). It has to be mentioned that three of the four MWEs that are present at *tezaurs.lv* are formed by a verb in a person form followed by a noun in the locative form.

Frequency	Mi-score	MWE candidate
2006	44.740771	<b>stāties V spēks N</b> (come into force)
623	24.896466	<b>pieņemt V lēmums N</b> (to make decision / decide)
290	16.915062	<b>dot V iespēja N</b> (to enable)
247	15.360174	<b>tikt V gals N</b> (to manage)
218	14.595137	veikt V pētījums N (to do research)
141	11.759475	<b>sniegt V informācija N</b> (provide information)
130	11.393867	<b>pievērst V uzmanība N</b> (pay attention)
115	10.505124	tiesības N saņemt V (rights to receive)
104	10.187839	<b>ienākt V prāts N</b> (to come into one's head)
92	9.581201	aizvērt V acs N (to close eyes)

Figure 2: Frequency and mi-score for the top 10 verbal phrases consisting of verb (V) and noun (N), with the MWEs in bold.

We can conclude that the Balanced Corpus of Modern Latvian is a good source for automatic identification of named entities (people's names and their occupations, as well as the related organization names) and terminological units from different domains. The increase in the threshold allows us to obtain good terminological entries with high precision. When a specific phrase pattern is considered, the result depends on that particular phrase's construction and the quality of the pattern. We can also see that the size and balancing criteria of this corpus limits the ability of automatic methods with regard to finding idiomatic expressions.

### 3.2 Latvian-Lithuanian Corpus

To investigate the applicability of our methods for identification of other types of MWEs, not only named entities and terms, we applied the same strategy to the Latvian part of the Latvian-Lithuanian parallel corpus. Although it is also a rather small corpus, it contains more general domain texts than the Balanced Corpus of Modern Latvian, including modern fiction and news texts. Our hypothesis was that such a corpus could contain more frequently used fixed phrases and idiomatic expressions than the Balanced Corpus of Modern Latvian. However, as shown in Table 5, the results obtained are similar to the previous ones. When MWE candidates are ordered by mi-score, seven MWE candidates are named entities, one is part of a longer phrase (*Ventspils peldbaseina relaksācija – Ventspils swimming pool of relaxation*), one is a fixed phrase (*peldbaseins relaksācija komplekss – complex of swimming pools for relaxation*) and one is a character string (W/m). If MWE candidates are ordered by frequency, then four MWEs are terms (*apkure katls – central heating boiler, apkure iekārta – central heating boiler, sāls istaba – salt room, relaksācijas komplekss – complex of relaxation*), one is a named entity (*Ventspils peldbaseins – Ventspils swimming pool*), one is a complex function word (*ne tikai – not only*), while four other MWE candidates are parts of longer phrases.

Table 5: Top 10 MWE candidates extracted from the Latvian-Lithuanian parallel corpus and ordered by mi-score and frequency (MWEs are in bold).

mi-score	Frequency
SIA “AD BALTIC” (company)	<b>apkure katls</b> (central heating boiler)
Ventspils peldbaseins relaksācija komplekss	apkure iekārta (heating system)
izstāde “Tech Industry” (event)	<b>Ventspils peldbaseins</b> (Ventspils swimming pool)
“Tech Industry” (event)	koksne granula (wooden pellet)
“AD BALTIC” (named entity)	katls iekārta (boiler equipment)
SEALEY POWER products (named entity)	granula apkure (pellet heating)
W / m	informācija par (information about)
peldbaseins relaksācija komplekss (swimming pool relaxation complex)	sāls istaba (salt room)
Ventspils peldbaseins relaksācija (Ventspils swimming pool relaxation)	relaksācija komplekss (relaxation complex)
REN TV Baltija (named entity)	<b>ne tikai</b> (not only)

In contrast to the previous experiment, the Latvian-Lithuanian parallel corpus was too small to obtain good terminological units when a higher threshold for the t-score was set. Our hypothesis that we could identify idiomatic expressions in this corpus was thus not supported, and perhaps idiomatic expressions are quite rare in this collection.

### 3.3 Open Subtitles Corpus

As idiomatic expressions were not found in two previously used corpora, we turned to the last on our list – the Open Subtitles corpus. As in the previous experiments, we used a lemmatized corpus and

applied the t-score for identification and mi-score as the second filter. In addition, MWE candidates were filtered by frequency (in the previous experiments a threshold of only five was used). The results of these experiments are summarized in Table 6, and they differ from those of the earlier ones – besides named entities, different idiomatic expressions are also identified.

Similar to the previous experiments, many (four when the frequency is at least five and six if it is at least 10 or 15) of the extracted MWEs are named entities (e.g., *viesnīca “dižena Budapešta”* - hotel “great Budapest”, *Bārts Šērmens* – Bart Shermen, “*zēns un ābols*” - “boy and an apple (painting)). However, different idiomatic expressions are identified too (e.g., *dzīvot laimīgi līdz mūžs gals* – live happily to the end of his days, *gulēt saldi* – sleep well, *daudz laimes dzimšanas diena* – happy birthday; *ar tas nebūt nekāds sakars* – nothing to do with this,).

Table 6: List of MWE candidates extracted from the Open Subtitles corpus, with the idiomatic expressions in bold.

Freq>=5	Freq>=10	Freq>=15
it ‘s not going	it ‘s not going	Bārts Šērmens (named entity)
viesnīca “dižena Budapešta” (hotel “Great Budapest”)	Bārts Šērmens (named entity)	“Pearson Hardman
<b>dzīvot laimīgi līdz mūžs gals</b> (live happily till end of life)	dzeršana no zābaks (drink from boot)	“Folsom foods”
misis boss (named entity)	paskriet , paostīt , sarauties (run, sniff, cringe)	“SouthJet” 227
Bārts Šērmens (named entity)	Vašingtona māksla noziegums nodaļa (Washington Arts Crime Division)	“Delta psi” (named entity)
Rikijs Pontings (named entity)	laimīgi līdz mūžs gals (happily till end of life)	“mežonīgs vepris” (“wild hog” - name of bar)
dzeršana no zābaks (drink from boot)	“SouthJet” 227	pakārt viņš (hang him)
paskriet , paostīt , sarauties (run, sniff, cringe)	“zēns ar ābols” (“Boy with Apple” – painting)	daudz laime dzimšana diena (happy birthday)
<b>gulēt saldi</b> (sleep well)	“Wayne enterprise”	ar tas nebūt nekāds sakars (nothing to do with it)
kosmos kuģis (spaceship)	“Pearson Hardman”	dzeršana no zābaks (drink from boot)

## 4 Conclusion

In this paper we discussed possible strategies for extraction of MWE candidates from different corpora – a balanced, parallel corpus that contains mainly fiction and a corpus of a specific genre (subtitles). We demonstrated that in case of a small amount of general domain (balanced) data, automatic methods can be used to find good MWE candidates – terms or named entities. However, finding idiomatic expressions in small, general domain corpora is looking for a needle in a haystack: only a larger or more expressive corpus could help in the identification process.

In the case of a small parallel corpus, the most reliable results are obtained for named entities. Terms and complex function words could be also identified, but in this case more careful manual inspection is necessary. Therefore, our next task is to investigate the possibility of applying an automatically extracted bilingual dictionary as an additional filter to improve the precision of MWE candidates.

If the aim of the MWE identification is to identify idiomatic expressions that have recently appeared in a language, then the corpus needs to represent more everyday language and to be rather large, because idiomatic expressions are rare in balanced corpora that represent literary language and carefully edited texts.



## References

- Baldwin, T. and Kim, S.N. (2010). Multiword Expressions. In *Handbook of Natural Language Processing*, pp. 267-292.
- Barlow, M. (2004). Collocate 1.0: Locating collocations and terminology.
- Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, (2017). A.: Multiword expression processing: a survey. In *Computational Linguistics*, 43(4), pp. 837-892.
- Deksne, D., Skadins, R. & Skadina, I. (2008). Dictionary of Multiword Expressions for Translation into Highly Inflected Languages. In: *Proceedings of the International Conference on Language Resources and Evaluation LREC 2008*, pp. 1401-1405.
- Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., Paikens, P. (2018). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 4506-4513.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, pp. 923-929.
- Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge.
- Paikens, P., Rituma, L. & Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*.
- Levāne-Petrova K. (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpus un tā tekstu atlases kritēriji. In *Baltistica VIII priedas*, Vilnius, pp. 89-98.
- Pinnis, M. & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In *Proceedings of the Fifth International Conference Baltic HLT 2012*, pp. 176-184.
- Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Recent Advances in Natural Language Processing*, pp. 562-570.
- Pinnis, M., Ljubešić, N., Ștefănescu, D., Skadiņa, I., Tadić, M. & Gornostay, T. (2012). Term Extraction, Tagging and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*.
- Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. Theory and Applications of Natural Language Processing series XIV, Springer.
- Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: a pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'02*, pp. 1-15.
- Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegård, G., Parra, C., Waszczuk, J., Constant, M., Osenova, P., Sangati, F. (2015) “PARSEME – PARSing and Multiword Expressions within a European multilingual network”. In *Proceedings of the 7th Language & Technology Conference (LTC 2015)*, pp. 27-29.
- Skadiņa, I., Veisbergs, A., Vasiljevs, A., Gornostaja, T., Keiša, I. & Rudzīte, A. (2012). *Latvian Language in the Digital Age*. Springer.
- Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L. & Saulite, B. (2016). Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pp. 2568-2571.
- Utkā, A., Levane-Petrova, K., Bielskiene, A., Kovalevskaite, J., Rimkute, E. and Vevere, D. (2012). Lithuanian-Latvian-Lithuanian parallel corpus. In *Proceedings of the Fifth International Conference Baltic HLT 2012*, pp. 260-264.
- Veisbergs, A. (2005). *Jaunā latviešu-angļu vārdnīca*. Zvaigzne ABC.

## Acknowledgements

This work was supported by the European Regional Development Fund grant No. 1.1.1.1/16/A/219 “Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian”.



# **Terminology, Terminography and Specialised Lexicography**



# Semantic-based Retrieval of Complex Nominals in Terminographic Resources

**Melania Cabezas-García, Juan Carlos Gil-Berrozpe**

University of Granada

E-mail: [melaniacabezas@ugr.es](mailto:melaniacabezas@ugr.es), [jcgilberrozpe@ugr.es](mailto:jcgilberrozpe@ugr.es)

## Abstract

In English, specialized concepts frequently take the form of complex nominals (CNs), e.g. *greenhouse gas emissions*. The syntactic-semantic complexity of these multi-word terms (MWTs) highlights the need for a systematic treatment in specialized resources. This paper explores how semantic patterns in CNs can be applied to retrieve information in terminological knowledge bases, specifically in EcoLexicon (<http://ecolexicon.ugr.es>), the practical application of Frame-based Terminology (Faber 2012). For that purpose, we extracted the 250 most frequent CNs in an English wind power corpus. Structural disambiguation was performed to identify the internal groups linked by semantic relations. *Ad-hoc* semantic categories were then assigned to the elements of CNs with a view to studying the formation of CNs and allowing semantic-based queries in EcoLexicon. Then, the semantic relations between the CN constituents were analyzed by means of knowledge patterns and paraphrases. Our preliminary results showed recurrent semantic patterns in CN formation. This facilitates the inference of semantic relations, which is one of the main difficulties of MWTs. Furthermore, a semantic-based view of the CN module of EcoLexicon is presented, which allows different types of semantic query.

**Keywords:** complex nominals, semantic patterns, semantic categories, terminological knowledge bases

## 1 Introduction

In English, specialized concepts frequently take the form of complex nominals (CNs), e.g. *greenhouse gas emissions*. These multi-word terms (MWTs) are characterized by their syntactic and semantic complexity, which underlines the need for their systematic treatment in lexicographic and terminographic resources (Cabezas-García and Faber 2017a). Before describing CNs, their meaning must be specified, usually with a set of semantic relations (e.g. in *wind erosion*, wind *causes* erosion) (Rosario et al. 2002; Girju et al. 2005; *inter alia*). Nakov and Hearst (2006) propose the use of verb paraphrases (e.g. *olive oil* is an *oil that comes from/is squeezed from olives*). This indicates that semantic patterns in CN formation (Maguire et al. 2010) should be addressed before including CNs in specialized resources, e.g. *beach erosion* represents a LANDFORM (*beach*) that is the *patient\_of* a LOSS PROCESS (*erosion*). These semantic patterns are closely related to the underlying conceptualization of the domain and the semantic relations in CNs, whose non-specification usually poses comprehension problems.

This paper explores how semantic patterns in CNs can be applied to retrieve information in terminological knowledge bases. EcoLexicon (<http://ecolexicon.ugr.es>) (Faber et al. 2016) is a terminological knowledge base on environmental science, which is currently being redefined to describe CNs. The design of its new phraseological module focuses on the inclusion of different kinds of data regarding CNs, such as the combinations derived from a given term, definitions, translations in English and Spanish, syntactic combinations and semantic co-occurrences. For this purpose, we extracted the 250 most frequent CNs in an English wind power corpus. Structural disambiguation was performed to identify the internal groups between which semantic relations had to be established. *Ad-hoc* semantic categories (e.g.



LANDFORM, WATER BODY) were then assigned to the internal groups in CNs with a view to conceptually analyzing the formation of CNs and allowing semantic-based queries in EcoLexicon. The assignment of semantic categories also facilitated the identification of the semantic relations between CN constituents with paraphrases and knowledge patterns (Meyer 2001; Marshman 2006). Thus, this search allowed users to perform queries based on terms, semantic categories, and semantic relations present in CNs.

The rest of this paper is organized as follows. Section 2 provides a review of the most relevant theoretical aspects concerning CNs, as well as a description of how they are usually dealt with in various terminographic resources. Section 3 explains the methodology applied to this study, focusing on the extraction of CNs and their semantic analysis. Section 4 shows the results of this semantic analysis and displays the most relevant semantic patterns in our wind power corpus. In Section 5 we describe the semantic-based view of the CN module of EcoLexicon. Finally, Section 6 gives the conclusions that can be derived from this research.

## 2 Complex Nominals in Linguistic Resources

### 2.1 Complex Nominals and their Semantics

Complex nominals (CNs) are expressions with a head noun modified by one or more elements, usually nouns or adjectives (Levi 1978), e.g. *significant wave height*. CNs reflect the preference of Germanic languages for condensed structures in which economy of expression overcomes clarity of expression (Štekauer et al. 2012). In particular, stacking modifiers on the left of the head (i.e. pre-modification) is the most frequent formation pattern in English (e.g. *water vapor*), although prepositional modification of the head (i.e. post-modification) can also be found, often combined with premodifiers (e.g. *general circulation of the atmosphere*). This permits the expression of specialized knowledge in semantically condensed structures (Sager et al. 1980; Sanz-Vicente 2012).

According to Lauer and Dras (1994), CNs can pose problems, namely insofar as their identification in texts, and their syntactic and semantic analysis. First, identifying them in texts can be challenging since they are often formed by general language words that may not be recognized as part of the CN. Furthermore, CNs are composed of more than two constituents, which also entails the need to disambiguate their internal structure (e.g. *renewable [energy source]*). Accessing the semantics of CNs is not an easy task, since these MWTs convey a relation between two or more elements that requires further knowledge (Maguire et al. 2010; Smith et al. 2014). Additional difficulties can arise when translating CNs, since term formation has different patterns in different languages (e.g. the pre-modification in English is not possible in Romance languages). Finally, the representation of CNs in linguistic resources has also been the subject of debate since they should be treated more systematically (Cabezas-García and Faber 2017a).

The semantic analysis and representation of CNs are the focus of this study. The semantics of CNs can be analyzed in terms of predicate nominalization (e.g. a system *transmits* AC > AC *transmission system*) and predicate deletion (an industry *produces* wind power > *wind power industry*) (Levi 1978). These concealed predicates largely correspond to semantic relations (e.g. AC *transmission system* > a system *has\_function* [transmitting] AC; and *wind power industry* > an industry *has\_function* [producing] wind power). Semantic relations are essential for the conceptualization of CNs, because they are part of the micro-context of these MWTs. The semantics of the head of a CN determines the conceptual nature of its modifiers (e.g. *energy* is usually referred to in terms of the resource that is used for its production, as in *wave energy*, *solar energy*, *wind energy*, etc.). In this micro-context, the internal semantic relations in CNs are relevant because they show how the elements of the micro-context are linked (e.g. *energy caused\_by* waves/sun/wind).

## 2.2 Representation of Complex Nominals

As for the representation of CNs, a major problem is that they are rarely defined (e.g. *Vocabulaire et cooccurents de la comptabilité* [Caignon 2001]). Furthermore, they are often listed alphabetically (e.g. *Dictionary of Energy* [Cleveland & Morris 2015]; *A Dictionary of Translation Technology* [Chan 2004]). However, a representation of domain structure should reflect the relations of the CN with other terms (e.g. *Elsevier's Dictionary of Medicine Spanish-English English-Spanish* [Hidalgo 2014]). There are also resources that show the modifiers and their possible heads in different lines, instead of including the entire CN (e.g. *Diccionario técnico inglés-español español-inglés* [Beigbeder 2006]). Other resources display a modifier along with a list of different and not necessarily related CNs that contain the modifier (e.g. *Routledge French Technical Dictionary* [Arden 2013]), such as *complementary angle*, *complementary code* and *complementary color* (Arden 2013: 141).

For rapid knowledge acquisition (Faber 2012), a conceptual approach seems to be best, since it reflects domain structure, facilitates understanding, and provides the basis for translation (Cabezas-García & Faber 2017a). An increasing number of resources take a semantic approach (e.g. FrameNet [Baker et al. 1998]; WordNet [Fellbaum 2005]; VerbNet [Kipper-Schuler et al. 2006]; DiCouèbe [OLST 2013]; DicoInfo [OLST 2018]; DicoEnviro [OLST 2018]; *inter alia*). However, the conceptual information of units is often based on semantic roles (e.g. AGENT, PATIENT), while other relevant information is not recorded. This is the case of semantic categories (e.g. SUBSTANCE, LANDFORM), which permit generalizations about concepts and conceptual organization as well as semantic-based queries. Resources that allow such queries include the *Pattern Dictionary of English Verbs* (Hanks 2014) where a list of semantic categories can be queried to obtain the verbs with which they co-occur, or vice versa (e.g. the COLOR category establishes verbs such as *shade*, *paint* or *dye*). The *DELAC Dictionary of Serbian Compounds* (Krstev et al. 2006) also allows the retrieval of compounds based on semantic categories (e.g. the search for the +Zool category shows Serbian compounds including an animal). Furthermore, in the new version of the *Diccionario de términos médicos* of the Spanish *Real Academia Nacional de Medicina* (2012) users can search for terms included in definitions (e.g. if *piel* [skin] is queried, entries including *skin* in their definitions will be shown). Additionally, one of the views of the *Diccionario Ideológico de la lengua española* (Casares 2013) organizes concepts in semantic categories (e.g. the MOLLUSK category includes *clam*, *oyster*, *squid*, etc.). Therefore, frame-like representations (Fillmore 1985) (e.g. EcoLexicon) are a good option since they combine conceptual and linguistic representations (L'Homme 2014).

## 3 Complex Nominal Extraction and Semantic Analysis

For our study, we compiled a corpus of English specialized texts on wind power, composed of approximately 1.8 million words. The corpus, which consisted of specialized articles and PhD dissertations, was uploaded to Sketch Engine (<http://www.sketchengine.co.uk>) (Kilgariff et al. 2004), a corpus analysis tool that is used for CN extraction and semantic analysis. The different forms of English CNs were obtained with Corpus Query Language (CQL) regular expressions. Two lists of 1,000 CNs were extracted. One was composed of CNs formed by pre-modification<sup>1</sup> of the head by nouns, adjectives and/or adverbs (e.g. *tip speed*), and a second list composed of CNs formed by post-modification<sup>2</sup> of the head by a prepositional phrase (e.g. *angle of attack*). Both lists were combined and the 250 most frequent CNs were selected. All CNs that were part of a longer CN or which belonged to other domains such as Statistics, Mathematics, or Economics, were discarded.

1 The regular expression used for such query was the following one: [tag="N.\*|JJ.\*|RB.\*"]{1,}[tag="N.\*"].

2 The regular expression used for such query was the following one: [tag="N.\*|JJ.\*|RB.\*"]{0,}[tag="N.\*"]{1,}[tag="IN"]{1}[tag="DT"]?[tag="JJ.\*|RB.\*"]{0,}[tag="N.\*"]{1,}.

The first step in semantic analysis was the internal structure disambiguation of CNs with more than two constituents (e.g. [*wind power*] [*generation system*]). Indicators of the internal groupings in CNs are the conceptual nature of the possible combination (e.g. *wind power* is a concept in linguistic resources), the existence of monolexical variants or equivalents in another language (e.g. *generation system* is also referred to as *generator*), or the co-occurrence of this CN with different modifiers (e.g. *conventional generation system*, *diesel generation system*, or *electric generation system*).

Conceptual or semantic categories were then used to specify the semantics of CNs, taking into account different categorization levels (Murphy & Lassaline 1997) and conceptual similarity (Hahn & Chater 1997). This categorization was based on concept definitions as well as on the contextual information in the wind power corpus. After determining the characteristics shared by concepts, categories were manually established and organized hierarchically from general to specific. In this way, the 250 CNs were classified in 47 conceptual categories, distributed in four categorization levels. The most general level was composed of the three starter ontological categories: ENTITY (i.e. physical and mental objects), PROCESS (i.e. events extending over time and involving different parties), and ATTRIBUTE (i.e. characteristics of entities or processes). The CNs were then classified in one of four more specific levels. For example, CNs with a generic meaning could only be categorized in one level (e.g. *energy source* [ENTITY]), whereas CNs with a specific meaning could be categorized in all four levels (e.g. *carbon dioxide* [ENTITY > MATTER > SUBSTANCE > CHEMICAL SUBSTANCE]). However, for the sake of simplifying such categorization, we only refer to the most specific category (e.g. *carbon dioxide* [CHEMICAL SUBSTANCE]). Furthermore, this categorization was applied to the CNs in three stages: (1) to the whole CN; (2) to its internal groupings; (3) to its individual constituents. For instance, *offshore wind turbine* as a whole was classified as an INSTRUMENT. Semantic categories were then assigned to the internal groupings in the CN (*offshore* [LOCATION] *wind turbine* [INSTRUMENT]), as well as to each of its constituent parts (*offshore* [LOCATION] *wind* [WIND MOVEMENT] *turbine* [INSTRUMENT]).

The next step involved specifying the semantic relation between CN constituents. Knowledge patterns (KPs) were used to extract the internal relations in CNs in the form of KP-based sketch grammars (León-Araúz et al. 2016). Figure 1 shows an extract of the results of a query that targets the sentences annotated as word sketches between *power* and *plant*, where *ws* means word sketch, “power-n” and “plant-n” are the terms that have been annotated as part of a word sketch in the corpus; and “\”%w\”.\*” means any relation defined in the KP-based sketch grammars. As can be seen, these KPs show that the function of plants is power generation.

fresh water use. Conventional *plants* generate *power* from fossil fuels and nuclear materials, which the “representative” wind *plant* produces zero *power* approximately 2000 h/y, and full power about the wind source. In fact, as far as the amount of *power* produced by the wind *plant* is small in . In particular, the higher the fraction of *power* produced by a power *plant* in comparison with the following formula, where: P is the total active *power* generated by the wind power *plant* ; Q is the total by the wind power plant; Q is the total reactive *power* generated by the wind power *plant* ; r is the not be controlled, at the rated frequency, the *power* produced by a wind power *plant* can be and availability of the *plant* to generate *power* that is expected at a particular period. The in large, central power *plants* produce *power* at high voltage (up to 25,000 V). These

Figure 1: KPs between *power* and *plant* obtained with the following query:  
[ws(“power-n”, “\”%w\”.\*”, “plant-n”)].

The analysis of semantic relations with KPs was complemented by searching for verb paraphrases in order to access the concealed or nominalized predicates in CNs (Levi 1978; Nakov & Hearst 2006, 2013), which add further semantic precision to semantic relations (Nakov & Hearst 2013; Cabezas-García & Faber 2017b). For instance, *power curve* was found to designate a curve that represents, calculates, simulates or provides the output power of a wind turbine, as shown in Figure 2.



ble to the rated speed. By considering the **curve it is possible to obtain the desired power**, which, limited in the upper part by the  
 The rated power is 3.6 MW and the power **curve, which provides an indicator of the power** as a function of the average wind speed a  
 Wind turbine manufacturers provide power **curves representing turbine power** output as a function of wind speed (see C  
 respectively. In Figures 3 and 4, the green **curve represents practical output power**; the blue curve is for the prediction resu  
 and Zender 2009). Vestas' published power **curve was used to calculate the electric power** output of a single wind turbine. The smoc  
 11.25 ; B2 =1.20). Wind **power is finally obtained by use of a variable speed classical power curve** for each park. In order  
 ilar wind speeds. Wind **power is simulated from the historical wind speed data using a turbine power curve** based on the Vestas V11  
 is: where  $P(U_i)$  is the **power output defined by a wind machine power curve**. 5) The energy from a  
 lthousen 1994), and the **power output was calculated with an interpolated power curve** of the V90 turbine. The

Figure 2: Verb paraphrases for *power curve* obtained with the following queries: [lemma="power"][]{0,10}  
 [tag="V.\*"][]{0,10}[lemma="curve"]within <s/>  
 [lemma="curve"][]{0,10}[tag="V.\*"][]{0,10}[lemma="power"]within <s/>.

However, verb paraphrases are not always easy to find. One reason for this is that CNs usually have concealed elements, whose omission complicates the extraction of verb paraphrases. Instead, many CNs include adjectives<sup>3</sup> or other units that refer to the hidden noun and thus are not easily linked to the head by means of predicates. For instance, in *renewable generation*, no verb joining *renewable* and *generation* was found. Therefore, we used free paraphrases, i.e. co-occurrences of the constituents of a CN in a sentence, as a way of accessing the meaning of those CNs whose semantics had not been ascertained by means of KPs or verb paraphrases. As shown in Figure 3, there was a missing element in *renewable generation* that complicated the extraction of KPs and verb paraphrases, namely the sources from which power is generated.

a company's federal income tax based on the	<b>generation of electricity with renewable resources</b>	, such as wind. As discussed in the following
a specified percentage of the total power	<b>generation from renewable sources</b>	within a certain date. In order to meet the RPS, a
goal is to achieve 30% of its electrical power	<b>generation from renewable sources</b>	by 2020 with a long term goal of 50% by 2050 [30].
up or demonstrated large, small and micro power	<b>generation systems using exploitable renewable resources</b>	. In Bangladesh, to establish the use of
of the CVs constitutes the incentive for the	<b>generation of electric power from renewable energy</b>	sources, except for photovoltaics, for which
. 1. Introduction. During the last years, the	<b>generation of electric power from renewable energy</b>	sources has increased potentially to reduce
Decree on the promotion of electricity	<b>generation from renewable energy</b>	source provides additional incentives, inter
priority connection of installations for the	<b>generation of electricity from renewable energies</b>	and from mine gas to the general electricity
with renewables is the variable and uncertain	<b>generation of electric power from renewable energy</b>	resources. This dissertation focuses on the

Figure 3: Free paraphrases for *renewable generation* obtained with the following query:  
 [lemma="generation"][]{0,10}[word="renewable"][]{0,10}[lemma!="generation"]within <s/>.

In summary, the non-specification of the semantic relation in CNs was addressed by combining different procedures – namely KPs, verb paraphrases and free paraphrases – which offered further semantic insights into these MWTs.

## 4 Semantic Patterns in Complex Nominal Formation

In CNs, two or more concepts converge. However, these combinations are not random, but are the result of semantic constraints (Štekauer 1998). In CNs, these semantic constraints take the form of micro-contexts, which refer to the opening of slots by the head (Maguire et al. 2010). These slots are filled by semantic categories and roles. The semantics of the head thus determines possible modifiers (Rosario et al. 2002). For instance, *emission* opens two slots. One of them is usually filled by the semantic category of SUBSTANCE, which has the role of PATIENT or the entity being emitted (as in *greenhouse gas emissions* and *CO<sub>2</sub> emission*). The second slot normally represents the category of SPACE, which has the role of DESTINATION or place where substances are emitted (as in *air emission* and *atmospheric emission*). The combination of both slots is also possible and produces longer CNs, such as *atmospheric emission of CO<sub>2</sub>*.

<sup>3</sup> Some adjectives have an underlying noun. In that case, they can be replaced by this noun, which facilitates the retrieval of verb paraphrases. For instance, in *fluvial sediment*, verbs linking *river* and *sediment* can be queried (e.g. *deposit*, *carry*, *transport*, etc.).

This study focused on semantic categories in CN formation because they capture regularities and allow generalizations (Hoey 2005). The CNs in our study mostly designated the categories of QUANTITY (28 out of 250 CNs, e.g. *net load*), MAGNITUDE (21 CNs, e.g. *wind speed*), SYSTEM (19 CNs, e.g. *distribution system*), ACTIVITY (19 CNs, e.g. *wind project*), ENERGY (18 CNs, e.g. *renewable energy*), and INSTRUMENT (16 CNs, e.g. *offshore wind turbine*). In addition, the most frequent categories used to form CNs were the following: (i) WIND MOVEMENT (67 times, e.g. *wind resource*); (ii) ENERGY (64 times, e.g. *power production*); (iii) QUANTITY (67 times, e.g. *feed-in tariff*); (iv) INSTRUMENT (30 times, e.g. *rotor disc*); (v) ACTIVITY (29 times, e.g. *system operation*); (vi) SYSTEM (28 times, e.g. *power generation system*); (vii) MAGNITUDE (22 times, e.g. *water depth*); (viii) LOCATION (20 times, e.g. *offshore market*). The combination of these categories reflected recurrent semantic patterns in the formation of CNs, which can be useful for the inference of the semantic relations between the components of CNs (Rosario et al. 2002; Maguire et al. 2010; Smith et al. 2014; Cabezas-García & León-Araúz 2018). Table 1 shows the most productive semantic patterns for term formation in our study, as well as some of the CNs derived from these conceptual combinations and the semantic relation between their constituents.

Table 1: Semantic patterns for term formation in the domain of wind power.

Semantic Pattern	Number of CNs	Examples	Semantic Relation
ENERGY + ACTIVITY	8	<i>wind-energy project</i> <i>wind power development</i> <i>wind power forecast</i>	<i>has_function</i> <i>has_patient</i>
ORIGIN + ENERGY	8	<i>electrical power</i> <i>kinetic energy</i> <i>renewable energy</i>	<i>caused_by</i> <i>has_attribute_origin</i>
ENERGY + FORMATION	7	<i>wind power generation</i> <i>energy production</i> <i>electricity generation</i>	<i>has_result</i>
ENERGY + SYSTEM	7	<i>wind power generation system</i> <i>electric power system</i> <i>energy storage system</i>	<i>has_function</i>
LOCATION + ACTIVITY	5	<i>offshore wind market</i> <i>offshore wind industry</i> <i>offshore project</i>	<i>has_attribute_location</i>
QUANTITY + QUANTITY	5	<i>total installed capacity</i> <i>net load</i> <i>load demand</i>	<i>has_attribute_quantity</i> <i>has_patient</i> <i>represents</i>
WIND MOVEMENT + FACILITY	5	<i>wind power plant</i> <i>wind farm</i> <i>wind park</i>	<i>uses_resource</i>

As can be observed, regularities were found between CNs formed by the same conceptual categories and the semantic relations encoded. Namely, the following semantic patterns established the same semantic relation in all of their CNs: ENERGY + FORMATION (*has\_result*); ENERGY + SYSTEM (*has\_function*); LOCATION + ACTIVITY (*has\_attribute\_location*); and WIND MOVEMENT + FACILITY (*uses\_resource*). However, other patterns had more than one semantic relation, which were indicative of semantic constraints. For example, ENERGY + ACTIVITY activated two semantic relations: *has\_patient*, when the CN was formed by predicate nominalization (e.g. *wind power development*, *wind power forecast*), and *has\_function*, when the CN was formed by predicate deletion (e.g. *wind-energy project*, *electricity market*). This highlighted the role of predicates in CN formation (Levi



1978; Cabezas-García & Faber 2017b). As for ORIGIN + ENERGY, the *caused\_by* relation was mostly established (as in *electrical energy* and *kinetic energy*) though the *has\_attribute\_origin* relation was found in CNs with *renewable* (e.g. *renewable energy*, *renewable electricity*), which actually modifies the source from which energy can be produced. Finally, QUANTITY + QUANTITY tended to encode the *has\_attribute\_quantity* relation, when it included adjectives determining how the quantity must be understood (as in *total installed capacity* and *net load*). Additionally, it can activate the *represents* relation (e.g. *capacity credit*), or the *has\_patient* relation, when the CN is formed by predicate nominalization (as in *load demand*). These semantic relations correlated with the most frequent relations in all of the CNs of our study, which were *has\_patient* (in 35 CNs), *has\_function* (in 32 CNs), and *represents* (in 18 CNs). The broad ontological distinction of conceptual categories between entities, processes, and attributes also shed light on the relevance of certain relations. Namely, the *has\_function* relation was prevalent in entities (29 out of 177 CNs), whereas *has\_patient* was the most frequent relation in processes (29 out of 60), and *attribute\_state* was the most recurrent relation in the attributes (three out of 13 CNs). This was not surprising, since entities are usually described in terms of their use; processes often emphasize the concept that receives the action; and attributes specify properties, such as the state. Therefore, analyzing the semantic aspects of the formation of CNs can help to identify recurrent patterns in the production of these MWTs. The application of this semantic information to knowledge resources facilitates understanding of concepts, as well as domain and frame reconstruction (Sager et al. 1980).

## 5 Phraseological Module of EcoLexicon: The Semantic Combinations View

### 5.1 EcoLexicon and its Phraseological Module

EcoLexicon (Faber et al. 2016), based on the theoretical premises of Frame-based Terminology (Faber 2012), is a multidimensional terminological knowledge base on environmental science. It targets user knowledge acquisition through different types of multimodal and contextualized information to respond to cognitive and communicative needs. Its public is any user group interested in broadening their knowledge of the environment for text comprehension and/or generation (e.g. environmental experts, technical writers, translators). This resource is currently available in English and Spanish, though five more languages (German, Modern Greek, Russian, French and Dutch) are being gradually implemented. To date, its database consists of a total of 3,950 concepts and 21,720 terms.

EcoLexicon has a visual interface with different modules for conceptual, linguistic, and graphical information. Because of the importance of phraseological information in terminological resources, a new phraseological module is currently under construction. Although the module describes verbal collocations and CNs, this paper only deals with the CN submodule. This submodule contains four views: (1) the Modifiers + Head view; (2) the EN-ES view; (3) the Syntactic Combinations View; and (4) the Semantic Combinations view. The Modifiers + Head view focuses on CN formation and allows users to search a list of CNs that contain a given term (e.g. if the user looks for *turbine*, the query will show a list of CNs including *horizontal axis wind turbine*, *fixed speed wind turbine*, *wind turbine foundation*, etc.). Moreover, the search tool in the EN-ES view offers bilingual results in English and Spanish (e.g. the user can search *wind turbine* and then find *aerogenerador*, or vice versa). The Syntactic Combinations view offers advanced queries according to the part of speech (e.g. if the user looks for N+N+N, the results will include *sea level rise*, *incident wave height*, *sediment transport rate*, etc.). Regarding the Semantic Combinations view, it allows users to perform queries based on the semantics of CNs in terms of their conceptual categories and internal semantic relations, as will be shown in Section 5.2.

## 5.2 The Semantic Combinations View

In the Semantic Combinations view of the CN submodule, users can perform a simple or an advanced query. Figure 4 shows the query screen and the results screen of the simple query “wind power”. The simple query box can be used to perform a proximity search. As shown in the results screen, the system automatically converts the user’s search into a query expression (“wind[TERM] AND power[TERM]”) and displays a list of results in EcoLexicon that include the terms (e.g. *offshore wind power*, *wind power generation*). The alphabetically listed results show the CN divided into its components with the bracketing, along with the conceptual category of each constituent (below, in red) and the internal semantic relation that links these components (on the right, in blue). For instance, the first result in Figure 4 is *offshore wind power*, which is described as *offshore* [LOCATION] *wind power* [ENERGY] and contains the semantic relation *has\_attribute\_location*. To the right edge of each result, there is a “+” symbol that displays a box with additional information about the entry: definition, semantic information, corpus concordances, verbal collocations, access to the full entry in EcoLexicon, and notes. The definition describes the concept. Additionally, the semantic information section provides conceptual information, such as the semantic category of the CN, each of its constituents, and the internal relations in CNs formed by more than two terms (e.g. in *wind power forecast*, *forecast has\_patient* wind power, and *power is\_caused\_by* wind). The corpus concordances allow users to access the EcoLexicon corpus in Sketch Engine. Furthermore, the verbal collocations option links the two sections of the phraseological module, namely CNs and verbal collocations, because many collocations have a nominal equivalent in the form of a CN (e.g. *a volcano erupts* > *volcano eruption*). Finally, the term entry in EcoLexicon provides synonyms of the CN, equivalents in different languages, images, conceptual networks, etc.

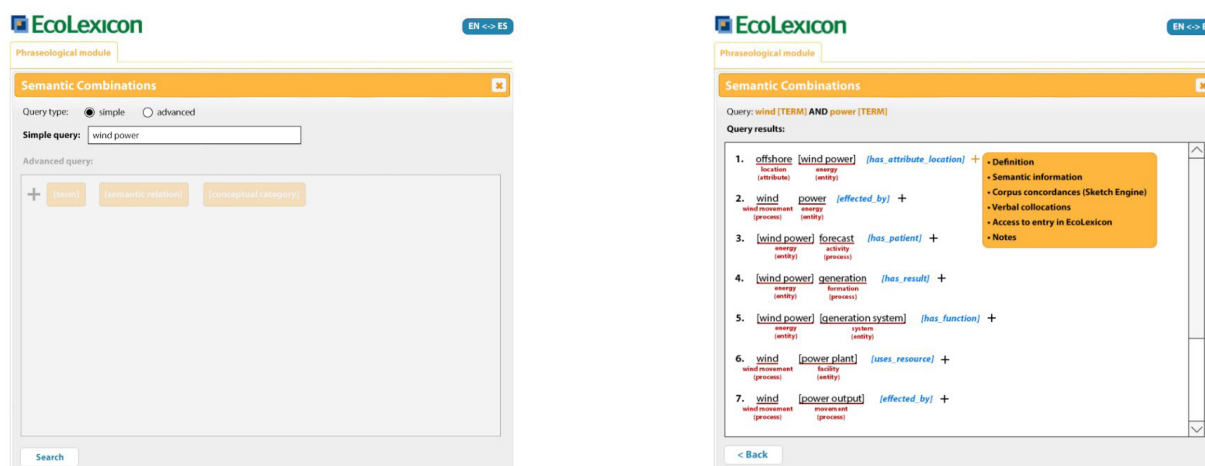


Figure 4: Simple query (left side) and results (right side) in the Semantic Combinations view using the following expression: wind[TERM] AND power[TERM].

The advanced query presents a series of particularities that allow users to perform more complicated searches. As shown in Figure 5, the advanced query is based on three elements: (i) terms; (ii) semantic relations; (iii) conceptual categories. By clicking on the orange bubbles next to the “+” symbol, the user can add as many elements to the query as they want and in any order, since this query allows for free element combination (e.g. CATEGORY + CATEGORY, term + CATEGORY, CATEGORY + relation + CATEGORY, etc.). In the same way, any element can also be deleted. The term bubble has a free text box to type anything, whilst the semantic relation and the conceptual category bubbles display a picklist showing all the relations or categories contained in EcoLexicon. However, it is also possible

to choose the options “ANY RELATION” or “ANY CATEGORY”. In fact, displaying all the possibilities with a picklist is the simplest way for users to easily find and choose the most suitable option for their query. In addition, each bubble contains an “ADD” and an “OR” button, which are useful if the user wants to look for more than one term, relation and/or category found in the same position.

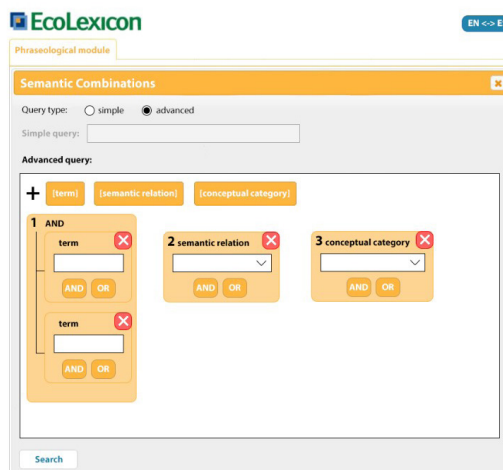


Figure 5: Advanced query in the Semantic Combinations view, showing the free element combination.

Figure 6 shows the query screen and the results screen of the advanced query “offshore[TERM] + [ANY CATEGORY]”. In order to perform this search, the user must select the option “advanced” next to “Query type”, and this will activate the advanced query box, where the user will then create a term bubble in order to type “offshore”, and a conceptual category bubble in order to select “ANY CATEGORY”. As a consequence, this expression displays a series of results that include *offshore capacity*, *offshore market* and *offshore project*, to name a few examples. As in the simple query, the results are also listed alphabetically, showing the conceptual categories of the groupings in the CN and the internal semantic relation.

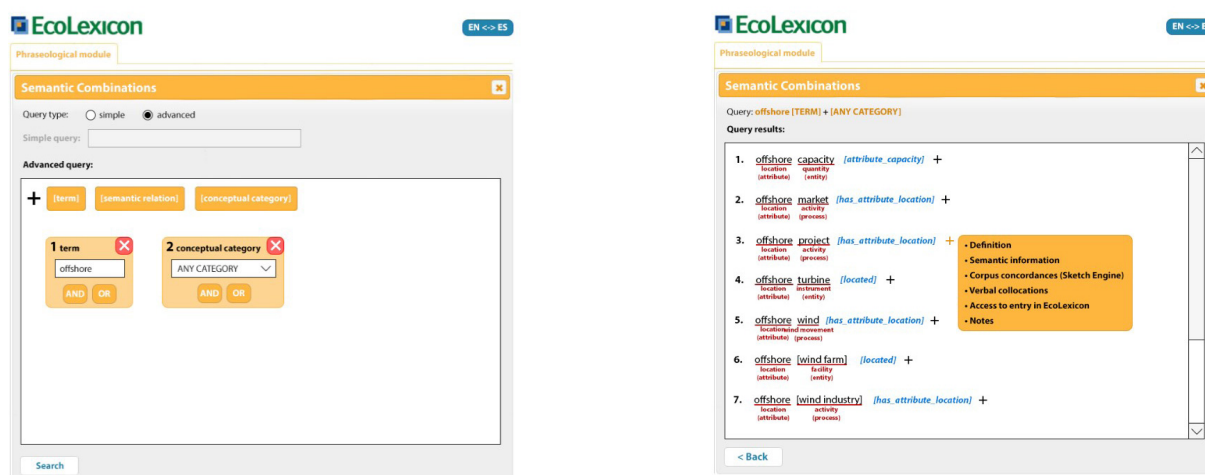


Figure 6: Advanced query (left side) and results (right side) in the Semantic Combinations view with the following expression: offshore[TERM] + [ANY CATEGORY].

Finally, another feature of the results screen is the possibility of offering similar results based on the search criteria. In Figure 7 is shown an advanced query of CNs made of two conceptual categories [“energy (entity)[CATEGORY] + formation (process)[CATEGORY]”], which has six results (e.g.

*electricity generation, electricity production, energy generation*). In this way, the Semantic Combinations view allows the user to see more results by clicking on “+ Show similar results”, which would display those CN that meet the search criteria entered by the user, but in reverse order (e.g. *generated power, generated energy, produced power*). Accordingly, what matters in CN formation is meaning and content, not shape (Štekauer 1998).

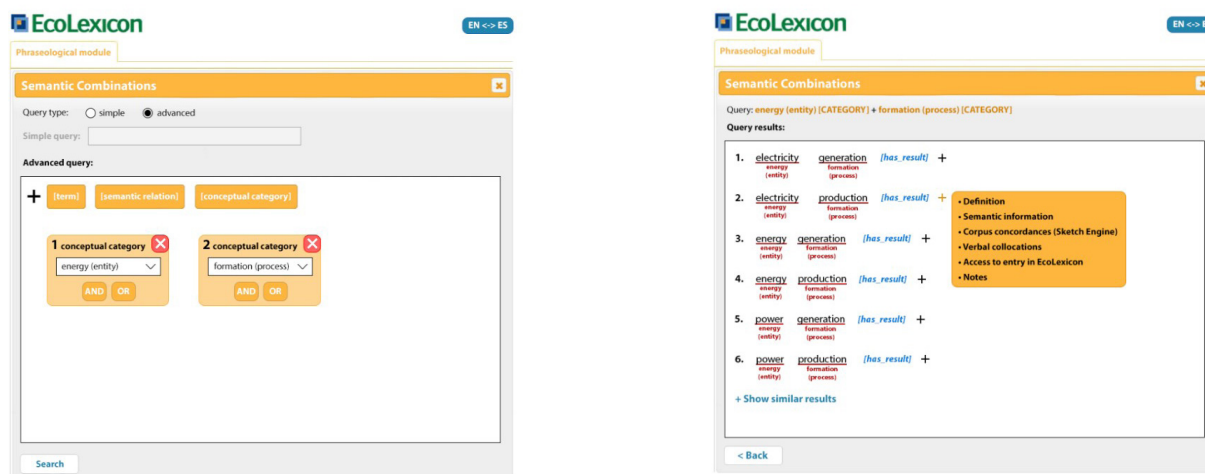


Figure 7: Advanced query (left side) and results (right side) in the Semantic Combinations view using the following expression: energy (entity)[CATEGORY] + formation (process)[CATEGORY].

This Semantic Combinations view enhances the CN submodule by adding a conceptual approach to the queries that users can perform inside the new phraseological module. Since specialized knowledge is not conceptualized in isolation, but rather as part of a context or frame (Faber 2012), the semantic and conceptual features contained in this tool help to describe concept structure and interrelationships within the environmental domain.

## 6 Conclusion

Complex nominals (CNs) pose different problems, such as their identification in texts, their syntactic and semantic description, their translation into different languages, their representation in lexicographic and terminographic works, *inter alia*. Accordingly, we carried out a semantic analysis of a set of CNs extracted from a wind energy corpus. This analysis was performed by assigning conceptual categories both to the CNs as a whole and to their constituent parts. The semantic relations within the CNs were also described. As part of this process, the most recurring knowledge patterns in CN formation were detected with a view to applying this semantic information to the CN module in EcoLexicon. In this way, users could perform queries based on the meaning of these linguistic units.

Our results showed that the constituent concepts of CNs did not combine randomly, but rather as a result of semantic constraints. In particular, within the wind energy domain the most frequent patterns of CN formation were the following combinations: ENERGY + ACTIVITY, ORIGIN + ENERGY, ENERGY + FORMATION, ENERGY + SYSTEM, LOCATION + ACTIVITY, QUANTITY + QUANTITY, and WIND MOVEMENT + FACILITY. Since certain semantic relations tended to be established between these category pairs, this helped to infer the internal semantic relations in the CNs composed of these categories. Therefore, the existence of a certain relation in a category pair indicated that most of the CNs with these same categories would also have the same internal semantic relation (Rosario et al. 2002; Maguire et al. 2010; Smith et al. 2014; Cabezas-García & León-Araúz 2018).



Our ultimate objective was to apply the semantic information extracted from CNs to one of the new views of the CN module in EcoLexicon: the Semantic Combinations view, which is complemented by the Modifiers + Head view (showing CNs from a given term); the EN-ES view (allowing users to perform bilingual queries); and the Syntactic Combinations view (offering the possibility to combine syntactic categories in order to obtain CNs with specific syntactic patterns). The Semantic Combinations view permits users to enter different search elements, namely terms, semantic relations and conceptual categories, and combine them in any order or number. Examples of possible queries include “energy (entity)[CATEGORY] + formation (process)[CATEGORY]” (e.g. *electricity generation, energy production*), “offshore[TERM] + [ANY CATEGORY]” (e.g. *offshore turbine, offshore wind farm*), and “[ANY CATEGORY] + has\_function [RELATION] + [ANY CATEGORY]” (e.g. *control system, power plant*).

In conclusion, this new tool based on semantic information is adapted to the new tendencies in linguistic resources (Krstev et al. 2006; RANM 2012; Casares 2013; Hanks 2014), where semantics plays a key role. EcoLexicon, a resource characterized by its emphasis on conceptualization and knowledge structuration, will be enhanced with the creation of a specific section for CNs, allowing users to perform a wide range of queries. Although the set of semantic categories may vary in different domains, the semantic-based queries presented in this paper can be implemented in any electronic resource with a view to offering an enhanced user experience.

In future research, we plan to annotate every CN in EcoLexicon based on its domain or subdomain within the environment, as well as their annotation based on their semantic roles (Cabezas-García & San Martín 2017). Furthermore, semi-automatic annotation of CNs will be explored and user feedback in the CN module of EcoLexicon will be assessed. Additionally, since hyponymy is the semantic relation that is at the core of CN formation, further research will also include the analysis of the hyponymic nuances within CNs and their relation to the decomposition of hyponymy into subtypes (Gil-Berrozpe et al. 2017).

## References

- Arden, Y. (2013). *Routledge French Technical Dictionary / Dictionnaire technique anglais*. London/New York: Routledge.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '98)*, 10–14 August 1998, pp. 86–90. Université de Montréal, Canada.
- Beigbeder, F. (2006). *Diccionario técnico inglés-español español-inglés*. Madrid: Díaz de Santos.
- Cabezas-García, M., Faber, P. (2017a). A Semantic Approach to the Inclusion of Complex Nominals in English Terminographic Resources. In R. Mitkov (ed.), *Computational and Corpus-Based Phraseology*, pp. 145–159. Cham: Springer.
- Cabezas-García, M., Faber, P. (2017b). Exploring the Semantics of Multi-word Terms by Means of Paraphrases. In M.A. Candel-Mora, C. Vargas-Sierra (eds.), *Temas actuales de terminología y estudios sobre el léxico*, pp. 193–217. Granada: Comares.
- Cabezas-García, M., León-Araúz, P. (2018). Towards the Inference of Semantic Relations in Complex Nominals: A Pilot Study. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, 7–12 May 2018, pp. 2511–2518. Miyazaki, Japan: ELRA.
- Cabezas-García, M., San Martín, A. (2017). Semantic annotation to characterize contextual variation in terminological noun compounds: a pilot study. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 4 April 2017, pp. 108–113. Valencia: Association for Computational Linguistics.
- Caignon, P. (2001). *Vocabulaire et cooccurrents de la comptabilité*. Montréal: Linguattech.
- Casares, J. (2013). *Diccionario ideológico de la lengua española (de la idea a la palabra y de la palabra a la idea)*. Madrid: Gredos.



- Chan, S.W. (2004). *A Dictionary of Translation Technology*. Hong Kong: Chinese University Press.
- Cleveland, C., Morris, C. (2015). *Dictionary of Energy*. Amsterdam: Elsevier.
- Faber, P. (2012) (ed.). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.
- Faber, P., León-Araúz, P., and Reimerink, A. (2016). EcoLexicon: new features and challenges. In I. Kernerman., I. Kosem Trojina, S. Krek, and L. Trap-Jensen (eds.), *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, pp. 73–80. Portorož, Slovenia.
- Fellbaum, C. (2005). WordNet and wordnets. In K. Brown et al. (eds.), *Encyclopedia of Language and Linguistics*, pp. 665–670. Oxford: Elsevier.
- Fillmore, C.J. (1985). Frames and the semantics of understanding. In *Quaderni di Semantica*, 6(2), pp. 222–254.
- Gil-Berrozpe, J.C., León-Araúz, P., and Faber, P. (2017). Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. In *Proceedings of the Fifth International Conference on Electronic Lexicography in the 21st Century (eLex 2017)*, 19–21 September 2017, pp. 63–92. Leiden, The Netherlands.
- Girju, R., Moldovan, D., Tatu, M., and Andantohe, D. (2005). On the semantics of noun compounds. In *Computer Speech and Language*, 19(4), pp. 479–496.
- Hahn, U., Chater, N. (1997). Concepts and Similarity. In K. Lamberts and D. Shanks (eds.), *Knowledge, Concepts, and Categories*, pp. 93–131. Cambridge (MA)/London: MIT Press.
- Hanks, P. (2014). *Pattern Dictionary of English Verbs*. Accessed at: <http://pdev.org.uk/> [27/03/2018].
- Hidalgo, A. (2014). *Elsevier's Dictionary of Medicine Spanish-English English-Spanish*. Amsterdam: Elsevier.
- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Abingdon: Routledge.
- Kilgariff, A., Rychlý, P., Smrž, P. and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EU-RALEX International Congress*, 6–10 July 2004, pp. 105–116. Lorient, France.
- Kipper-Schuler, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 22–28 May 2006, pp. 1027–1032. Genoa, Italy: ELRA.
- Krstev, C., Vitas, D., and Savary, A. (2006). Prerequisites for a Comprehensive Dictionary of Serbian Compounds. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala (eds.), *Advances in Natural Language Processing: Lecture Notes in Computer Science*, 4139, pp. 552–563. Berlin/Heidelberg: Springer.
- L'Homme, M.C. (2014). Terminologies and taxonomies. In J.R. Taylor (ed.), *Handbook of the Word*. Oxford: Oxford University Press.
- Lauer, M., Dras, M. (1994). A probabilistic model of compound nouns. In *Proceedings of the 7th Australian Joint Conference on Artificial Intelligence*, pp. 474–481. Singapore, Singapore.
- León-Araúz, P., San Martín, A., and Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm 2016)*, 12 December 2016, pp. 73–82. Osaka, Japan.
- Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.
- Maguire, P., Wisniewski, E.J., and Storms, G. (2010). A corpus study of semantic patterns in compounding. In *Corpus Linguistics and Linguistic Theory*, 6(1), pp. 49–73.
- Marshman, E. (2006). Lexical Knowledge Patterns for Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts: A Comparative Study of English and French. PhD Thesis. Université de Montréal, Montréal, Canada.
- Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, and M.C. L'Homme (eds.), *Recent Advances in Computational Terminology*, pp. 279–302. Amsterdam/Philadelphia: John Benjamins.
- Murphy, G.L., Lassaline, M.E. (1997). Hierarchical Structure in Concepts and Basic Level of Categorization. In K. Lamberts and D. Shanks (Eds.), *Knowledge, Concepts, and Categories*, pp. 93–131. Cambridge (MA)/London: MIT Press.
- Nakov, P., Hearst, M. (2006). Using Verbs to Characterize Noun-Noun Relations. In *Artificial Intelligence Methodology Systems and Applications*, 4183, pp. 233–244.
- Nakov, P., Hearst, M. (2013). Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases. In *ACM Transactions on Speech and Language Processing*, 10(3), pp. 1–51.
- Observatoire de Linguistique Sens-Texte (OLST) (2013). *DiCouèbe: Dictionnaire en ligne de combinatoire du Français*. Accessed at: <http://olst.ling.umontreal.ca/dicouebe/index.php> [27/03/2018].

- Observatoire de Linguistique Sens-Texte (OLST) (2018). *DiCoInfo: Le dictionnaire fondamental de l'informatique et de l'Internet*. Accessed at: <http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi> [27/03/2018].
- Observatoire de Linguistique Sens-Texte (OLST) (2018). *DiCoEnviro: Dictionnaire fondamental de l'environnement*. Accessed at: [http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search\\_enviro.cgi](http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi) [27/03/2018].
- Real Academia Nacional de Medicina (RANM) (2012). *Diccionario de términos médicos*. Madrid: Panamericana.
- Rosario, B., Hearst, M., and Fillmore, C. (2002). The Descent of Hierarchy, and Selection in Relational Semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, 7–12 July 2002, pp. 247–254. Philadelphia, Pennsylvania, United States.
- Sager, J.C., Dungworth, D., and McDonald, P.F. (1980). *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.
- Sanz-Vicente, L. (2012). Approaching secondary term formation through the analysis of multiword units: An English–Spanish contrastive study. In *Terminology*, 18(1), pp. 105–127.
- Smith, V., Barratt, D., and Zlatev, J. (2014). Unpacking noun-noun compounds: interpreting novel and conventional food names in isolation and on food labels. In *Cognitive Linguistics*, 25(1), pp. 99–147.
- Štekauer, P. (1998). *An Onomasiological Theory of English Word-Formation*. Amsterdam/Philadelphia: John Benjamins.
- Štekauer, P., Valera, S., and Körtvélyessy, L. (2012). *Word-formation in the world's languages*. Cambridge/New York: Cambridge University Press.

## Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by the FPU grants given by the Spanish Ministry of Education to both authors. Finally, we would like to thank the anonymous reviewers for their useful comments.



# Towards a Glossary of Rum Making and Rum Tasting

**Cristiano Furiassi**

*University of Turin*

*E-mail: cristiano.furiassi@unito.it*

## Abstract

A lexicographic work exclusively dedicated to the making and tasting of rum has not been published to date. With the ambitious aim of filling this editorial gap in mind, this article focuses on the implementation stage of a specialized glossary of rum-related terms in the English language. Preceded by an overview of the historical, geographical and linguistic factors that made rum a renowned global product, the computer-assisted terminology acquisition procedures applied in order to extract rum-related terms from an *ad hoc* corpus are described. By merging computer-assisted term extraction with data collected from experts' knowledge, fieldwork and the existing specialized literature on rum, a list of candidate headwords was drafted. The replicability of the methodology applied makes this pilot study generalizable, thus fostering the compilation of specialized glossaries connected to other fields or disciplines.

**Keywords:** computer-assisted term extraction, glossary, rum, specialized lexicography

## 1 Introduction

A lexicographic product exclusively focusing on rum, namely a reference tool where both rum amateurs and connoisseurs can look up notions on the making and tasting of rum, is still missing on the market. Following an introductory section on the various historical and geographical aspects concerning rum, which also includes linguistic information about the word *rum* itself, the main aim of this article is to describe the implementation stage of a specialized glossary of rum-related terms, namely “[a] type of REFERENCE WORK which lists a selection of words or phrases, or the terms in a specialised field, usually in alphabetical order, together with minimal definitions or translation equivalents” (Hartmann & James 2002: 63).

More precisely, the article deals with the selection of headwords, the most salient macrostructural feature of any glossary, by showing how a list of candidate items may be obtained by exploiting a specialized corpus containing texts about rum written in English through the combination of (partly) automatic, namely “corpus-driven” (Krishnamurthy 2008: 231), and (mostly) semi-automatic, namely “corpus-based” Tognini Bonelli 2001: 65), techniques.<sup>1</sup>

The term-extraction procedures described are limited to specialized texts about rum written in English. However, the fact that rum production has spread on a large scale, also involving the French-speaking and the Spanish-speaking Caribbean, makes rum a global product *par excellence*. Therefore, by applying the same procedures, an additional step would lead to the compilation of a multilingual glossary of rum.

<sup>1</sup> A similar, though more sophisticated, approach to specialized lexicography regarding alcoholic beverages, namely wine, was adopted by Leroyer (2015; 2018) for his *Oenolex Wine Dictionary*.

## 2 Rum as a Global Product

In many rum-producing countries and especially in its Caribbean birthplace, Barbados, rum often represents a national symbol deeply rooted in the local culture.<sup>2</sup> In fact, in the 15<sup>th</sup> century it was Christopher Columbus who brought to the Caribbean a large amount of sugarcane from Spain, specifically the Canaries. At the beginning of its production, in the 17<sup>th</sup> century, rum was considered a drink of little value and unpleasant taste, lacking the prestige of more refined distillates made in Europe. However, by the 18<sup>th</sup> century, besides having become a precious export, the importance of rum grew both in the North American continent and Europe.

As opposed to other world-famous spirits, such as, for instance, cognac, gin, vodka and whisk(e)y, readily associable with France, England, Russia and Scotland or Ireland respectively, rum – an icon of the “cultural fragmentation” (Furiassi 2014: 91) typical of the Caribbean – spread throughout the world and eventually reached all continents to the point that nowadays the general public seems to ignore its exact origin and can hardly associate it with a particular territory. Within North America, a few distilleries may be found in the United States. In the Caribbean, most of the Greater Antilles and virtually all the Lesser Antilles are renowned for producing rum. Moreover, various mainland territories of Central America, including Belize, Costa Rica, Guatemala, Nicaragua and Panama, produce rum. In South America, rum is distilled in Argentina, Brazil, Colombia, French Guiana, Guyana, Paraguay, Peru, Suriname and Venezuela. In Asia, Japan, Nepal, the Philippines and Thailand are involved in the making of rum, while in Oceania, Australia, Fiji and New Zealand are rum-producing countries. In Africa, Madagascar, Mauritius, Réunion, South Africa and Saint Helena are also known for the production of this drink. In Europe, rum distillation is limited to Las Palmas, one of the Canary Islands. Finally, it is worth mentioning that even remote islands, such as Bermuda, are celebrated rum makers.

Lexicographically, the word *rum* has been defined alternatively as an alcoholic drink, liquor or spirit, as the following quotations show: “[a]n alcoholic spirit distilled from molasses and other sugar-cane products, prepared chiefly in the Caribbean and parts of Central and South America; a serving or variety of this” (*OED*); “an alcoholic liquor prepared by fermenting molasses, macerated sugarcane, or other saccharine cane product, distilling, coloring with caramel, and aging” (*Merriam-Webster*); “[a]n alcoholic drink industrially distilled from the juice of the sugar-cane, blended and cured in barrels” (*DCEU*).<sup>3</sup>

As far as its earliest attestation is concerned, the first written account of the word *rum* in English dates from 1654 (*OED*, *Merriam-Webster*). Nonetheless, it must be noted that the culture surrounding rum is also ingrained in the French- and Spanish-speaking Caribbean. In fact, its French and Spanish cognates, namely *rhum*, first recorded in 1688 (*TLFi*), and *ron*, dating from about 1770 (*DECH*), derive from English *rum*, as attested in authoritative lexicographic sources such as, for instance, the *FEW* and the *TLFi* for French and the *DECH* and the *DRAE* for Spanish.

## 3 Data Retrieval and Methodology: The Caribbean Rum Corpus (CRC)

Among others, Gamper and Stock (1998: 147) claim that “[t]he manual acquisition of terminological material from the domain-specific text material is a very time-consuming task. [...] Computer-assisted

2 As reported by Smith (2008: 13), “[...] evidence indicates that the British island of Barbados and the French island of Martinique were the cradles, if not the birthplaces, of Caribbean rum”.

3 Originally the shortening of *rumbullion*, an Early Modern English word perhaps originated in Devonshire and meaning ‘a great tumult’ or ‘uproar’, over the centuries rum was called by many different names, mostly referring to its close association with seafaring, buccaneering and the infernal regions: *Barbados water*, *devil’s death*, *grog*, (*hot*) *hellish liquor*, *kill-devil*, *navy neaters*, *Nelson’s blood*, *pirate’s drink*, *rumbullion*, *rumbustion* and *taffia* or *tafia* – all included in the glossary headword list (see table 1).



term acquisition improves both the quantity and the quality of terminological work”. Consequently, the first action taken to obtain a wordlist of rum-related terms was to design and compile a specialized – or “special” (Tognini Bonelli & Sinclair 2006: 210) – corpus, namely the *Caribbean Rum Corpus (CRC)*.

Texts contained in the *CRC* include material from websites created by rum experts,<sup>4</sup> that is the official websites of 25 rum makers throughout the English-speaking Caribbean – all listed in the reference section, thus providing “adequate coverage of the field in question” (Bowker 2003: 162).<sup>5</sup> Here follows the list of rum makers grouped by territory: *Anguilla Rums* (Anguilla); *Antigua Distillery* (Antigua and Barbuda); *Bacardi, Todhunter-Mitchell Distillery* (Bahamas); *Cockspur, Foursquare Rum Distillery* (where *Doorly’s*, *E.S.A. Field*, *Mahiki*, *Old Brigand*, *The Real McCoy* and *R.L. Seale* are produced), *Mount Gay Distillers*, *St. Nicholas Abbey* (Barbados); *Gosling* (Bermuda); *Arundel Estate Callwood Distillery*, *Pusser’s* (British Virgin Islands); *Clarke’s Court*, *Grand Havana Rum*, *Westerhall Estate* (Grenada); *Appleton Estate*, *Blackwell*, *Captain Morgan*, *Coruba*, *Myers’s*, *Worthy Park Estate* (Jamaica); *Elements Eight Rum*, *St. Lucia Distillers* (St. Lucia); *10 Cane*, *Angostura*, *Caroni*, *Zaya* (Trinidad and Tobago).

Depending on the degree of usability of each website – how practical it was to extract plain text, most of the makers listed above were considered except for *Caroni*, whose website is non-existent since the distillery closed in 2002, and *Anguilla Rums*, *Mount Gay Distillers* and *St. Nicholas Abbey*, whose websites could not be exploited for technical reasons – the automatic extraction of texts was not allowed. At the end of the collection procedure, the *CRC* amounted to 33,625 tokens and 4,202 types: for the task at hand, the choice of texts and number of running words seem to meet both the representativeness and reliability requirements which a specialized corpus must satisfy in order to be useful for linguistic and lexicographic investigation (Biber 2008: 63-64; Bowker & Pearson 2002: 45).<sup>6</sup>

## 4 Computer-assisted Term Extraction

Once the *CRC* was collected, the data gathered were processed by means of *WordSmith Tools*.<sup>7</sup> The *CRC* wordlist (4,202 types), obtained via the *WordList* tool, was compared with two wordlists extracted from two general corpora of the English language, i.e. the *Freiburg-Lancaster-Oslo-Bergen Corpus of British English (FLOB)* and the *Freiburg-Brown Corpus of American English (FROWN)*, the former containing texts typical of British English, the latter based on American English. Although compiled much earlier, namely in the early 1990s, the two reference corpora selected were considered functional for the lexicographic purpose at hand, as the size, granularity and text types included made term extraction viable. In addition, the *FLOB* and the *FROWN*, albeit somehow dated, were considered instead of the *British National Corpus (BNC)* and the *Corpus of Contemporary American English (COCA)*, among others, because of usability criteria: in practice, the availability of all texts belonging to the *FLOB* and the *FROWN* allowed both corpora to be processed by *WordSmith Tools*.

Two separate non-lemmatized keyword lists were thus obtained using the *KeyWords* tool: one, containing 472 positive keywords, resulting from the comparison between the *CRC* and the *FLOB* wordlists, and another, containing 475 positive keywords, resulting from the comparison between the *CRC*

4 See Bergenholtz (1995: 19-20), Bowker & Pearson (2002: 27-28) and Gotti (2011: 25-28) for a distinction of levels of expertise in the encoding/decoding of specialized texts.

5 Atkins & Rundell (2008: 80) suggest that “a carefully designed web corpus can provide reliable language data”.

6 Although, at present, the corpus may look small, it must be noted that the domain under investigation is highly specialized. However, the *CRC* could be expanded in a future phase by extending the range of websites considered to those of rum producers in other English-speaking parts of the world, such as Australia, Fiji, New Zealand, the Philippines, South Africa and the United States.

7 Despite the existence of various types of text analysis software such as, for instance, *AntConc* and *TextSTAT*, *WordSmith Tools* is among the few – to the author’s knowledge – which allow the analyst to conveniently compare corpus wordlists in order to detect the keyness of certain items, and was specifically selected to help in such endeavors (see footnote 8).

and the *FROWN* wordlists.<sup>8</sup> After being merged, the positive keywords generated by *WordSmith Tools* (512 tokens) were of paramount importance for extracting single- and/or multi-word terminological units which will then become headword candidates of the rum glossary.

Keyword lists were expressly drawn to highlight “topic sensitivity” (Ringbom 1998: 48): it was essential to detect topic-sensitive items, that is “words that are closely linked to the topics dealt with” (Furiassi 2004: 194) in the *CRC*, namely rum. However, in order to keep only topic-sensitive content words, the resulting list, which still included some noise, i.e. undesired items, had to be further reduced (328 items) by manually eliminating function words and proper nouns related to rum brands/makers and toponyms.<sup>9</sup>

While most of these items can be intuitively associated with the specialized language of rum, e.g. *barrel*, *distillation*, *molasses*, others are also common in general English, e.g. *gold*, *scent*, *wood*, though obviously acquiring a specialized meaning in a rum-oriented context. Therefore, via the *Concord* tool, a concordance output was provided for each item in the wordlist thus obtained in order to establish whether it should qualify as a headword in the glossary.

In addition, the aim of the present study was not restricted to the selection of single words contained in the *CRC* keyword list (see footnote 9). Indeed, in order not to miss recurrent “collocations” (Sinclair 1991: 109-121), high-frequency word clusters were also obtained for each positive keyword: clusters range from a minimum of a two-word combination to a maximum of a four-word combination. Finally, from both the noise-free *CRC* keyword list and the manually-selected clusters, a wordlist of candidate items suitable for inclusion as headwords in a glossary of rum was gathered.<sup>10</sup>

## 5 The Selection of Headwords

Since LSP lexicography cannot rely entirely on corpus data, a final list of candidate headwords was drafted only after combining computer-assisted term extraction from the *Caribbean Rum Corpus* (*CRC*) – and the *Guyana Rum Corpus* (*GRC*), expressly collected at a later stage (see Section 5.2) – with data gathered from experts’ knowledge, fieldwork and the existing specialized literature on rum published in English.

### 5.1 Headword Selection from the *CRC*

Alongside the initial corpus-based term-extraction procedures, carried out through the *KeyWords* tool provided by *WordSmith Tools*, the wordlist of candidate headwords selected from the *CRC* was mostly the result of a semi-automatic procedure since the following decisions were made:

- 8 Chung (2003: 221) states that “[...] the corpus comparison approach using word types is a reasonably simple and practical way of identifying terms”. More specifically, Furiassi (2004: 201), maintains that “[t]he comparison of two wordlists provides information about the keyness of each word in a corpus [...]. Positive keywords are items that occur more often than would be expected by chance in comparison with the reference corpus”.
- 9 Unfortunately, the *KeyWords* tool provided by *WordSmith Tools* only extracts single-word units automatically. However, although function words, e.g. *of*, were discarded at this stage, they may still be included in the rum glossary as part of multi-word units gathered manually by selecting their typical clusters detected through the *Concord* tool, e.g. *gram of alcohol* (see Table 1).
- 10 Groundbreaking NLP processing tools for term extraction, which work on lemmatized, POS-tagged wordlists extracted from corpora through statistical methods, were made available after the present research was conceived. Indeed, term extractors such as *OneClick Terms*, powered by *Sketch Engine*, and *TermoStat*, which exploit a hybrid method, i.e. statistical plus linguistic, to identify candidate terms, would undoubtedly contribute to the implementation stage of a glossary of rum. In particular, as far as terms to be considered as candidate headwords are concerned, it would then be mandatory to verify whether the same corpus data processed by fully-automatic term extractors produce similar outputs or, most certainly, the glossary is improved by including additional headwords. Moreover, once a final list of headwords is obtained, term extractors are also likely to enrich the lexical information for each headword in the glossary, e.g. word-class assignment.

- all words linked to rum making and rum tasting were included;
- abbreviations and acronyms were taken into account only if closely connected to the specialized language under scrutiny;
- headword status was also granted to multi-word lexical units resulting from cluster selection.

Following these criteria, a list, which contains 295 headwords and 81 sub-headwords, was drafted.

## 5.2 Headword Selection from the *GRC*

A glossary of rum would not be complete without considering distillers based in English-speaking Guyana, another famous rum-producing territory connected with the Caribbean.<sup>11</sup> Therefore, a smaller corpus including texts extracted from the websites of Guyanese rum producers, namely *Demerara Distillers* (maker of award-winning *El Dorado Rum*), and *XM*, was compiled. Despite the fact that the *XM* website could not be exploited since the automatic extraction of texts was not allowed, the *Guyana Rum Corpus* (*GRC*), consisting of 5,327 tokens and 1,323 types, underwent the same semi-automatic term-extraction procedures applied to the *CRC*, thus allowing the retrieval of 10 new headwords – including acronyms, i.e. *Authentic Caribbean Rum*<sup>TM</sup> (*ACR*<sup>TM</sup>), *butterscotch*, *Savalle still*, *signature rum*, *texture*, *toffee*, *uncrystallised sugar* and *West Indies Rum & Spirits Producers' Association* (*WIRSPA*), and two new sub-headwords, i.e. *exotic fruit* and *flavo(u)rful*.

## 5.3 Headword Selection from Experts' Knowledge, Fieldwork and Specialized Literature

Corpus-based LSP lexicography, also known as computer-assisted term extraction or computer-assisted terminology acquisition, must be complemented by information retrieved from experts' knowledge, fieldwork and specialized literature.

Therefore, experts' knowledge gathered from detailed visits of Barbadian rum distilleries represented a valuable source of information, as it provided a number of headwords that could not be selected otherwise and that could only be extrapolated by interviewing tour guides, master blenders and master distillers, watching the documentaries shown as part of the guided tours and analyzing the various signs and posters on display inside distilleries. In addition, the specialized literature on the topic published to date was taken into account, namely Barty-King and Massel (1983), Arkell (1999), Plotkin (2001), Ruthström (2001), Broom (2003), Coulombe (2004), Williams (2006), Curtis (2007), Smith (2008), Miller et al. (2009), Liberman (2010), Laurie (2011), Foss (2012), Hopkins (2012), Maier (2013) and Smiley, Watson and Delevante (2014).

Consequently, with the aid of experts' knowledge and specialized literature, 19 additional headwords, namely *Coffey still*, *condensation*, *cooper machine*, *cut*, *de-ionized water*, *de-mineralized water*, *earthy*, *harmonious*, *head*, *heart*, *master distiller*, *pastry*, *pepper*, *peppermint*, *reduction*, *single cask*, *single distillation*, *subtle* and *tail*, and four additional sub-headwords, *oaky*, *ripe fruit*, *toasted wood* and *woody*, were collected.

## 6 Findings: The Rum Glossary Headwords

Eventually, it is worth mentioning that the “lexicographer's intuition” (Sinclair 2003: 167) was of paramount importance to decide whether a lexical item or collocation had to be included or excluded

11 Even though Guyana, officially the Co-operative Republic of Guyana, is geographically in South America, politically, culturally and linguistically it is considered part of the Caribbean. Guyana is also among the founder members of the *Caribbean Community of Commonwealth States* (*CARICOM*); indeed, the headquarters of *CARICOM* are in Guyana, in the capital city of Georgetown, within the Demerara-Mahaica region.

from the final glossary headwords. Headwords in Table 1 are listed in alphabetical order (horizontally, from left to right) with grey shading signaling the first term for each letter of the alphabet. Each headword appears in bold; alternative spelling variants are shown in italics next to the headword. Some headwords required the insertion of sub-headwords: sub-headwords are shown in roman below the corresponding headword. Headwords (and sub-headwords) which are semantically linked to other headwords (and sub-headwords) included in Table 1 are cross-referenced: any cross reference is indicated by an arrow, i.e. →, followed by the respective headword (or sub-headword).

Table 1: Rum Glossary Headwords.

<b>absolute alcohol</b>	<b>ABV</b> → alcohol by volume → vol.	<b>aged rum</b>	<b>ageing (process)</b> <i>aging (process)</i>
<b>alcohol</b>	<b>alcohol by volume</b> → ABV → vol.	<b>alcohol recovery column</b>	<b>alcoholic fermentation</b>
<b>alcoholic strength</b>	<b>aldehyde</b>	<b>almond</b>	<b>amber</b>
<b>apricot</b> dried apricot	<b>aroma</b> → nose aroma profile	<b>aromatic</b>	<b>ACR™</b> → Authentic Caribbean Rum™
<b>Authentic Caribbean Rum™</b> → ACR™	<b>balanced</b> balanced rum	<b>Barbados water</b>	<b>barrel</b> → cask
<b>batch</b>	<b>batch distillation</b> → pot still distillation	<b>batch number</b>	<b>batch rum</b>
<b>black rum</b> → dark rum	<b>blend</b>	<b>blending</b> blending information blending instruction blending process	<b>boiling pot</b>
<b>Boston glass</b>	<b>bottle</b>	<b>bottling</b> bottling strength	<b>bounty rum</b>
<b>bouquet</b>	<b>bourbon barrel</b> → bourbon cask	<b>bourbon cask</b> → bourbon barrel	<b>brand</b>
<b>brand positioning</b> → product range	<b>bronze</b>	<b>brown</b>	<b>brown sugar</b>
<b>butterscotch</b>	<b>buttery</b>	<b>by-product</b>	<b>cane sugar</b>
<b>capacity</b> → size	<b>caramel</b>	<b>carbon dioxide</b>	<b>carbon filtration</b>
<b>cask</b> → barrel	<b>champagne glass</b>	<b>character</b>	<b>charcoal filtration</b>
<b>charred oak barrel</b> → charred oak cask	<b>charred oak cask</b> → charred oak barrel	<b>chocolate</b>	<b>cinnamon</b>
<b>citrus</b>	<b>clean</b>	<b>clove</b>	<b>cocktail glass</b>
<b>cocoa</b>	<b>coconut</b>	<b>coconut rum</b>	<b>coffee</b>
<b>Coffey still</b>	<b>Collins glass</b>	<b>colour</b> <i>color</i>	<b>column distillation</b> → continuous still distillation
<b>column still</b>	<b>complexity</b> → sophisticated complex	<b>compound</b>	<b>concentrated alcohol</b>
<b>condensation</b>	<b>congener</b>	<b>connoisseur</b>	<b>content</b>

<b>continuous still</b> → pot still continuous still rum	<b>continuous still distillation</b> → column distillation	<b>cooper machine</b>	<b>copper alembic pot</b>
<b>copper kettle</b>	<b>copper pot still</b>	<b>coupette glass</b>	<b>cream</b>
<b>crushed cane</b>	<b>cut</b>	<b>dark rum</b> → black rum	<b>dash</b>
<b>de-ionized water</b> → de-mineralized water	<b>de-mineralized water</b> → de-ionized water	<b>devil's death</b>	<b>distillation</b> distillation method distillation process
<b>distilled drink</b>	<b>distinct</b> <i>distinctive</i>	<b>double distillation</b> double distillate double distilled rum	<b>earthy</b>
<b>estate</b>	<b>ethyl</b>	<b>exact</b>	<b>expertise</b>
<b>external water jacket</b>	<b>extra old</b> → XO extra old rum	<b>factory</b>	<b>fermentation</b> fermentation process
<b>fermented wash</b>	<b>fertile soil</b>	<b>filtration</b> filtration process	<b>finish</b>
<b>first press</b>	<b>flavour</b> <i>flavor</i> → taste flavourful <i>flavorful</i>	<b>flavour compound</b> <i>flavor compound</i>	<b>flavouring agent</b> <i>flavoring agent</i>
<b>fresh</b>	<b>fruit</b> exotic fruit fresh fruit fruity honeyed fruit ripe fruit	<b>full</b> full-bodied	<b>gentle</b>
<b>gentle filtration</b> → light filtration	<b>ginger</b>	<b>glass</b>	<b>gold</b> gold rum
<b>golden</b>	<b>gram of alcohol</b>	<b>grog</b>	<b>hand blend</b>
<b>hand-crafted</b>	<b>harmonious</b>	<b>harshness</b>	<b>harvested</b> hand harvested
<b>hazelnut</b>	<b>head</b>	<b>heart</b>	<b>heavy</b> heavy bodied heavy rum
<b>heavy pot still</b> heavy pot still rum → light pot still	<b>hellish liquor</b> → hot hellish liquor	<b>highball glass</b>	<b>high proof rum</b>
<b>high wine retort</b> → low wine retort	<b>hint</b>	<b>honey</b>	<b>hot hellish liquor</b> → hellish liquor
<b>hurricane glass</b>	<b>infused</b>	<b>ingredient</b>	<b>instruction</b>
<b>intense</b>	<b>International Wine &amp; Spirits Competition</b> → IWSC	<b>International Wine &amp; Spirits Festival</b> → IWSF	<b>IWSC</b> → International Wine & Spirits Competition
<b>IWSF</b> → International Wine & Spirits Festival	<b>juice</b>	<b>kill-devil</b>	<b>label</b>
<b>labour-intensive crop</b> <i>labor-intensive crop</i>	<b>legacy</b>	<b>lemon</b> lemon peel lemon rind	<b>light</b> light bodied light rum



<b>light filtration</b> → gentle filtration	<b>light pot still</b> light pot still rum → heavy pot still	<b>lime</b> lime peel → lime rind lime rind → lime peel	<b>limited edition</b> → limited reserve
<b>limited number</b> → limited production	<b>limited production</b> → limited number	<b>limited reserve</b> → limited edition	<b>liqueur</b> <i>liquor</i>
<b>long</b>	<b>long drink</b>	<b>low wine retort</b> → high wine retort	<b>making process</b>
<b>manufacturing process</b>	<b>maple</b>	<b>margarita glass</b>	<b>market</b> European market local market mass market mid-market top market US market
<b>marrying process</b>	<b>mash</b>	<b>master blender</b>	<b>master distiller</b>
<b>maturation process</b>	<b>medal</b>	<b>medium</b> medium bodied medium rum	<b>mellow</b>
<b>milled</b>	<b>minimum aged rum</b>	<b>mixed</b> mixed drink	<b>mixing glass</b>
<b>mixing rum</b>	<b>molasses</b>	<b>naturally filtered</b>	<b>navy neaters</b>
<b>neat</b>	<b>Nelson's blood</b>	<b>nose</b> → aroma	<b>note</b>
<b>nut</b> nutty toasted nut	<b>nutmeg</b>	<b>oak</b> oaky	<b>oak barrel</b> → oak cask
<b>oak cask</b> → oak barrel	<b>old</b> old rum	<b>old-fashioned glass</b> → rocks glass	<b>orange</b> orange peel → orange rind orange rind → orange peel
<b>organic compound</b>	<b>original</b>	<b>overproof rum</b>	<b>oxidation</b>
<b>packaging</b> packaging detail	<b>painkiller</b>	<b>palate</b>	<b>part</b>
<b>passion fruit</b>	<b>pastry</b>	<b>peach</b>	<b>pepper</b>
<b>peppermint</b>	<b>pirate's drink</b>	<b>plant</b>	<b>plantation</b> plantation distillery plantation rum
<b>platinum</b> → PT platinum rum	<b>pot</b>	<b>pot still</b> → continuous still pot still rum	<b>pot still distillation</b> → batch distillation
<b>premium</b> premium rum	<b>primary water treatment system</b> → secondary water treatment system	<b>production capacity</b>	<b>product line</b> → production line
<b>production line</b> → product line	<b>product range</b> → brand positioning	<b>profile</b>	<b>PT</b> → platinum

<b>quality rum</b>	<b>raisin</b> honeyed raisin ripe raisin sweet raisin	<b>raw</b>	<b>recipe</b>
<b>recovery column</b>	<b>red</b>	<b>reduction</b>	<b>reserve</b>
<b>residual impurity</b>	<b>rich</b>	<b>rim</b>	<b>rocks glass</b> → old-fashioned glass
<b>rounded</b> rounded rum	<b>rum</b>	<b>rumbullion</b>	<b>rumbustion</b>
<b>rum making</b> rum making process	<b>rum steam</b>	<b>saccharomyces</b>	<b>Savalle still</b>
<b>scent</b>	<b>seal</b>	<b>secondary water treatment system</b> → primary water treatment system	<b>select</b> <i>selected</i>
<b>sherry cask</b>	<b>short glass</b>	<b>shot glass</b>	<b>signature drink</b>
<b>signature rum</b>	<b>single barrel</b> → single cask	<b>single cask</b> → single barrel	<b>single distillation</b> single distillate single distilled rum
<b>single-label</b>	<b>single rum</b>	<b>sipping rum</b> → tasting rum	<b>size</b> → capacity
<b>smoky</b>	<b>smoothness</b> smooth	<b>soft</b>	<b>soil</b>
<b>sophisticated</b> → complexity	<b>spice</b> → spiced	<b>spiced</b> <i>spicy</i> → spice spiced rum <i>spice rum</i>	<b>spirit</b> spirity
<b>stalk</b>	<b>steam</b> steam engine	<b>still</b> still maturation	<b>storage</b> storage container storage facility storage tank
<b>straight</b> straight rum	<b>strain</b>	<b>strength</b>	<b>subtle</b>
<b>sugar</b> sugar factory	<b>sugar cane</b> <i>sugarcane</i> sugar cane juice <i>sugarcane juice</i> sugar cane plantation <i>sugarcane plantation</i>	<b>sulphate</b>	<b>sultana</b>
<b>superior</b> superior rum	<b>super premium</b> super premium rum	<b>sweetness</b> sweet	<b>taffia</b> <i>tafia</i>
<b>tail</b>	<b>tall glass</b>	<b>taste</b> → flavour <i>flavor</i> taste profile	<b>tasting note</b>
<b>tasting rum</b> → sipping rum	<b>terroir</b>	<b>texture</b>	<b>toasted</b>
<b>tobacco</b>	<b>toffee</b>	<b>tot</b>	<b>triple distillation</b> triple distillate triple distilled rum
<b>tropical ageing</b> <i>tropical aging</i> tropically aged	<b>tropical fruit</b>	<b>uncrystallized sugar</b>	<b>vanilla</b>

<b>velvety</b>	<b>versatile</b> versatile rum	<b>vibrant</b>	<b>vintage</b> vintage blend vintage rum
<b>vol.</b> → ABV → alcohol by volume	<b>volatile sulphur compound</b>	<b>wash</b>	<b>water treatment system</b>
<b>West Indian rum</b>	<b>West Indies Rum &amp; Spirits Producers' Association</b> → WIRSPA	<b>wheat bread</b>	<b>white</b> white rum
<b>WIRSPA</b> → West Indies Rum & Spirits Producers' Association	<b>wood</b> toasted wood woody	<b>wooden distillation</b>	<b>World Spirits Competition</b> → WSC
<b>WSC</b> → World Spirits Competition	<b>XO</b> → extra old	<b>yeast</b> yeast strain	<b>zest</b>

## 7 Conclusion

The initial steps of the “implementation” stage (Svensén 2003: 99) of a specialized glossary of rum – limited to the English language – were described. All combined, the compilation of the *CRC* and the *GRC*, the application of corpus-based term-extraction procedures, the exploitation of experts' knowledge through fieldwork, the analysis of specialized literature and the subsidy of the lexicographer's insight proved fruitful. Consequently, among thousands of candidate items, 324 headwords and 87 sub-headwords were eventually considered for inclusion, thus accomplishing the main goal of this piece of research. At a later stage, it will be possible to move towards the microstructure of the glossary, that is the editing of each entry: a more detailed treatment of headwords – not yet produced – implies that all entries will be provided with a definition, instances of usage in authentic texts and, where necessary, especially in cases that require the illustration of highly specialized appliances used in rum distillation, images will be added – at present, since this article is mostly a work-in-progress report on a glossary-making project, the plans to make the resource available and the strategy for distributing it through the appropriate channels are not yet underway.

The procedures implemented in this pilot study focusing on the macrostructure of a specialized glossary of rum-related terms, meant to be the starting point for the compilation of a specialised glossary of rum, seem generalizable. The same methodology, possibly complemented by the application of fully-automatic term extractors (see footnote 10), may indeed be replicated to enable the compilation of specialized glossaries connected to other subjects or domains, as already successfully attempted, among others, by Gamper and Stock (1998), Cabré Castellví (1999); Bourigault, Jacquemin and L'Homme (2001), Peñas, Verdejo and Gonzalo (2001), and Chung (2003).

## 8 Desiderata

The ubiquitous nature of rum, especially its popularity throughout the French- and Spanish-speaking Caribbean in addition to the English-speaking Caribbean, naturally calls for a further, more ambitious project, that is the compilation of a multilingual glossary of rum. Following the same criteria adopted for the selection of rum-related terminology in the English language, a French and

Spanish supplement should be considered in order to appeal to the worldwide audience of rum enthusiasts.

As for French, after compiling an analogous specialized corpus based on texts retrieved from the websites of rum distillers based in the French-speaking Caribbean, as well as Madagascar, Mauritius, Réunion and Seychelles, the wordlist provided may be compared to the wordlist produced from the *Corpus français* (CF) or the forthcoming *Corpus de référence du français contemporain* (CRFC), to be considered as reference corpora of general French (see Siepmann, Bürgel & Diwersy 2015: 64). As far as Spanish is concerned, the wordlist obtained from the specialized corpus gathered by extracting texts from the websites of rum makers in the Spanish-speaking Caribbean, as well as in Central and South America and the Canaries, may be set against the wordlist triggered by the *Corpus de referencia del español actual* (CREA), a general corpus of the Spanish language. After implementing the appropriate semi-automatic procedures, applying automatic term extractors and including the pertinent specialized literature written in French and Spanish respectively, the keywords obtained would lead to the drafting of a French and Spanish list of candidate headwords, thus collocating the content of the glossary in a multilingual perspective.

## References

- 10 Cane. Accessed at: <http://www.10cane.com> [25/05/2018]  
 [AIS] Associazione Italiana Sommelier Accessed at: <http://www.aisitalia.it> [25/05/2018]  
 Angostura. Accessed at: <http://www.angostura.com> [25/05/2018]  
 Anguilla Rums. Accessed at: <http://www.pyratrum.com> [25/05/2018]  
 [AntConc] Anthony, L. (2018). *AntConc* 3.5.7. Tokyo: Waseda University. Accessed at: <http://www.laurenceanthony.net/software/antconc/> [25/05/2018]  
 Antigua Distillery. Accessed at: <http://antiguadistillery.com> [25/05/2018]  
 Appleton Estate. Accessed at: <http://www.appletonestate.com> [25/05/2018]  
 Arkell, J. (1999). *Classic Rum*. London: Prion Books.  
 Arundel Estate Callwood Distillery. Accessed at: [http://www.bareboatsbvi.com/cgb\\_callwood\\_distillery.html](http://www.bareboatsbvi.com/cgb_callwood_distillery.html) [25/05/2018]  
 Atkins, B. T. S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.  
 Bacardi. Accessed at: <http://www.bacardi.com> [25/05/2018]  
 Barty-King, H., Massel, A. (1983). *Rum: Yesterday and Today*. London: Heinemann.  
 Bergenholtz, H. (1995). Basic Issues in Specialized Lexicography. In H. Bergenholtz, S. Tarp (eds.) *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam & Philadelphia: John Benjamins, pp. 14-47.  
 Biber, D. (2008). Representativeness in Corpus Design. In T. Fontenelle (ed.) *Practical Lexicography: A Reader*. Oxford: Oxford University Press, pp. 63-87.  
 Blackwell. Accessed at: <http://www.blackwellrum.com> [25/05/2018]  
 [BNC] Davies, M. (ed.). (2004-2018). *British National Corpus*. Provo: Brigham Young University. Accessed at: <http://corpus.byu.edu/bnc> [25/05/2018]  
 Bourigault, D., Jacquemin, C. & L'Homme, M.-C. (eds.). (2001). *Recent Advances in Computational Terminology*. Amsterdam & Philadelphia: John Benjamins.  
 Bowker, L. (2003). Specialized Lexicography and Specialized Dictionaries. In P. Van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Amsterdam & Philadelphia: John Benjamins, pp. 154-164.  
 Bowker, L., Pearson, J. (2002). *Working with Specialized Language. A Practical Guide to Using Corpora*. London & New York: Routledge.  
 Broom, D. (2003). *Rum*. San Francisco: Wine Appreciation Guild.  
 Captain Morgan. Accessed at: <http://www.captainmorgan.com> [25/05/2018]  
 Cabré Castellví, M. T. (1999). *Terminology: Theory, Methods and Applications*. Amsterdam & Philadelphia: John Benjamins.

- [CF] *Corpus français*. Leipzig: Universität Leipzig & Neuchâtel: Université de Neuchâtel. [http://wortschatz.uni-leipzig.de/ws\\_fra](http://wortschatz.uni-leipzig.de/ws_fra) [25/05/2018]
- Chung, T. M. (2003). A Corpus Comparison Approach for Terminology Extraction. In *Terminology*, 9(2), pp. 221-246. *Clarke's Court*. Accessed at: <http://www.clarkescourttrum.com> [25/05/2018]
- [COCA] Davies, M. (ed.). (2008-2018). *Corpus of Contemporary American English*. Provo: Brigham Young University. Accessed at: <http://corpus.byu.edu/coca> [25/05/2018]
- Cockspur*. Accessed at: <http://www.cockspurrum.com> [25/05/2018]
- Coruba*. Accessed at: <http://www.coruba.co.nz> [25/05/2018]
- Coulombe, C. A. (2004). *Rum: The Epic Story of the Drink that Conquered the World*. New York: Citadel Press.
- [CREA] *Corpus de referencia del español actual*. Madrid: Real Academia Española. Accessed at: <http://corpus.rae.es/creanet.html> [25/05/2018]
- Curtis, W. (2007). *And a Bottle of Rum: A History of the New World in Ten Cocktails*. New York: Three Rivers Press.
- [DCEU] Allsopp, R. (ed.). (2003) [1996]. *Dictionary of Caribbean English Usage*. Mona: University of the West Indies Press.
- [DECH] Corominas, J., Pascual, J. A. (1983). *Diccionario crítico etimológico castellano e hispánico*. Madrid: Gredos.
- Demerara Distillers*. Accessed at: <http://demeraradistillers.com> [25/05/2018]
- [DRAE] (2001). *Diccionario de la lengua española*, 22<sup>nd</sup> edn. Madrid: Real Academia Española. <http://lema.rae.es/drae> [25/05/2018]
- El Dorado Rum*. Accessed at: <http://theeldoradorum.com> [25/05/2018]
- Elements Eight Rum*. Accessed at: <http://www.e8rum.com> [25/05/2018]
- [FEW] Von Wartburg, W. (2003). *Französisches Etymologisches Wörterbuch*. Accessed at: <http://apps.atilf.fr/lecteurFEW> [25/05/2018]
- [FLOB] Hundt, M., Sand, A. & Siemund, R. (eds.). (1998). *Freiburg-Lancaster-Oslo-Bergen Corpus of British English*. Freiburg: Albert-Ludwigs-Universität Freiburg. Accessed at: <http://icame.uib.no/flob> [25/05/2018]
- Foss, Richard. (2012). *Rum: A Global History*. London: Reaktion Books.
- Foursquare Rum Distillery*. Accessed at: <http://foursquarerum.com> [25/05/2018]
- [FROWN] Hundt, M., Sand, A. & Skandera, P. (eds.). (1999). *Freiburg-Brown Corpus of American English*. Freiburg: Albert-Ludwigs-Universität Freiburg. Accessed at: <http://icame.uib.no/frown> [25/05/2018]
- Furiassi, C. (2004). Spoken and Written Learner English: A Quantitative Analysis of ICLE-IT and LINDSEI-IT. In M. T. Prat Zagrebelsky (ed.) *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learner's Interlanguage*. Alessandria: Edizioni dell'Orso, pp. 193-208.
- Furiassi, C. (2014). Caribbean English Vocabulary: Setting a Norm through Lexicographic Practice. In A. Molino, S. Zanotti (eds.) *Observing Norm, Observing Usage: Lexis in Dictionaries and in the Media*. Bern: Peter Lang, pp. 89-107.
- Gamper, J., Stock, O. (1998). Corpus-based Terminology. In *Terminology*, 5(2), pp. 147-159.
- Gosling*. Accessed at: <http://www.goslingsrum.com> [25/05/2018]
- Gotti, M. (2011). *Investigating Specialized Discourse*, 3<sup>rd</sup> edn. Bern: Peter Lang.
- Grand Havana Rum*. Accessed at: <http://www.grandhavanarum.com> [25/05/2018]
- Hartmann, R. R. K., James, G. (2002). *Dictionary of Lexicography* London & New York: Routledge.
- Hopkins, T. (2012) [2004]. Rum. In A. F. Smith (ed.) *The Oxford Encyclopedia of Food and Drink in America*, 2<sup>nd</sup> edn. Oxford: Oxford University Press, vol. II, pp. 158-160.
- Krishnamurthy, R. (2008). Corpus-driven Lexicography. In *International Journal of Lexicography*, 21(3), pp. 231-242.
- Laurie, P. (2011) [2001]. *The Barbadian Rum Shop: The Other Watering Hole*, 2<sup>nd</sup> edn. Oxford: Macmillan.
- Leroyer, P. (2015). Turning the Corpus into a Functional Component of the Dictionary: The Case of the Oenolex Wine Dictionary. In *Procedia – Social and Behavioral Sciences*, 198, pp. 257-265.
- Leroyer, P. (2018). The Oenolex Wine Dictionary. In P. A. Fuertes-Olivera (ed.) *The Routledge Handbook of Lexicography*. London & New York: Routledge, pp. 438-454.
- Liberman, A. (2010). The Rum History of the Word "Rum". In *OUPblog*, 6<sup>th</sup> October 2010. Accessed at: <https://blog.oup.com/2010/10/rum> [25/05/2018]
- Maier, E. (2013). *OED Word Stories: 'rum'*. Accessed at: <http://public.oed.com/aspects-of-english/word-stories/rum> [25/05/2018]



- [Merriam-Webster] Gove, P. B. (ed.). (2002). *Webster's Third New International Dictionary Unabridged*. Springfield: Merriam-Webster. Accessed at: <http://unabridged.merriam-webster.com> [25/05/2018]
- Miller, A., Brown, J., Broom, D. & Strangeway, N. (2009). *Cuba: The Legend of Rum*. London: Mixellany. *Mount Gay Distillers*. Accessed at: <http://www.mountgayrum.com> [25/05/2018]
- Myers's. Accessed at: <https://www.diageo.com/en/our-brands/brand-explorer/#myers> [25/05/2018]
- [OED] Simpson, J., Weiner, E. (eds.). (1989-2018). *The Oxford English Dictionary*. Oxford: Oxford University Press. Accessed at: <http://www.oed.com> [25/05/2018]
- OneClick Terms. (2016-2018). Brno & Brighton: Lexical Computing. Accessed at: <https://terms.sketchengine.co.uk> [25/05/2018]
- Peñas, A., Verdejo, F. & Gonzalo, J. (2001). Corpus-based Terminology Extraction Applied to Information Access. In P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster: UCREL, pp. 458-465.
- Plotkin, R. A. (2001). *Caribe Rum. The Original Guide to Caribbean Rum and Drinks*. Tucson: BarMedia. *Pusser's*. Accessed at: <http://www.pussers.com> [25/05/2018]
- Ringbom, H. (1998). Vocabulary Frequencies in Advanced Learner English: A Cross Linguistic Approach. In S. Granger (ed.) *Learner English on Computer*. London & New York: Longman, pp. 41-52.
- Ruthström, B. (2001). *Tafia, ratafia and rum – liquor words of dizzy origin*. In *Indogermanische Forschungen*, 106(1), pp. 262-275.
- Siepmann, D., Bürgel, C. & Diwersy, S. (2015). The *Corpus de référence du français contemporain (CRFC)* as the first genre-diverse mega-corpus of French. In *International Journal of Lexicography*, 30(1), pp. 63-84.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2003). Corpora for Lexicography. In P. Van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Amsterdam & Philadelphia: John Benjamins, pp. 167-178.
- [Sketch Engine] Kilgariff, A. (2003-2018). *Sketch Engine*. Brno & Brighton: Lexical Computing. Accessed at: <http://www.sketchengine.co.uk> [25/05/2018]
- Smiley, I, Watson, E. & Delevante, M. (2014). *The Distiller's Guide to Rum*. Hayward: White Mule Press.
- Smith, F. H. (2008) [2005]. *Caribbean Rum. A Social and Economic History*, 2<sup>nd</sup> edn. Gainesville: University Press of Florida.
- St. Lucia Distillers*. Accessed at: <http://www.saintluciarums.com> [25/05/2018]
- St. Nicholas Abbey*. Accessed at: <http://www.stnicholasabbey.com> [25/05/2018]
- Svensén, B. (2003). Dictionary Projects. In R. R. K. Hartmann (ed.) *Lexicography: Critical Concepts*. London & New York: Routledge, vol. I, pp. 97-108.
- [TermoStat] Drouin, P. (2010-2018). *TermoStat*. Montréal: Université de Montréal – Observatoire de linguistique Sens-Texte. Accessed at: <http://termostat.ling.umontreal.ca> [25/05/2018]
- [TextSTAT] Hüning, M. (2015). *TextSTAT 3.0*. Berlin: Freie Universität Berlin. Accessed at: <http://neon.niederlandistik.fu-berlin.de/en/textstat/> [25/05/2018]
- [TLFi] (1994). *Le trésor de la langue française informatisé*. Paris: CNRS editions. Accessed at: <http://atilf.atilf.fr> [25/05/2018]
- Todhunter-Mitchell Distillery*. Accessed at: <http://www.burnshouse.com> [25/05/2018]
- Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.
- Tognini Bonelli, E., Sinclair, J. (2006) [1993]. Corpora. In K. Brown (ed.) *Encyclopedia of Language & Linguistics*. Boston: Elsevier, vol. III, pp. 206-219.
- Westerhall Estate*. Accessed at: <http://www.westerhallrums.com> [25/05/2018]
- Williams, I. (2006) [2005]. *Rum: A Social and Sociable History of the Real Spirit of 1776*, 2<sup>nd</sup> edn. New York: Nation Books.
- [WordSmith Tools] Scott, M. (2018). *WordSmith Tools 7.0*. Liverpool: Lexical Analysis Software. Accessed at: <http://www.lexically.net/wordsmith> [25/05/2018]
- Worthy Park Estate*. Accessed at: <http://www.worthyparkestate.com> [25/05/2018]
- XM*. Accessed at: <http://www.xmrumguyana.com> [25/05/2018]
- Zaya*. Accessed at: <http://www.infiniumspirits.com> [25/05/2018]

## Acknowledgements

The author would like to thank Jeannette Allsopp and Jason Siegel (*University of the West Indies* at Cave Hill, Barbados) for their careful revision of the manuscript and advice on key bibliographic material. Thanks are also due to Marek Łukasik (*Pomeranian University* in Słupsk, Poland) for his valuable comments and precious suggestions on the methodological implant. The author is grateful to Cristina Castielli, a student at the *University of Turin* (Italy), who contributed to the compilation of the *Caribbean Rum Corpus (CRC)* as part of her MA thesis. A final word of thanks goes to the personnel and tour guides at *Mount Gay Distillers* visitor center (St. Michael, Barbados) and distillery (St. Lucy, Barbados), *St. Nicholas Abbey* (St. Peter, Barbados) and *Foursquare Rum Distillery* (St. Philip, Barbados). Cristiano Furiassi's investigation was made possible by a three-month research stay – from September 2015 to December 2015 – at the *Richard and Jeannette Allsopp Centre for Caribbean Lexicography* of the *University of the West Indies* at Cave Hill, Barbados, jointly sponsored by the *University of Turin* and *Fondazione CRT* through the second edition of the *World Wide Style (WWS)* fellowship. In May 2016 Cristiano Furiassi was also awarded a certified sommelier diploma by the *Associazione Italiana Sommelier (AIS)*, which provided him with more thorough knowledge of world spirits, including rum.

# Russian Borrowings in Greek and Their Presence in Two Greek Dictionaries

**Zoe Gavriilidou**

*Democritus University of Thrace*

*E-mail: zoegab@otenet.gr*

## Abstract

This paper focuses on Russian loanwords, loanblends and loanshifts that entered the Greek lexicon during various historical periods and how they have been recorded into two major dictionaries of Modern Greek (MG), the *Dictionary of Standard Modern Greek* (DSMG) (1998) and the *User's Dictionary* of the Academy of Athens (UD) (2014). It also aims at the analysis of the semantic fields of all documented Russian borrowings in the history of Greek, following a classification scheme which was originally used for typological comparison (Haspelmath & Tadmor 2009c). In the first part of the paper, we consider the contact situations that led to borrowing from Russian into Greek and the reasons for borrowing, which include, among others (a) response to major political events, such as the October 1917 Revolution, the Soviet era, the 1987 Perestroika; (b) literary translations of Russian masterpieces in Greek and Greek literature, such as the work of Nikos Kazantzakis or Miltiadis Karagatsis; and (c) religious affinity. Then we compare how these borrowings are recorded in DSMG and UD. In the next section, we offer a morphophonological analysis of borrowings. The semantic fields in which the borrowings belong to are also studied. Finally, the paper provides experimental data for supporting Anastassiadis's (1994) claim that lexical fields, in which loanwords abound, reflect a stereotypic image of the country where the donor language is spoken.

**Keywords:** loanword, borrowing, loanshift, loanblend, internationalism, loan translation, calque, structural borrowing

## 1 Introduction

Borrowing is an interesting phenomenon of language contact which leads to language change and has been extensively studied in the recent literature (Thomason & Kaufman 1992; Thomason 2001; Johanson 2002; Haspelmath 2008; Haspelmath & Tadmor 2009). It sometimes reveals the type of relations between people speaking the donor and the recipient language, reflects the stereotypes established in a given culture about the 'other civilization' and "symbolizes the foreign and the strange" (Stubbs 1998: 19).

Previous research on language contact and borrowing focuses on:

- a) The typology of borrowings (Haugen 1950; Humbley 1990; Anastassiadis 1994; Matras & Sakel 2007);
- b) The reasons (cultural influence, historical events, stereotypes, denomination needs) that motivate borrowing (Haspelmath 2008);
- c) The type or intensity of linguistic contact that leads to borrowing (Thomason & Kaufman 1992);
- d) The parts of speech that are more easily borrowed among languages (Van Hout & Muysken 1994; Myers-Scotton 2002; Matras 2007);
- e) The borrowability scales (Matras 1998);
- f) The connection between borrowing and lexical meaning (Haspelmath & Tadmor 2009);
- g) The synchronic or diachronic analysis of loanwords;
- h) The adaptation and inclusion of borrowings in the receiving language.

There is also a large amount of previous research on lexical borrowing from English or French in Greek (Contossopoulos 1978; Apostolou-Panara 1991; Anastassiadis 1994), however no previous study has focused on borrowings from Russian into Greek, even though the phenomenon is not peripheral. Thus, the purpose of this paper is to investigate Russian borrowings that entered Greek lexicon and also shed light to the reasons for borrowing and account for the social and attitudinal factors that affect it. Our aim is also to investigate how these borrowings are included in two major dictionaries of Modern Greek (MG), the *Dictionary of Standard Modern Greek* (DSMG) (1998) and the *User's Dictionary* of the Academy of Greece (UD) (2014).

In the first part of the paper, after a brief presentation of the borrowing classification adopted in the present study, we consider the contact situations that led to borrowing from Russian into Greek and the reasons for borrowing. Then, we present the data collected from the two above-mentioned dictionaries: More specifically, a) We classify Russian borrowings following the typology of Haugen (1950) and Anastassiadis (1994) and provide quantitative data. b) Furthermore, we elaborate on their phonological and morphological adaptation and integration into Greek. c) The semantic description of Russian borrowings is also investigated. d) Then, we identify frequent Russian loanwords that are absent from the macrostructure of DSMG and UD, and propose new entries for these words to be included in the macrostructure of these dictionaries. e) Finally, we investigate, through a brainstorming experiment, the stereotypes that Greek speakers have with regard to Russian civilization and Russians, in order to provide experimental data for supporting Anastassiadis' (1994) claim that lexical fields, in which loanwords abound, reflect an image of the country where the donor language is spoken.

In this study, the term *borrowing* is used to refer to “the incorporation of foreign elements into the speakers' native language” (Thomason & Kaufman 1992). Additionally, we will refer to the language from which a loanword has been borrowed as *donor language* or L2, and the language into which it has been borrowed as the *recipient language* or L1. Finally, borrowing is a historical dimension, which can be studied if information on the linguistic diachrony of the involved languages is available. Therefore, in this paper we will use the methods of diachronic linguistics. However, we will perform a synchronic analysis of the output of borrowing from Russian to Greek.

## 2 Theoretical Issues

One of the best known typologies of lexical borrowing, adopted in the present study, is that of Haugen (1950), who distinguishes among *loanwords*, *loanblends* and *loanshifts*. *Loanwords* are words that, at some point in the history of a language, entered its lexicon as a result of language contact (e.g. καπίκι [capiki] ‘kopeck’). A specific category of loanwords are *internationalisms*, which are loanwords that entered simultaneously in different recipient languages (e.g. Perestroika). Loanwords belong to what is known as *material borrowing* (Matras & Sakel 2007).

*Loanblends*, on the other hand, are words constructed in the recipient language by a native and a foreign part (e.g. γυφτέ [jifte] (gipsy-like) from the word γύφτος ‘gipsy’ and the borrowed from French suffix –έ). Finally, *loanshifts* include *loan translations* or calques (e.g. αναθεωρητισμός [anaθeoritizmos] ‘revisionism’), which are complex lexical units, either monolexical or polylexical, that are created by item-by-item translation of the source-term, and *semantic borrowings* (e.g. the new meaning ‘political officer of the communist party’ that was added in the medieval word κοιμισάριος [komisarios] ‘commissary’). Semantic borrowings are borrowings of the *signified* of a word of the donor language (L2) that attaches to a semantic field of an already existing word in the recipient language (L1). Loan translations and semantic borrowings are indirect or partial borrowings (Humbley 1990) and belong to what is known in the literature as *structural borrowing* (Matras & Sakel 2007).

### 3 Linguistic Contact Between Greek and Russian

Borrowings are connected with the history of a nation and its relations with others more closely than any other part of the lexical inventory. Furthermore, the duration and intensity of language contact, the cultural or linguistic affinity between the L1 and L2, political, religious or other types of bond between people speaking L1 and L2, the roles and status of these languages, the attitudes and stance of people speaking the recipient language towards those speaking the donor language or other sociolinguistic factors play a vital role in determining the degree and outcome of borrowing between L1 and L2. Thus, only a complex study of linguistic items and historical events can facilitate the answering of important questions, such as the time of introduction of a borrowing from L2 to L1, the donor and intermediary language or any changes in form and meaning that may happen.

Greek has integrated in its vocabulary Russian borrowings. In order to understand the process of borrowing, one must consider the following facts: Greek and Russian coexisted in a bilingual context from the end of the 1980's and after in Northern Greece in bilingual communities of people repatriated to Greece from the former Soviet Union. However, during that period Russian did not leave any noteworthy traces in Greek. Consequently, there was no direct contact between Greek and Russian that motivated the introduction of Russian loanwords, but only an indirect relation between the two languages. This relation is reciprocal, since Greek as a donor language, in the past, gave more loanwords to Russian, especially in the ecclesiastical or everyday vocabulary, than it borrowed from Russian. During the 20<sup>th</sup> century, Greek borrowed from Russian mainly due to:

- a) Historical or political events, such as the October Revolution which in 1917 established the Soviet regime in Russia, or Perestroika in the late 80's. As Stubbs (1998) claims, words embody facts of history and are often borrowed into a language in response to world political events;
- b) Literary translations of Russian masterpieces (e.g. the work of Pushkin or Dostoyevsky) in Greek, which introduced words referring to culturally bound terms such as ντάτσα [datsa] 'dacha', ίζμπα [izba] 'isba', ουσάνκα [usanka] 'ushanka', βότκα [votka] 'vodka';
- c) The publication of literary masterpieces of Greek authors such as *Russia: A Chronicle of Three Journeys in the Aftermath of the Revolution* (1928) by Nikos Kazantzakis or *Junkermann* (1939) by Miltiadis Karagatsis, which also introduced everyday vocabulary from Russian;
- d) Terms referring to Russian technology such as σπούτνικ [sputnik] 'sputnik', κοσμοναύτης [kosmonaftis] 'cosmonaut', etc.

Russian borrowings were introduced in Greek either directly through literary translations (e.g. βότκα [votka] 'vodka', γιάφκα [jafka] 'javka', καπίκι [kapici] 'kopeck', πιροσκι [piroski] 'piroschki', μπαλαλάικα [balalaika] 'balalaika'), or indirectly with the intermediate of French (e.g. ινστρούχτορας [instruxtoras] 'instructor', κολεκτιβισμός [kolektivizmos] 'collectivism', πογκρόμ [pogrom] 'pogrom'), English (e.g. περεστρόικα [perestroika] 'perestroika', μολότοφ [molotof] 'molotov', σπούτνικ [sputnik] 'sputnik', σφυροδρέπανο [sfirodrepano] 'hammer and sickle'), or rarely from Turkish (τελατίνι [telatini] 'veal skin').

In the following section, we will focus on how these borrowings are recorded in the macrostructure of two major dictionaries of Modern Greek, the *Dictionary of Standard Modern Greek* and the *User's Dictionary* of the Academy of Athens.

### 4 Russian Loanwords in DSMG and UD

#### 4.1 The DSMG

The *Dictionary of Standard Modern Greek* (DSMG) is a modern and comprehensive definitional, orthographic, and etymological dictionary of Modern Greek. It was published in December 1998 by the



Institute for Modern Greek Studies of the Aristotle University of Thessaloniki, and is the product of many years of methodical labor. It is the first dictionary of Modern Greek to set forth lexicographical principles. It was first released in 1998 in paper form and then as an online application available at [http://www.greek-language.gr/greekLang/modern\\_greek/tools/lexica/triantafyllides/](http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/).

By performing a number of advanced searches in the online DSMG, sixty-five borrowings of Russian were retained. All words including the label ‘Russian’ as language of origin in the field of etymology or in the lexicographic article were considered. It was noted that DSMS provided the primary etymon for each entry and also the transition of the term – via an intermediate language – from the donor language to Greek. This principle was applied systematically to all entries.

The majority of these were loanwords (58 out of 65) (e.g. *σαμοβάρι* [samovari] ‘samovar’, *κεφίρ* [kefir] ‘kefir’, *αταμάνος* [atamanos] ‘ataman’, *κνούτο* [knuto] ‘knot, whip’, *κοσμοδρόμιο* [kozmodromio] ‘cosmodrome’, *κουλάκος* [kulakos] ‘kulak’, *μαζούτ* [mazut] ‘mazut’, *μαμούθ* [mamuθ] ‘mammoth’. Among them there were a lot of internationalisms (e.g. *βότκα* [votka] ‘vodka’, *περεστρόικα* [perestroika] ‘perestroika’). The loan translations were less frequent (nine out of 65): e.g. *ερυθροφρουρός* [erithrofruros] ‘red guard’, *σφυροδρέπανο* [sfirodrepano] ‘hammer and sickle’, *Λευκορωσία* [lefkorosia] ‘Byelorussia’, *υπερσιβηρικός* [ipersivirikos] ‘Trans-Siberian’, etc. The majority of loan translations found in DSMG macrostructure entered Greek through the intermediate of French language, except for the word *σφυροδρέπανο* [sfirodrepano] ‘hammer and sickle’ which entered Greek vocabulary through English.

In some cases, in the dictionary’s word list parallel couples of loanwords and loan translations with the same meaning were attested in the DSMG: e.g. *ρεβιζιονιστής* [revizionistis] ‘revisionist’ vs. *αναθεωρητής* [ana theoritis] ‘revisionist’. As loan translations are more transparent and adapted in Greek language, they are preferred by Greek speakers.

Semantic borrowings from Russian were extremely rare in DSMG: characteristic examples are the medieval word *κομισάριος* [komisarios] ‘official’ and the savant word *επίτροπος* [epitropos] ‘commissioner’ in which the new meaning of ‘political officer of the communist party’ was added in the 20<sup>th</sup> century.

No loanblends were found in DSMG. Actually some occasional, ludic creations in –vski, –ov and –its found in literature or advertisements are never included in dictionaries.

The above mentioned data confirm that both material and structural borrowing occurred from Russian to Greek, although material borrowing is much more frequent than the structural one. This is in line with previous research which found that at the lowest level in the borrowability scale borrowing is limited to the lexical level, and mainly to content words. Structural borrowing is only found at the higher levels. According to the scale, the existence of structural borrowing in a language generally implies that words have also been borrowed (Matras 1998, 2011).

Fifty-eight (58) out of sixty-five (65) borrowings found in DSMG were nouns, five (5) were adjectives and only two (2) verbs. This finding is in line with previous research on borrowability scales (Van Hout & Muysken 1994; Myers-Scotton 2002; Matras 2007). Van Hout and Muysken (1994: 42) account for the greater ease of nouns than verbs to be integrated in other languages by stating that “A very important factor involves one of the primary motivations for lexical borrowing, that is, to extend the referential potential of a language. Since reference is established primarily through nouns, these are the elements borrowed most easily”. Matras (2007: 48) claims that the difficulty of verbs to be integrated in another language

“lies in the conceptual complexity of the verb, and the fact that when borrowed and integrated, the verb is expected to perform two operations: The first is to serve as a referential lexical

item – a content word, not dissimilar to a noun, adjective, or descriptive adverb. The second is to initiate the predication and so to serve as the principal anchor point for the entire proposition of the utterance. This latter function constitutes its *verbness*. It appears that borrowing of verbs is motivated by a similar need for modifying the inventory of lexical-referential expressions as the borrowing of nouns (and no doubt various specific semantic motivations could be postulated for groups of lexical content words). Speakers thus allow the lexical component of the verb to “cross” the mental demarcation boundary between languages, i.e. they license themselves to employ the same action/ event signifier in any speech interaction. The bare lexical stem, however, is not always sufficient in order to assume the role of predication-initiator. A great number of languages therefore require this additional, crucial function to be explicitly marked out in the verbal expression; in other words, they need to transform the strictly “lexical” depiction of an action/event into a predicate”.

## 4.2 The UD

The *User's Dictionary* of the Academy of Athens (UD) is a user's definitional, orthographic, and etymological paper dictionary of Modern Greek. It was published in 2014 by the Academy of Athens, and is the product of almost ten years of compilation. It includes 75,000 entries, 5,000 neologisms, and more collocations than any other Greek dictionary.

The UD includes only 38 words with Russian as language of origin in the field of etymology. All 38 are direct loanwords; no loan translations from Russian are attested in UD, because this dictionary considers English as the donor language of all loan translations after 1950 and French as the donor language of loan translations before this. Actually, UD includes in its entry list only direct loanwords from Russian, while DSMG includes both direct and indirect borrowings (loanwords, calques or semantic loans) providing the initial etymon in Russian. Thus the etymon of the entry *μενσεβίκος* [mensevikos] ‘Menshevik’ in the UD is the French word *Menshevik*, while in DSMG the etymon is the Russian word *men'shevik*. This discrepancy in the lexicographic practice between the two dictionaries reflects a methodological difference in the description of word origin, and explains why more words in DSMG are characterized of Russian origin than in UD.

From the thirty-eight borrowings from Russian in UD, thirty-six are nouns and only two adjectives.

## 4.3 Comparing DSMG and UD Entries

Definitions of the common entries in the two dictionaries are not divergent. This is probably due to the use of more or less the same textual sources. Additionally, no variability in spelling of the head entries was attested between DSMG and UD.

Differences were mainly found in the details in the etymology part; DSMG provides more exhaustive etymological information than UD, while UD includes in some cases, parallel types in other languages with a chronology of the first appearance of these words (e.g. in the etymology part of the word *ματριόσκα* [matrioska] the user can read [< Rus. matrëshka, Engl. matrioshka, 1964]). The etymology part of both dictionaries should be ameliorated by incorporating information on word-forms and meanings from donor languages, together with dates of attestation in those languages wherever possible, in order to provide a much fuller picture of the process of integration of individual words into Greek. These sources will allow users to realize the systematicity in the borrowing process and identify contemporaneous borrowings into other languages in Europe, showing that Greek is part of a network of languages which share in the process of borrowing and semantic development.

As far as coverage is concerned, from the sixty-eight Russian borrowings recorded in DSMG, the following eleven are absent from the UD macrostructure:

αγκιτάρω [agitaro] ‘agitate’  
 αταμάνος [atamanos] ‘ataman’,  
 ερυθροφρουρός [erithrofruros] ‘red guard’,  
 κουλάκος [kulakos] ‘kulak’  
 λευκορωσικός [lefkorosikos] ‘Byelorussian’  
 ουκάζιο [ukazio] ‘ukase’  
 ρασκόλνικος [raskolnikos] ‘Raskolnik’  
 σπούτνικ [sputnik] ‘sputnik’  
 σταχανοφισμός [staxanofizmos] ‘stakhanovism’

According to Podhajecka (2006: 132), the word ουκάζιο [ukazio] ‘ukase’ is mis-etymologized as Russian, since it is French. So the etymology of that word in DSMG should be revised. The other ten borrowings should be added in UD.

Similarly, the entry list of DSMG has to be augmented with the following twelve words found only in UD:

απαράτ [apparat] ‘apparat’  
 γκλασνοστ [glaznost] ‘glasnost’  
 δούμα [duma] ‘douma’  
 καλάσνικοφ [kalasnikof] ‘kalashnikof’  
 ΚGB [kaïebe] ‘KGB’  
 ματριόσκα [matrioska] ‘matrioshka’  
 μπάμπουσκα [babushka] ‘babushka’  
 πάβλοβα [pavlova] ‘pavlova’  
 πολίτ-μπιρό [politburo] ‘Politburo’  
 σοβχόζ [sonxoz] ‘sovkhoz’  
 στάρετς [starets] ‘starets’  
 τάιγκα [taiga] ‘taiga’

Furthermore, the entry list of both dictionaries has to be supplemented by the following words which are not included either in DSMG’s or UD’s macrostructure, even though they are quite widespread in oral or written language:

απαράτσικ [aparatsik] ‘aparatchik’  
 βογιάρος [vojaros] ‘boyar’  
 γκουλάγκ [gulag] ‘Gulag’  
 ίζμπα [izba] ‘isba’  
 κομσομόλ [komsomol] ‘komsomol’  
 Κρεμλίνο [kremlino] ‘Kremlin’  
 μπαλακλάβα [balaklava] ‘balaklava’  
 μελούγκα [beluga] ‘Beluga’  
 μπλινί [blini] ‘blini’  
 μπορτς [borts] ‘borsch’  
 ναρόδνικος [narodnikos] ‘narodnik’  
 ΝΚΒΔ [nikavede] ‘NKVD’  
 ντάτσα [datsa] ‘dacha’,  
 ουσάνκα [usanka] ‘ushanka’  
 σαμιζντάτ [samizdat] ‘samizdat’

Finally new senses in already existing entries should also be added (e.g. the sense ‘in the former Soviet Union and other communist countries a member of a children’s movement that aimed to foster communist ideals’ in the entry *πιονιέρος* [pioðeros] ‘pioneer’ or the sense ‘form of address to Stalin’ in the entry *πατερούλης* [paterulis] ‘dear father’.

## 5 Phonetic and Morphological Adaptation of Loanwords in L1

The source words of loanwords often include morphophonological properties in the donor language that do not fit into the morphophonological system of the recipient language. This is the reason why often loanwords undergo changes in order to adapt to the recipient language. The procedure of changing in order to better fit to the morphophonological system of the recipient language is called *adaptation* or *intergration* (Haspelmath 2009).

For instance, Russian [S] becomes [s] in Greek in order to adapt to the Greek phonological system which does not have a [S] sound: *μενσεβίκος* [mensevikos] ‘menshevik’. However, the degree of adaptation varies according to the age of the loanword, the knowledge of the donor language by recipient language speakers, and their attitude toward the donor language. Thus, loanwords found in the DSMG and the UD, contain rare consonant clusters and word endings that oppose to the phonological constraints and the phonotactic patterns of Greek: *μπολσεβίκος* [bolsevikos] ‘bolshevik’, *καλάσνικοφ* [kalasnikof] ‘kalashnikof’, *νομενκλατούρα* [nomenklatura] ‘nomenclature’, *γιάφκα* [jafka] ‘javka’, *ρασκόλνικος* [raskolnikos] ‘Raskolnik’, *σπούτνικ* [sputnik] ‘sputnik’, *βότκα* [votka] ‘vodka’, *πογκρόμ* [pogrom] ‘pogrom’, *σοβιέτ* [soviet] ‘soviet’. These are cases of primary adaptation (Anastassiadis 1994) and behave like foreignisms. Instable signifiers (e.g. [kefir] vs. ([kefiri], [tundra] vs. [tundra]), are also found in our data. They may denote that the loanword is in a process of adaptation. These formal variants can coexist for a considerable stretch of time, although the prevailing direction of phonological adaptation is from polyformity to uniformity (Baldunčiks 1991).

Loanword adaptation makes loanwords easily usable in the recipient language. For instance, languages with inflection and gender classes, such as Greek, need to assign verbs a person and tense inflection, and nouns a gender and inflection class in order to be used in syntactic constructions which require gender agreement. Thus, all verbs coming from Russian were integrated in the Greek inflectional system with the use of the suffix *-άρω* [aro] as in *αγκιτάρω* [agitaro] ‘agitate’ (for the morphological adaptation of loan verbs see Wohlgemuth 2009). All animate nouns were classified in the category of masculine nouns by the use of the ending *-ος* [os] or *-ας* [as], as in *ρασκόλνικος* [raskolnikos] ‘Raskolnik’, *μπολσεβίκος* [bolsevikos] ‘bolshevik’, *αταμάνος* [atamanos] ‘ataman’, *κουλάκος* [kulakos] ‘kulak’, *ινστρούχτορας* [instruxtoras] ‘instructor’ (for gender assignment in loanwords see Anastassiadis 1994; Stolz 2009). Inanimate nouns ending in the vowel *-a* were classified in the category of feminine nouns: e.g. *νομενκλατούρα* [nomenklatura] ‘nomenclature’, *γιάφκα* [jafka] ‘javka’. A number of inanimate nouns ending in a consonant were classified in the category of neutral with the attachment of the ending *-ο* or *-ι* as in *κνούτο* [knuto] ‘knut’ or *σαμοβάρι* [samovar] ‘samovar’. Only a number of inanimate nouns ending in a consonant remained morphologically non-adapted to Greek, and these were classified to the class of neutrals. These are uninflected words, e.g. *σπούτνικ* [sputnik] ‘sputnik’, *καλάσνικοφ* [kalasnikof] ‘kalashnikof’. Out of the 58 loanwords in DSMG, 16 are uninflected. Six out of 36 are uninflected loanwords in UD. These words are in the phase of primary morphological adaptation (Anastassiadis 1994).

## 6 Semantic Fields of Borrowings

In order to provide a systematic and comparative approach to the study of loanwords from Russian to Greek and the semantic categories they belong to, the Haspelmath and Tadmor (2009) semantic fields catalogue from their study *Loanwords in the languages around the world* will be used. This catalogue contains the following 24 semantic fields (see Haspelmath & Tadmor 2009): ‘The physical world’, ‘Kinship’, ‘Animals’, ‘The body’, ‘Food and drink’, ‘Clothing and grooming’, ‘The house’, ‘Agriculture and vegetation’, ‘Basic actions and technology’, ‘Motion’, ‘Possession’, ‘Spatial relations’, ‘Quantity’, ‘Time’, ‘Sense perception’, ‘Emotions and values’, ‘Cognition’, ‘Speech and language’, ‘Social and political relations’, ‘Warfare and hunting’, ‘Law’, ‘Religion and belief’, ‘Modern world’, and ‘Miscellaneous function words’. According to Tadmor (2009), the semantic field in which a word belongs affects the probability of that word being borrowed. In other words, certain semantic fields are better candidates for borrowing than others. For instance, semantic fields like ‘Religion and belief’, ‘Social and political relations’, ‘Clothing’ or ‘The house’ correspond to domains which have been affected by intercultural influences (Tadmor 2009: 64). These fields are more prone to borrowing. On the other hand, semantic fields like ‘Sense perception’ or ‘Spatial relations’ are least amenable to borrowing, since practically every language is expected to have indigenous words for such concepts.

Our data are distributed in the following semantic fields, in descending order:

- social and political relations (e.g. φράξια [fraksɣa] ‘fracsija’, αγκιτάτσια [aĩitatsɣa] ‘agitation’, κομισάριος [komisarios] ‘political officer of the communist party’, πογκρόμ [pogrom] ‘pogrom’, περεστρόικα [perestroika] ‘perestroika’, αταμάνος [atamanos] ‘ataman’);
- religion and belief (e.g. σαμάνος [samanos] ‘saman’, ουνία [unia] ‘unja’, ουνίτισσα, [unitisa] ‘female supplet of unja’, ρασκόλνικος [raskolnikos] ‘Raskolnik’);
- food and drink (e.g. βότκα [votka] ‘vodka’, πιροσκι [piroski] ‘piroshki’, κεφίρ [kefir] ‘kefir’, πάβλοβα [pavlova] ‘pavlova’);
- the house (e.g. σαμοβάρι [samovar] ‘samovar’, μπάμπουσκα [babushka] ‘babushka’);
- clothing and grooming (e.g. αστρακάν [astrakan] ‘astrakan’, ουσάνκα [usanka] ‘ushanka’);
- basic actions and technology (e.g. κοσμοναύτης [kozmonaftis] ‘cosmonaute’, κοσμοδρόμιο, [kozmodromio] ‘cosmodrome’ σπούτνικ [sputnik] ‘sputnik’, μαζούτ [mazut] ‘mazut’);
- physical world (e.g. στέπα [stepa] ‘steppe’, τούντρα [tundra] ‘tundra’, τάιγκα [taiga] ‘taiga’);
- animals (κουτάβι [kutavi] ‘puppy’, μαμούθ [mamuθ] ‘mammoth’);
- warfare and hunting (e.g. κνούτο [knuto] ‘wip’, μολότοφ [molotof] ‘molotof’).

In accordance with Tadmor (2009), the majority of our data are loanwords which refer to ‘Social and political relations’, ‘Religion and belief’, ‘Food and drink’, ‘The house’.

The most prolific field is that of politics, with many loanwords relating to the period of the Soviet Union. The great impact of the Russian Revolution and the subsequent Communist regime marked the end of the old autocratic rule (czarism) and largely influenced modern communities and their languages with the new borrowings which refer to new forms of social organization, new institutions, and new ranks. These words were introduced in the Greek language as loans with a denotative meaning; however, they acquired specific positive connotations in the leftist political vocabulary, and were used with the aim to declare a left-wing political identity and ideological proximity to the Soviet regime and communism. In some other cases, they had negative connotations expressing the depreciation of speakers towards the Communist regime: ιντελιγκέντσια [inteliĩentsia] ‘intelligentia’ (vs. διανόηση [dianoisi] ‘intellectuals’), αγκιτάτορας [aĩitatoras] ‘agitator’, ρεβιζιονιστής [revizionistis] ‘revisionist’, προβοκάτσια [provokatsɣa] ‘provocation’ (vs. πρόκληση [proklisi] ‘provocation’).



The data that belong to other categories are products of borrowing for external reasons (they result from the extra-linguistic realm and are related to material-economic or cultural reasons). More specifically, they are borrowings used to denominate new notions or objects which refer to Russian culture and belong to the semantic fields of ‘Food and drink’, ‘The house’ or ‘Clothing’ (e.g. βότκα [votka] ‘vodka’, πιροσكى [piroski] ‘piroshki’, σαμοβάρι [samovar] ‘samovar’, μπαλαλάικα [balalaika] ‘balalaika’). These three semantic fields are similar, since globalization and continued migration have contributed to the spread and adoption of such words worldwide.

In line with Tadmor (2009), no Russian loanwords belonging to the semantic fields of the physical world, kinship, the body, motion, possession, spatial relations, quantity, time, sense perception, emotions and values, cognition, speech and language were found in Greek.

### 6.1 Stereotypes and Borrowing

Anastassiades (1994) argues that by studying the semantic fields of L1 where borrowings from L2 abound it is possible to account for the reasons for borrowing, since borrowings reflect the image that L1 speakers have of the country where the donor language is spoken or their stance towards it. This image does not represent reality (in the sense that it is not a photographic imprint of it), but it has a symbolic value.

In order to collect quantitative data for studying the possible correlation between a) the stereotypic representations that L1 speakers have in mind about L2 speakers, and b) the semantic fields of Russian loanwords in Greek, a brainstorming experiment was held on Facebook. The stimulus word was ‘Russia’. More particularly, the participants were asked to post which idea, word or image was recalled first when they heard the word ‘Russia’.

Five hundred and six subjects, aged 19 to 70 years old participated in the research by posting their answer in a period of 48 hours. Table 1 shows their answers.

Table 1: Frequency of the recalled words, ideas or images

Words/ideas/images	Frequency
Putin	48
Communism	38
Vodka	37
Red Square	34
cold	33
red	30
baboushka	27
Orthodoxy, St. Petersburg	23
Moscow	21
Tsar	20
Stalin, Dostoyevsky	10
ballet	8
revolution	7
Red Army, perestroika, CCCP	6
Soviet Union, Kremlin, steppe, Gorbachev, bear	5
literature, Lenin, Romanov	4
Tchaikovsky/Tolstoy/ushanka/Bolshoi/Siberia	3
hammer and sickle, piroshki, Chekhov	2
Cold war, Pushkin, oligarchs, mafia, Gulag, Rasputin, Volga, proletarian, samovar, Dr. Zhivago, caviar	1

As one can see from Table 1, the majority of answers refer to different political regimes of Russia (tsarist period, Soviet era or Perestroika). Actually, the correlation of the representations and stereotypes emerging from the brainstorming, with the semantic fields where there is high borrowing from Russian, indicates that most borrowing takes place in the field of politics, whether it concerns the tsarist (e.g. κουλάκος ‘kulak’), the Soviet (e.g. κομισάριος ‘commissary’), or the post-Soviet period (e.g. Perestroika, oligarchs).

A lot of stereotypes refer to toponyms and geography (vast unpopulated areas), Russian literature, ballet, food and climate conditions (cold), and the Russian mafia. These stereotypical perceptions are associated with everyday vocabulary words that refer to typical Russian clothing (μπαλακλάβα ‘balaklava’, ουσάνκα ‘ushanka’) drink (βότκα ‘vodka’, σαμοβάρι ‘samovar’), music (μπαλαλάικα ‘balalaika’), geographical terms (τάιγκα ‘taiga, τούνδρα ‘tundra’, στέπα ‘steppe’).

Russian borrowings enriched the political vocabulary of the left Part in Greek. Russian loanwords are used by L1 speakers who want to express their sympathy to the Left or ironically by others who want to criticize it. This vocabulary stereotypically reflects the image of a country that was marked by the Communist regime or Perestroika and its positive (glasnost) or negative results (oligarchs, mafia, etc.). The results revealed a pattern of responses with older people (aged 45-70) associating the word ‘Russia’ mainly with its Soviet history, and the younger ones showing more negative stereotypes and associating Russia with corruption, lack of democracy and the rise of oligarchs.

## 7 Conclusions

This paper investigated Russian loanwords, loanblends and loanshifts that entered Greek lexicon during various historical periods and how they have been recorded into two major dictionaries of Modern Greek. The comparison of the two dictionaries showed that the routes of loanwords did not always overlap. There were discrepancies between the two works in the number of entries of Russian origin as a result of different lexicographic practices adopted during the dictionary compilation process. The paper also offered a morphophonological analysis of Russian loanwords in Greek. The study of the semantic fields in which Russian borrowings in Greek belong revealed that the most prolific semantic fields were ‘Social and political relations’, ‘Religion and belief’, ‘Food and drink’ and ‘The house’. These results confirmed the observation of Tadmor (2009: 64) that “different languages display a remarkable degree of consistency with regard to which fields are more or less affected by borrowing”. Finally, the paper provided experimental data supporting Anastassiadis’s (1994) claim that lexical fields, in which loanwords abound, reflect the stereotypic image of the country where the donor language is spoken.

## References

- Anastassiadis, A., (1994). *Neological borrowing in Greek* [IN GREEK]. Thessaloniki.
- Andreitsenko, E., (2014). *Greek loanwords with Orthodox religious content in the evolution of Russian* [IN GREEK], Epikairoitita.
- Apostolou-Panara, A-M., (1991). English loanwords in Modern Greek: An overview. *Terminologie et Traduction*, 1, pp. 45-58.
- Baldunčiks, J., (1991). West European loanwords in Latvian. In. Ivir V & Kalogjera D. (eds) *Languages in contact and contrast*, pp. 15-24, Berlin: Mouton de Gruyter.
- Contossopoulos, N., (1978). *L'influence du français sur le grec, emprunts lexicaux et calques phraséologiques*. Athènes.
- Haspelmith, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. In T. Stolz, D. Bakker & P.-R. Salas (eds.), *Aspects of language contact: New theoretical, methodological and empirical findings with special focus on Romancisation processes*, pp. 43–62. Berlin: Mouton de Gruyter.

- Haspelmath, M. & Tadmor, U., (2009). *Loanwords in the world's languages*. The Hague: De Gruyter Mouton.
- Haugen, E., (1950). The analysis of linguistic borrowing. *Language*, 26, pp. 210-231.
- Hout, R., & Muysken, P., (1994). Modelling lexical borrowability. *Language, variation and change*, 6(1), pp. 39-62.
- Humbley, J., (1974). Vers une typologie de l'emprunt lexical. *Cahiers de Lexicologie*, 25(2), pp. 46-70.
- Johanson, L., (2002). *Structural factors in Turkic language contact*. London: Curzon.
- Matras, Y., (1998). Utterance modifiers and universals of grammatical borrowing. *Linguistics*, 36, pp. 281-331.
- Matras, Y., (2007). The borrowability of structural categories. In G. Bossong, B. Comrie & Y. Matras (eds) *Empirical Approaches to Language Typology*, pp. 31-74, Berlin: Mouton de Gruyter.
- Matras, Y., & Sakel, J., (2007). *Grammatical borrowing in cross-linguistic perspective*. Berlin: Mouton de Gruyter.
- Matras, Y., (2011). Universals of structural borrowing. In P. Siemund *Linguistic Universals and language variation*, pp. 204-233. Berlin: De Gruyter.
- Myers-Scotton, C., (2002). *Language contact: bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.
- Podhajecka, M. (2006). Russian borrowings in English: similarities and differences in lexicographic description. In R. W. McConchie et al. (eds.) *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX)*, Sommerville, MA: Cascadia Proceedings Project, pp.123-134.
- Stolz, C., (2009). A different kind of gender problem: Maltese loanword gender from a typological perspective. In B. Comrie, R. Fabri, E. Hume, M. Mifsud, T. Stolz. M. Vanhod (eds) *Introducing Maltese linguistics*, pp. 321-353, Amsterdam: John Benjamins.
- Stubbs, M., (1998). German loanwords and cultural stereotypes. *English Today*, 14(1), pp. 19-26.
- Tadmor, U., (2009). Loanwords in the world's languages: Findings and results, In Haspelmath, M. & Tadmor, U., (eds). *Loanwords in the world's languages*, pp. 55-75, The Hague: De Gruyter Mouton.
- Thomason, S., (2001). *Language Contact*. Georgetown University Press, Washington.
- Thomason, S., & Kaufman, T., (1992). *Language Contact, Creolization and Genetic Linguistics*. Berkeley: University of California Press.
- Wohlgemuth, J., (2009). *A typology of verbal borrowings*. Berlin: Mouton de Gruyter.



# Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary

**Laura Giacomini**

Heidelberg University

E-mail: [laura.giacomini@iued.uni-heidelberg.de](mailto:laura.giacomini@iued.uni-heidelberg.de)

## Abstract

In this paper, a frame-based approach to terminological variation is presented along a model for presentation of multiword terms and their variants in a technical e-dictionary. A case study concerning terminology related to semiconductor diodes is the background against which methods and goals of a larger study on the technical language (Habilitation thesis at Hildesheim University) are illustrated and compared with those of existing resources. At the core of the proposed model are three interrelated description layers (ontology – frame – terminology), with the frame layer serving as the semantic interface between ontological classes and terms, as well as a variation typology accounting for orthographical, morphological and syntactic term variants. The microstructural properties of the envisaged e-dictionary, which aims at supporting text production in the native language, are illustrated by means of the example entry *diode in forward bias*. The addressed users, technical writers and professional translators, are able to access all types of data separately from each other, in a modular way. The paper closes with an outlook on how future developments could include application of the model to further technical domains.

**Keywords:** frame-based terminology, frame-based lexicography, term variation, technical language, LSP dictionary

## 1 Introduction

This paper describes a model for data presentation in a technical e-dictionary with special focus on variation of multiword terms. This is part of a larger corpus-based study on the modelling of a terminological database for lexicographic purposes<sup>1</sup>.

A multiword term is functionally understood as “a term containing two or more content words” (Jacquemin & Tzoukermann 1999: 26). The technical e-dictionary, which aims to fill a clear gap in lexicographic coverage of variation, is intended to support text production and specifically addresses professional translators and technical writers. The proposed data representation method and the related lexicographic presentation are explained by using English denominations employed in the field of electrical engineering and referring to semiconductor diodes (Figure 1). Generally speaking, semiconductors are “solids whose electrical conductance lies between that of good conductors and insulators” (Clouden 2014: 80).

A diode is a specialized electronic component with two electrodes called the anode and the cathode. Most diodes are made with semiconductor materials such as silicon, germanium, or selenium. [...] Diodes can be used as rectifiers, signal limiters, voltage regulators, switches, signal modulators, signal mixers, signal demodulators, and oscillators. The fundamental property of a diode is its tendency to conduct electric current in only one direction. (<http://whatis.techtarget.com>)

<sup>1</sup> Habilitation thesis at the Institute of Information Science and Natural Language Processing, Hildesheim University (Germany).



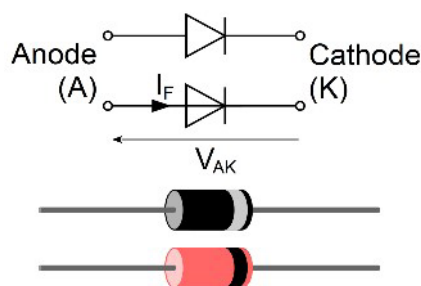


Figure 1: Semiconductor diode and diode symbol with indication of the anode and cathode.<sup>2</sup>

The terminology of electrical engineering is subject to a comparatively high degree of standardization, with national and international institutions and associations contributing to unification of terminology in its different subject areas (e.g. electromagnetism, circuit theory, and computer network technology, to name just a few that are part of the Electropedia classification; cf. <http://www.electropedia.org>). In comparison with other technological areas, this results in a lexically homogeneous domain. Still, synonymous variation appears to be a quantitatively and qualitatively relevant phenomenon which deserves careful attention. So far, many studies in computational linguistics (cf. Daille 2005, Daille 2017), terminology and translation (cf. Fernández-Silva & Kerremans 2011, Temmerman 2000) have systematically explored terminological variation, conveying a view of terminology that is quite different from its more traditional, monolithic interpretations.

After pointing out which notion of variation and which types of variants will be taken into consideration (Section 2), we will concentrate on a frame-based approach to data modelling (Section 3), and then finally on data presentation in a technical e-dictionary (Section 4). The last section of this paper recapitulates key findings and draws some conclusions on the significance of frame-based specialized lexicography and the applicability of the proposed description model to other technical domains.

## 2 Term Variation

### 2.1 Which Notion of Variation?

In the context of this study, terminological data is retrieved from a database that collects terms, variants and relations extracted from corpora of specialized texts and addressing experts and semi-experts. Multiword terms are at the center of discussion as terms which are most exposed to non-diasystemic variation, i.e. to (near) synonymous variation occurring at the same discourse level with no diachronic, diatopic, register-, or corporate language-related changes (Caro Cedillo 2004), e.g.

*depletion region*  
*depletion zone*  
*depletion layer*  
*space charge region*  
*space charge layer*.<sup>3</sup>

The study focuses on variants which match the following criteria:

<sup>2</sup> Erik Streb CC-BY-SA-3.0 [http://en.wikipedia.org/wiki/File:Diode\\_pinout\\_en\\_fr.svg](http://en.wikipedia.org/wiki/File:Diode_pinout_en_fr.svg)

<sup>3</sup> “Near the junction, a depletion region is created by electrons from the N-type material moving in to fill holes in the P-type material, and holes moving in the opposite direction (from the P-type material) to combine with available electrons. The depletion region is electrically neutral, but separates the N- and P-type materials, which have a difference in potential called the barrier potential (or junction voltage)” (Diffenderfer 2005: 41).

- Variants are totally or partially synonymous,
- Variants display a morphological similarity (similarity is hereby defined as the presence of shared lexical morphemes), and
- Variants build term clusters which mostly belong to the same systemic level. Diasystemic, e.g. geographical, variation is still accounted for if available, but does not represent the focus of this study.

Empirical observation of term behavior in technical language often provides evidence of variant clustering in the same text, with the use of variants motivated by discourse-related, functional, inter-linguistic, and cognitive factors (for a detailed coverage of variation grounds, cf. Freixa 2005). Here are a few examples of synonymous variation within the same text:

“a diode... is forward biased [...] the biasing is classified as Forward biasing and Reverse biasing of a diode [...] a diode is connected in a forward bias” (Godse & Bakshi 2010: 73)

“The shape of the charge density,  $p$ , depends upon how the diode is doped. Thus, the junction region is depleted of mobile charge carriers. Hence, it is called the depletion region (layer), the space charge region, or the transition region. The depletion region is of order 0.5  $\mu\text{m}$  thick. There are no mobile carriers in this very narrow depletion layer. Hence no current flows across the junction and the system is in equilibrium” (Salivahanan et al. 1998: 88)

“SEMICONDUCTOR DIODE AS A RECTIFIER Figure 7.15 depicts the rectifying action of a semiconducting diode. [...] In this way, the semiconductor diode has been able to do rectification, i.e., change ac into dc” (Joshi 2010: 7.12).

Moreover, the availability of a relatively large number of  $n$ -gram variants in various technical fields suggests that this is likely to be a common phenomenon in technical language. The presence of different degrees of extension due to field-specific properties does not refute these findings, it simply proves that variation is a natural process, at least in some LSP and that it is strictly interconnected with language- and text-dependent factors, for instance the type of communication involved (i.e. domain-internal or domain-external), and the specific register or source features.

## 2.2 Which Types of Variants?

The overall study deals with variant description at the single term and multiword term levels. Multiword terms, on which this paper concentrates, include both complex terms and phrasemes. We distinguish the following three types of variation:

- 1) MV, morphological variation (partial / total):  
changes in lexical morphemes (e.g. *depletion layer* vs. *depletion region*)
- 2) SV, syntactic variation:  
changes in the part of speech, word order, and sentence construction (e.g. *depletion region* vs. *region of depletion*)
- 3) OV, (ortho)graphical variation:  
changes in hyphenation and capitalization (e.g. *light-emitting diode* vs. *light emitting diode*).

In addition to the examples just mentioned, the three types often combine with each other, building complex patterns of variation.

## 3 Frame-based Data Modelling

In the present study, variant modelling is largely based on a frame-based approach to terminology, in which basic ideas deriving from Frame Semantics (Fillmore 1977, Ruppenhofer et al. 2006) and

Frame-based Terminology (Faber 2015) are adapted to the modelling of specialized discourse, with the purpose of representing terms of a certain domain according to the role they play in certain domain-specific scenarios. Some frame-oriented lexicographic and terminographic resources have already been published over the last ten years. Well-known examples are the multilingual EcoLexicon (Reimerink & Faber 2009), developed at the University of Granada and covering environmental terminology and the Kicktionary (Schmidt 2014), which deals with the language of football. The focus of our model with respect to the existing resources primarily lies in

- the inclusion of an extensive domain ontology which interfaces with the lexicon through the frame layer; the three layers of analysis (ontology – frame – terminology) are linked to and motivated by each other;
- the focus on terminological variation, with frame elements serving as identifiers of shared or distinct semantic roles in orthographical, morphological, and syntactic variants;
- the monolingual orientation of term and variant representation for supporting text production in the native language.

Data representation in the terminology database relies on a multi-layered model in which terms and variants undergo a top-down analysis process beginning with their conceptual background (ontological layer), going through their semantic content (frame layer) and ending with their morphological and syntactic features (see Table 1).

Table 1 – Multi-layered model and related procedural steps

LAYER	PROCEDURE	COMPONENTS
Domain ontology layer	Designing an ontology for semiconductor devices.	The ontology includes taxonomic and non-taxonomic relations between classes.
	Selecting a key ontological entity: SEMICONDUCTOR DIODE.	
Frame layer	Identifying possible frames related to the semiconductor diode, e.g. PRODUCTION or SALE.	
	Selecting a frame to be described in the model: FUNCTIONALITY.	The frame includes core and non-core frame elements.
Lexical layer	Modelling single terms, multiword terms and term variants.	Morphosyntactic, frame-related, ontological and variational features.

The top level of the proposed model is a domain ontology structured around a key entity, the SEMICONDUCTOR DIODE, which constitutes the topical focus of the available corpus texts. At the interface between the top ontological level and the bottom lexical level is a frame level in which the key entity is semantically accounted for in the sense of Frame Semantics (Fillmore 1977, Ruppenhofer et al. 2006) and Frame-Based Terminology (Faber 2015). The frame FUNCTIONALITY is selected among the possible frames describing a semiconductor diode, and each term or term component directly denoting or indirectly referring to a diode can be reduced to the identified frame elements (e.g. SEMICONDUCTOR MATERIAL, CONSTRUCTION FORM, APPLICATION TECHNIQUE). Figure 2 shows the combination of the three frame elements PRODUCT (PROD), GOAL (GOAL) and PROPERTY (PROP) in a set of synonymous multiword terms.

DIODE TYPE (Prod) + GOAL (Goal) + PROPERTY(Prop)

rectifier diode i-v characteristics

N<sub>Goal</sub> N<sub>Prod</sub> N<sub>Prop</sub>

i-v curve for a rectifier diode

N<sub>Prop</sub> p<sub>for</sub>N<sub>Goal</sub> N<sub>Prod</sub>

current-voltage characteristics of a rectifier diode

N<sub>Prop</sub> p<sub>of</sub>N<sub>Goal</sub> N<sub>Prod</sub>

i-v curve of a diode for rectification

N<sub>Prop</sub> p<sub>of</sub>N<sub>Prod</sub> p<sub>of</sub>N<sub>Goal</sub>

Figure 2 : Set of synonymous variants and corresponding syntactic-semantic annotation.

The frame-based approach to terminology, with its clear connection to cognitive linguistics (cf. Faber 2012), is at the core of the illustrated model. On the one hand, this approach establishes a link between the lexical and the ontological level, with frames seen as seme subsets and semes as semantic roles attached to ontological entities. On the other hand, this approach provides the necessary key for interpreting and describing the correspondence between morphosyntactic and semantic features of any multiword term.

The bottom level of the model envisages lexical analysis along morphosyntactic, conceptual (i.e. ontological and frame-related) and variational parameters, each supporting a different task in text production.

Any multiword term,

e.g. *rectifier diode i-v characteristics*,

in which the abbreviated form *i-v* stands for *current-voltage*, can be formally described in terms of

- I its morphosyntactic structure: AP (A *rectifier* + N *diode*) + NP (N *i-v* + N *characteristics*),
- II its rule-based relation to the basic morphosyntactic structures of its language: AP + NP,
- III the frame elements denoted by its constituents: GOAL + PRODUCT + PROPERTY, and
- IV the ontological classes to which these constituents are linked: FUNCTION: APPLICATION + SEMICONDUCTOR DIODE + MATERIAL: PROPERTY (see Table 2 for an overview of description I. to IV.).

Table 2: Term description based on the multi-layered model.

Domain ontology layer	FUNCTION: APPLICATION	SEMICONDUCTOR DIODE	MATERIAL: PROPERTY
Frame layer	GOAL	PRODUCT	PROPERTY
Lexical layer	A <i>rectifier</i>	N <i>diode</i>	N + N <i>i-v characteristics</i>

Furthermore, each variant of the given multiword term,

e.g. *i-v curve of a diode for rectification*,

is assigned to

- VI. a specific variant class and type: partial morphological variation (*characteristics > curve*) + syntactic variation (AP NP > NP PP), and
- VII. a specific variation template: paraphrase (*diode characteristics > curve of a diode, rectifier diode > diode for rectification*) + explicitation (*characteristics > curve*) + transposition (*rectifier > rectification*).

Variation is always identified in relation to a main term. The selection of a main term takes place by referring to existent standards and/ or to quantitative analysis in the available texts. However, the designation of a multiword term as a main term to which one or more variants are attached is a topic dealt with in the main study.

Frame-based data modelling has been developed, which pays special attention to the granularity of information. Descriptors, for instance frame elements, need to be specific enough to deliver a precise, unambiguous semantic characterization of terms, and general enough to be applicable to other technical domains. In particular, the feasibility of the model has been tested in other domains centered on technical artefacts (thermal insulation products and DIY-tools).

## 4 Data Presentation in a Technical E-dictionary

The parameters described in this section are the main building blocks of the abstract microstructure of a dictionary entry and will be discussed with the help of representative examples. The purpose of these examples is to comprehensively illustrate the model for lexicographic data presentation together with corresponding search options (semasiological and onomasiological access structures) and visualisation options.

Terminological variation is a phenomenon text producers have to deal with. The issue about the availability or adequacy of a given variant for a given context is well known among translators and technical writers. However, doubts cannot be removed by just using lexicographic resources, as lexicographic information tools (LSP dictionaries, glossaries and terminological databases) usually have the following characteristics:

- They cover only a small fraction of the commonly used variants;
- They rarely record longer multiword terms than bigrams;
- They usually contain possible variants at different levels of discourse (for instance geographical or chronological variants, e.g. *PNPN diode* vs. *Shockley diode*) or, in general, variants with no morphological affinity (e.g. *bias/ direction*);
- Their presentation produces coherency issues at macrostructural, microstructural and mediostructural level (variants may be lemmatized or not, may have their own search area within an entry or be indicated in different microstructural positions, or they may lack cross-referencing).

This results in the fact that time-consuming queries in parallel and comparable corpora are often required to obtain information concerning potential variants. Lexicographic resources are needed which provide users with terms together with relevant variants and variant-related information. A range of requirements can now be identified for lexicographic coverage of term variation:

- A need for systematic coverage of non-diasystemic variation;
- A need for syntactic and semantic information concerning variants;
- A need for pragmatic information concerning text sources, genres, type of communication;
- A need for a clear and coherent link between domain terminology and domain ontology.



As pointed out in Giacomini (2017), the operational and cognitive difference between the tasks of technical writers and professional translators do not change the fact that the main function of the envisaged dictionary should be to make variants and information about variants available in the native language of its users. Moreover, a lexicographic entry should consist of separate modules dedicated to the treatment of different information types. Each module should be separately accessible and information types should be combinable in order to enable users to perform targeted queries.

From a general structural perspective, a non-form-determined (conceptual) macrostructure and a form-determined macrostructure should be best combined (for classification of different types of macrostructures in electronic lexicography cf. Giacomini 2015). As a consequence, external data access should be made possible via both the ontological and frame-based path and the terminological path. In the proposed model, multiword terms and their variants appear both as a lemma and as another possible microstructural item (e.g. a variant or part of a corpus example) with related cross-references. Cross-referencing ensures a coherent representation of the different roles a term may play within the dictionary structure and, at the same time, reflects the relations existing between the different layers of the data model.

Given a multiword lemma as a main term, the abstract microstructure for a multiword term entry can be defined as follows:

#### ABSTRACT MICROSTRUCTURE

ontology-related data

frame-related data

language

lemma (main term, MT)

– syntactic and semantic structure

– corpus example(s)

– source(s)

– image

– variant

— syntactic and semantic structure

— variation type (O-, O+, M-, M~, M+, S-, S+)<sup>4</sup>

— corpus example(s)

— source(s)

Further items may apply to specific microstructural data, but will not be the object of this paper.

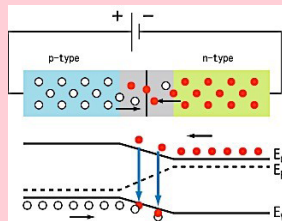
The concrete microstructure related to the lexicographic entry of the multiword term *diode in forward bias* can be visualized as the composition of three descriptive areas corresponding to the three layers of data analysis, i.e. the terminology layer, the frame layer and the ontology layer. All information concerning relevant frame elements and ontological classes refers to both a term and its variants. An image<sup>5</sup> is also integrated into the lexicographic entry and is attributed to the main term.

4 The - and + symbols indicate absence or presence of a certain trait, while the ~ symbol, which is only contemplated in the case of morphology, stands for partial morphological variation.

5 S-kei CC-BY-SA-2.5 <https://commons.wikimedia.org/wiki/File%3APnJunction-Diode-ForwardBias.PNG>

TERMINOLOGY: EN**MT: diode in forward bias**

$N_{Prod} p_{in} N_{Prop}$   
 The small signal model of a **diode in forward bias** is a resistance in parallel with a capacitance  
[\[inst.eecs.berkeley.edu\]](http://inst.eecs.berkeley.edu)



forward biased diode

$V_{Prop} N_{Prod}$   
 O- M- S+

In a **forward biased** p-n junction **diode**, the positive terminal of the battery is connected to the p-type semiconductor material and the negative terminal of the ...  
[\[physics-and-radio-electronics.com\]](http://physics-and-radio-electronics.com)

**diode under forward bias**

$N_{Prod} p_{under} N_{Prop}$   
 O- M- S-

Figure 4.4.5: Current-Voltage characteristics of a silicon **diode under forward bias**  
[\[ecee.colorado.edu\]](http://ecee.colorado.edu)

**diode in forward direction**

$N_{Prod} p_{in} N_{Prop}$   
 O- M~ S-

First produced by Clarence Zener in 1934. It is similar to normal **diode in forward direction**, it also allows current in reverse direction when the applied voltage reaches the breakdown voltage.  
[\[Electronicshub.org\]](http://Electronicshub.org)

FRAME:

Functionality

- Product:  
**diode**
- Property:  
**forward bias**

ONTOLOGY**SEMICONDUCTOR****DIODE**

- MATERIAL
- SEMICONDUCTOR MATERIAL
- HOUSING MATERIAL
- **PHYSICAL PROPERTY**
- FORM
- COMPONENT
- CONSTRUCTION FORM
- HOUSING TYPE
- FUNCTION
- MOUNTING TECHNOLOGY
- MOUNTING TECHNIQUE
- APPLICATION
- USER

Separate or combined queries involve each data type (i.e. microstructural item) available in the dictionary database. For instance, the output of a search query can be

- (i) all variants of a multiword term,
- (ii) specific orthographic, morphological or syntactic variants of a multiword term (e.g. O- M- S-),
- (iii) multiword terms corresponding to a given syntactic structure with given POS content (e.g. N pN),
- (iv) multiword terms matching a specific frame element or frame element combination (e.g. PRODUCT + PROPERTY),
- (v) multiword terms matching a specific ontological class or class combination (e.g. terms matching the class PHYSICAL PROPERTY), or
- (vi) frame elements and ontological classes matching a multiword term.

## 5 Conclusion and Further Work

This paper has introduced a frame-based description model for technical terms and their variants in an e-dictionary covering terminology related to the field of semiconductor diodes. The main goal of the paper was to provide information regarding methodology involved in developing the three layers of term and variant analysis (ontology – frame – terminology) and to introduce microstructural properties of the technical dictionary. Synonymous variation, especially in multiword terms, is at the center of discussion as a significant but still underestimated phenomenon in terminology. Electronic lexicography is a privileged area in which resources can be created to provide extensive coverage of this phenomenon. In this same area, an important contribution to the improvement of NLP procedures for term and variant extraction from specialized corpora can be made by exploring target users' needs and designing corresponding data models.

The applicability of the proposed approach to other technical domains is presently being tested on corpora containing texts about technical artefacts (thermal insulation products and DIY-tools) but referring to technical domains with different conceptualization, standardization and communicative features. Frame-based annotation of terms and variants turns out to be feasible provided that requirements for a coherent ontology and an exhaustive frame description with the right granularity (i.e. an appropriate level of semantic detail to ensure reliable annotation) are satisfied. Promising results obtained in the context of this paper as well as in the underlying project for what concerns frame-based data modelling and related corpus annotation lay a sound basis for future efforts towards a better lexicographic and terminographic coverage of multiword term variation in specialized language.

## References

- Caro Cedillo, A. (2004). *Fachsprachliche Kollokationen*. Tübingen: Gunter Narr.
- Clouden, L. (2014). *Physics: A Concise Revision Course for CXC*. Cheltenham: Stanley Thornes.
- Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology* 11.1.
- Daille, B. (2017). *Term Variation in Specialised Corpora*. Amsterdam: John Benjamins.
- Diffenderfer, R. (2005). *Electronic Devices: Systems and Applications*. Clifton Park: Thomson.
- Faber, P. (2015). Frames as a framework for terminology. In: Kockaert, H.J./ Steurs, F. (eds.), *Handbook of terminology* (Vol. 1). Amsterdam: John Benjamins, pp. 14-33.
- Faber, P. (ed.) (2012). *A cognitive linguistics view of terminology and specialized language*. Berlin/Boston: De Gruyter Mouton.
- Fernández-Silva, S. & Kerremans, K. (2011). Terminological variation in source texts and translations: A pilot study. In: *Meta: Journal des traducteurs. Meta: Translators' Journal*, 56(2).
- Fillmore, C. J. (1977). Scenes-and-frames semantics. In: A. Zampolli (ed.), *Linguistic Structures Processing*. Amsterdam: North-Holland Publishing Company, pp. 55-81.
- Freixa, J. (2005). Variación terminológica: ¿por qué y para qué?. In: *Meta: Journal des traducteurs. Meta: Translators' Journal*, 50(4).
- Giacomini, L. (2017). An ontology-terminology model for designing technical e-dictionaries: formalisation and presentation of variational data. In *Proceedings of eLex 2017*, September 2017, Leiden (NL).
- Giacomini, L. (2015). Macrostructural properties and access structures in LSP e-dictionaries for translation: the technical domain. In *Lexicographica* 31.2015, pp. 90-117.
- Godse, A.P. & Bakshi, U.A. (2010). *Electronic Devices and Circuits I*, Pune: Technical Publications.
- Jacquemin, C. & Tzoukermann, E. (1999). NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In: *Natural language information retrieval*. Dordrecht: Kluwer Academy Publishers, pp. 25-74.
- Reimerink, A. & Faber, P. (2009). Ecolexicon: A frame-based knowledge base for the environment. In *Proceedings of the International Conference Towards eEnvironment*, pp. 25-27.
- Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C.R. & Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

- Salivahanan, S., Suresh Kumar, N. & Vallavaraj, A. (1998). *Electronic Devices and Circuits*. New Delhi: Tata McGraw-Hill.
- Schmidt, T. (2014). *The Kicktionary: Combining corpus linguistics and lexical semantics for a multilingual football dictionary*. IDS Mannheim.
- Temmerman, R. (2000). *Towards new ways of terminology description: The sociocognitive-approach* (Vol. 3). Amsterdam: John Benjamins.

# Dictionaries of Linguistics and Communication Science / *Wörterbücher zur Sprach- und Kommunikationswissenschaft* (WSK)

**Stefan J. Schierholz**

*Friedrich-Alexander-Universität Erlangen-Nürnberg*

*E-mail: Stefan.Schierholz@fau.de*

## Abstract

The “WSK” is a German online dictionary series which will be published in print in 2019/2020. Each of the dictionaries is a terminological special field dictionary on the subject of “Linguistics and Communication Science”. The dictionaries will be partially translated to English, are intended for experts and semi-experts, and they will serve for comprehension of technical terms, information, and translation. Currently, 25 dictionaries are envisioned for the series, which will contain more than 50,000 lemmas. Eleven thousand of these dictionary articles have been published online since 2013. About 800 people are working on the project worldwide.

In this paper, the structure of the whole project, the organization and management, and the work flow of article writing will be presented. By taking volume 1, “Grammar”, as a basis, the text compound structure, the function of the systematic introduction, and the article structure will be introduced.

**Keywords:** special field lexicography, terminology, WSK, grammar, dictionary project

## 1 Preliminaries

In the book industry “WSK” has been a brand for a couple of years now. The WSK were established by the publishing house de Gruyter, which is responsible for the publication of the dictionaries. The abbreviation “WSK” is for the German title *Wörterbücher zur Sprach- und Kommunikationswissenschaft*, which you will find in the title of this contribution. Please consult the following website for more information: [www.wsk.uni-erlangen.de](http://www.wsk.uni-erlangen.de). WSK was founded by Herbert Ernst Wiegand and Stefan J. Schierholz in 2003, and was planned as a print dictionary series first. For more details on the project see Schierholz (2007, 2008, 2010, 2015), Schierholz/Wiegand (2004), and Wiegand (2002, 2003, 2004, 2006a, 2006b, 2009). Since 2013, the dictionary articles have been published in a WSK online version, and Stefan J. Schierholz has been the sole editor of the series since 2018.

## 2 General Description of the WSK

The WSK dictionaries will be published as print dictionaries and every one of them will be an autonomous, alphabetically organized terminological subfield special field dictionary, and will be partially bilingualized to English. Intended for experts and semi-experts, the books will serve as specialist dictionaries for reference and study. The terminological breakdown of the overall conceptual domain “Linguistics and Communication Science” and the distribution of the lemmas to the dictionaries will not and cannot be dictated by any abstract pre-assignment of disciplinary boundaries, whether theoretically or historically based. Rather, although being aware of such boundaries, the distribution will be fundamentally practical and pragmatic: who will actually use which of the subfield special field



dictionaries under what circumstances, and with what kind of information goals? In light of various possible answers to such questions, 25 volumes are envisioned and work on the following 18 subject fields is underway (volume editors in brackets):

Grammar with two sub-volumes: Morphology, Syntax (Stefan J. Schierholz/Pál Uzonyi)  
 Word Formation (Peter O. Müller/Susan Olsen)  
 Historical Linguistics (Mechthild Habermann/Ilse Wischer)  
 Phonetics and Phonology (Bernd Pompino-Marschall)  
 Writing (Martin Neef/Rüdiger Weingarten/Said Sahel)  
 Text Linguistics and Stylistics (Christina Gansel/Constanze Spieß)  
 Dialectology (Heiko Girth/Peter Rosenberg)  
 Quantitative and Formal Linguistics (Peter Grzybek/Reinhard Köhler/Sven Naumann)  
 Linguistic Theory and Methodology (Bernd Kortmann)  
 Lexicology and Phraseology (Christiane Fellbaum/Ingo Warnke/Daniel Schmidt-Brücken)  
 Language Typology (Johannes Helmbrecht/Dagmar Jung)  
 Cognitive Grammar (Constanze Juchem-Grundmann)  
 Philosophy of Language (Christoph Demmerling/Pirmin Stekeler-Weithofer)  
 Language Teaching: Native and Foreign Language (Jörg Kilian/Jutta Rymarczyk)  
 Onomastics (Kirstin Casemir/Eckhard Meineke)  
 Media Sciences (Michael Bachmann/Christoph Bläsi)  
 Terminology (Rute Costa/Fidelma Ní Ghallchobhair/Klaus-Dirk Schmitz).

So far, volume editors have not been found for the following subject fields: Language Technology and Computational Linguistics, Semantics and Pragmatics, Clinical Linguistics, Sociolinguistics, Science of the Translation, Lexicography and Dictionary Research, Language and Interaction, and Language Names.

All volumes will provisionally bear the same subtitle (here showing vol. 1 as an example):

“Grammar. Study and Reference Dictionary. With a Systematic Introduction and English translations” (*“Grammatik. Ein Lern- und Konsultationswörterbuch mit einer Systematischen Einführung und englischen Übersetzungen“*).

In the organization structure of the project four different roles can be defined: series editors, volume editors, authors, and publishing house.

The structure and format of the dictionaries have been set up by the two series editors. These have written an instruction manual for all members of the project. The dictionary volumes will all have the same structure, obligatorily conforming to certain formal specifications. All lemmas, namely the articles, are entered via a CMS to which the staff has password-protected access. The interactive form is set up in such a way that, upon entry, an XML version is created. The entry mask is part of a web-based editing system so that the authors can write the articles at any location. The series editors are responsible for the conception and organization of the whole series. They have arranged the frame conditions of the series and of the single volumes with the publisher, they decide upon the lists of sigla, abbreviations, etc., and they are responsible for the homogeneousness of the volumes in terms of their general structures. The series editors choose the volume editors and advise on general questions.

The volume editors are responsible for the compliance with the general rules in their own volume. The volume editors have to draw up an open list of lemma candidates, as well as provide the conceptual framework for the Systematic Introduction of their volume. The open list of lemma candidates

is provisional, and will continue to be regularly updated even after the work of the individual authors has begun. As a rule of thumb, it can be estimated that during the “article preparation phase” the number of lemmas will increase by approximately 20% compared to the provisional list drawn up at the end of the “editorial planning phase”. Individual authors will be contacted and invited to work on nets of conceptually linked lemmas. Another rule of thumb is that roughly 50 authors should work on 1,500 lemmas. This is the estimated number of lemmas of a single volume. Thus, the size of a printed dictionary in WSK would comprise approximately 800 pages.

By having 25 volumes, we will have about 1,000 authors involved in the project and the estimated sum of all lemmas we will have to treat in WSK is 50,000.

## 2.1 Specific Features of the WSK

In this section, selected features of the WSK will be discussed. The intended users of the WSK fall into three groups:

- (a) Students of all philological or linguistic disciplines, both in Germany and abroad, who need to read professional literature in German and English (semi-experts).
- (b) Academic instructors (in Germany and abroad) who teach students of group (a) (experts for some fields, semi-experts for others).
- (c) University graduates who have taken their degree in a philological and/or linguistic discipline and whose profession draws on their field(s) of study (experts for some fields, semi-experts for others).

However, the group of intended users of the WSK is not identical with the group of potential purchasers. In addition, purchasers will include the home institutions (universities, libraries, etc.) of the above-mentioned students and teachers ((a), (b)), as well as other, smaller categories (secondary school teachers, translators, academies, Goethe Institutes, etc.). At all events, the market for the WSK is clearly relatively large, and is moreover in a state of continuous renewal, given the constant (and presumably numerous) turnover of new students. Thus, lexicographic maintenance activities on the part of the dictionary publisher (new editions, updates, etc.) are being planned as well.

The WSK dictionaries will be multifunctional, with various dictionary functions weighted as follows (for the functions of dictionaries see Wiegand 2001):

- (i) Primary functions of the dictionaries include:
  - “comprehension function”: to help in the comprehension of texts
  - “informational function”: to provide specialized information on a given field.
- (ii) The secondary function of the dictionaries is:
  - “translation function”: to help in translating.

The purpose of each of the three dictionary functions is to cover a certain amount of usage situations for a particular dictionary of the WSK series.

The system of dictionary functions applies to every WSK dictionary, and the specific type of users of the respective dictionary determines the overall textual structure of each individual volume, its internal data distribution, the volume’s macrostructure, its cross-referencing system, the microstructure of the individual articles, and – last but not least – the dictionary basis for each volume.

## 2.2 The Dictionary Basis

The dictionary basis (i.e. the aggregate of all sources used in compiling the dictionary) will naturally differ by dictionary. It is determined for each volume by the volume editors. However, the dictionary

bases for all WSK dictionaries have the same structure. The dictionary basis for each dictionary includes, first of all, the group of *primary sources* for that dictionary. These are the German and English texts in which the terms dealt with in the dictionary actually appear. Additionally, there is the group of *secondary sources*. These include discipline-specific reference works in which the terms dealt with in the present dictionary have previously been dealt with. The used reference works will be listed in a register of secondary sources in the back-matter of the dictionary. The primary sources are given, always in identical format, in the “literature position” at the end of every dictionary article. In this way, the text basis of every dictionary is fully documented within the dictionary itself.

### 2.3 The Components of the Dictionary

Every WSK dictionary represents a heterogeneous composite text: a variety of component text-blocks belonging to different text genres are compiled together into a lexicographical whole. These components are displayed in the following list in the same order in which they will appear in all WSK dictionaries. (The symbol “o” indicates an *obligatory* component, while “f” denotes a *facultative* (optional) component.): Brief user introduction: in German, on the inside front cover (o); Title page (o); Table of contents: the dictionary’s component parts (o); Introduction by the volume editors (o); Preface by the series editors: only in WSK 1 (o); Detailed user introduction: laid out in identical format in all the WSK dictionaries (o); List of abbreviations: Alphabetical list of the sigla used in the Literature position (o), Other abbreviations: general abbreviations, abbreviations for dictionaries used in the texts of the articles (o); List of symbols (f); Phonetic transcription (f); Transliteration (f); List of illustrations which apply to several articles (f); Systematic Introduction (o); Alphabetical index to the Systematic Introduction (o); ALPHABETICAL WORD LIST (o); Alphabetical list of secondary sources (o); List of contributors (o); Index of English-German equivalents (o); Brief user introduction: in English, on the inside back cover (o).

Note that the above ordering scheme may be altered in minor respects prior to the printing of the first volume.

In the following, a few selected components of the dictionary will be briefly sketched (cf. also Schierholz/Wiegand 2004). Particular attention will be paid here to the Systematic Introduction and to the cross-referencing system (mediostrucuture) used in the main body of the dictionary (alphabetical word-list), for these are fundamental to the concept of the specialist “study and reference dictionary”.

Monolingual and bilingual special field dictionaries with a comprehensive Systematic Introduction can be found in many special fields, especially in the natural sciences. In linguistics, however, no such dictionary exists (cf. Haunstetter 2010; Schierholz/Wiegand 2004).

To prevent the following discussion from becoming too abstract, I will present the organizational scheme of the Systematic Introduction to WSK volume 1 (Grammar) here, as drawn up provisionally by Christa Dürscheid (Zürich), Pál Uzonyi (Budapest) and Stefan J. Schierholz (Erlangen). Christa Dürscheid was the volume editor of WSK 1 from 2004 to 2014, Pál Uzonyi has been the volume editor since 2016. The volume will comprise two sub-volumes (Morphology, Syntax); the Systematic Introduction, covering both sub-volumes, will appear only once, i.e. in the first volume “Morphology”. It must be noted that this Systematic Introduction is still in a work in progress, so that the subject division proposed here is not yet the final one. The Systematic Introduction of each WSK volume will be divided into numbered sections.

1. **What is grammar?** § 1 The term *grammar* in linguistic literature; § 2 Historical evolution of the term (grammar as system, grammar as theory, grammar as handbook); § 3 Grammar in WSK volume 1 (the most important components of grammar: *morphology* and *syntax*, justification for this distinction, broad and narrow conceptions of grammar).

2. **Possible ways of typologizing grammars** § 4 Diachronic vs. synchronic grammars; § 5 Prescriptive vs. descriptive grammars; § 6 Language-particular, comparative, and universal grammars; § 7 Scientific vs. practical grammars.
3. **Fundamental concepts of grammar** § 8 Form and function; § 9 Syntactic categories; § 10 Structures; § 11 Relations (syntagmatic and paradigmatic); § 12 Syntactic functions; § 13 Grammaticality and acceptability; § 14 Rules; § 15 Linguistic tests.
4. **Grammar as a system (here: narrowly conceived)** § 16 Morphology; § 17 Inflection (declension: noun, pronoun, article, declension/comparison: adjective, conjugation: verb); § 18 The word (syntactic, phonetic, and orthographic unity, morpheme, parts of speech (declinable, indeclinable), word groups); § 19 Syntax; § 20 The phrase; § 21 The sentence/clause (types of clauses, combinations of clauses, clause mood, main clauses, subordinate clauses, clause combinations: conjoined and embedded structures, sentence types: questions, requests, imperatives); § 22 Valence (concept of valence, government, verbal valence, valence of other parts of speech); § 23 Constituent and dependency structure (constituent structure, immediate and non-immediate constituents, dependency structure, head and dependent, valence).
5. **Grammar as a theory** (selected schools of grammatical theory) § 24 Traditional Grammar; § 25 Phrase-structure Grammar; § 26 Generative Grammar (context-free syntax, Transformational Grammar, X-Bar theory, Government and Binding Theory, Minimalist Program); § 27 Dependency Grammar (dependency grammar and valence grammar, valence theory, evolution of the concept: forerunners); § 28 Lexical Functional Grammar; § 29 Optimality Theory; § 30 Functional Grammar; § 31 Case Grammar; § 32 Pedagogical grammar (language acquisition, language competence, acquisition stages: native and foreign language).
6. **Grammar as a handbook:** § 33 Features of grammar books; § 34 Expectations of a grammar book (prescriptive: normative information, purposes: reference work, compilation of rules, reader); § 35 Users: students, professionals, experts, native speakers, foreign speakers; § 36 Examples of grammar books.
7. **Grammar and allied fields:** § 37 Phonology; § 38 Word formation; § 39 Text grammar; § 40 Pragmatics; § 41 Orthography/spelling.
8. **Alphabetical index to the Systematic Introduction**

The Systematic Introduction will probably consist of 60 to 70 pages, and is interwoven with the word list of the dictionary by a web of cross-references as follows. Each of the above 41 sections concludes with a cross-reference position, pointing to various lemmas of relevance in the word list of the dictionary. For example, at the end of § 7 (“Scientific vs. practical grammars”) the following cross-references appear in German alphabetical order (only a brief selection is given here): → *Abhängigkeitsgrammatik* [Dependency Grammar], *Generative Grammatik* [Generative Grammar], *Kategorialgrammatik* [Categorial Grammar], *Konstituentenstrukturgrammatik* [Constituent Structure Grammar], *Lernergrammatik* [Student Grammar], *Lexical Functional Grammar*, *Minimalistisches Programm* [Minimalist Program], *Optimalitätstheorie* [Optimality Theory], *Schulgrammatik* [school grammar], *traditionelle Grammatik* [traditional grammar], *Valenztheorie* [Valence Theory]. And at the end of § 14 (“Rules”) the following cross-reference address markers appear (brief selection in [German] alphabetical order): → *Generative Grammatik – Phrasenstrukturregel* [Generative Grammar – Phrase-Structure Rule], *Schulgrammatik* [school grammar], *Struktur* [structure], *Transformation*.

Users reading through the Systematic Introduction who may wish to learn more about topics pertaining to “scientific vs. practical grammars” can follow up on the cross-references given by the cross-reference mark. In most cases these will point to synopsis articles on various kinds of grammars. On the other hand, a user-*in-actu* looking up a specific term (e.g. “Lexical Functional Grammar”) may wish, after reading the article, to learn more about the broader conceptual field in which “Lexical Functional Grammar” is embedded. If this is the case, they will find the cross-referencing notation “® §



7” in the article, pointing to Section 7 of the Systematic Introduction. The WSK thus makes use of bidirectional cross-referencing, so that the Systematic Introduction is systematically interwoven with the dictionary articles: cross-references proceed from the former to the latter, and vice versa. This rich network of cross-referencing will serve to strengthen and enhance the “informational function” of the WSK; it will make it easy for readers to systematically familiarize themselves with whole blocks of terminology in any subfield they choose. On “Mediostrukturen” [cross-referencing networks] cf. Wiegand (2002).

The Systematic Introduction will significantly enhance the usefulness of the WSK dictionaries, especially in non-conflict-conditioned usage situations, e.g. in research contexts (e.g. in preparing seminar papers). Together with the cross-referencing network and the synopsis articles, the Systematic Introduction constitutes the essential components of the dictionary qua learning tool. The various possible functions of the Systematic Introduction to any given WSK volume can be summarized as follows:

- The Systematic Introduction can be read as an integral block of continuous text, providing the background knowledge necessary for dictionary consultation on points of detail in the given domain.
- By means of the alphabetical terminological index provided for each Systematic Introduction, readers can consult the Systematic Introduction to answer specific questions. Users who look up any particular issue in this way will be quickly led to further related terms.
- Since the dictionary articles refer to the relevant sections of the Systematic Introduction, the latter can be consulted selectively, in conjunction with the main dictionary, to obtain information on any specific point of interest.
- Since the sections of the Systematic Introduction refer to the articles in the dictionary, readers of the Systematic Introduction can augment their knowledge on any specific point by supplementary consultation of the dictionary articles.

## 2.4 The Dictionary Articles

In the following I will briefly discuss the format of the single dictionary articles, without giving a precise metalexicographical description of the different fields and structures that are entered into the format. Here I will lay out the relevant text fields, i.e. those which are fixed and invariant in form and those whose form is open.

All WSK dictionaries will comprise a heterogeneous collection of articles of three types:

- cross-reference articles
- partially condensed single articles
- partially condensed synopsis articles.

In the “single articles”, only a single terminological lemma will be dealt with, reflecting its normal denotation in texts. The single articles thus represent a maximal atomization of information. The “synopsis articles” will generally be devoted to broader terms, e.g. proper names which are relevant to the history of the discipline, etc. These articles will provide an overview of larger chunks of knowledge, with cross-referencing to lemmas that are dealt with in the single articles. The full abstract hierarchical microstructure is identical both in single articles and in synopsis articles, as are the corresponding text fields. For monosemous terms, both single articles and synopsis articles will have the following text fields (“o” = obligatory, “f” = facultative/optional): lemma position (o); position of English equivalents (o); position of the definition (o); position for further explanation and discussion (o); position for selected terminological references, including subfields; position for synonyms (f), position for antonyms (f); cross-reference position (o); position for the name of the author(s) (o); literature position (o).



The following article on the monosemous lemmas *binäres Merkmal* (*binary feature*) demonstrates the possible content of these article items (parts of it are translated into English):

lemma position	<b>binäres Merkmal</b>
German position of the definition	Merkmal eines Objekts, das auf der Kontradiktion zweier zueinander in Opposition stehender Eigenschaften beruht
position of English equivalent and English definition	<i>binary feature</i> : feature of an object which is based on the contradiction of two opposing properties.
Position for further explanation and discussion	<ul style="list-style-type: none"> <li>Traditionally found in phonology, the most basic type of feature is the distinctive feature, which, in contrast to a redundant feature, functions to distinguish meaning. In the structuralist framework, for example, phonemes are described as bundles of distinctive features, e.g. /p/ is [+consonant, -voiced, +bilabial, -nasal]. As such, it differs from /b/ solely according to voicing (the latter being [+voiced]). Since the inception of the phonological analysis of distinctive features in the 1950s, features have traditionally been specified by assigning them binary values, depending on whether the segment described by the feature possesses the property at hand or not. A positive value, [+], denotes the presence of a feature, while a negative value, [-], indicates its absence. In recent theory, however, phonologists have proposed the existence of single-valued, univalent features. These features, in contrast, can only describe the classes of segments that are said to possess them.</li> </ul>
position for synonyms	
position for antonyms	↔ monovalent feature
cross-reference position	→ feature // cf. WSK 1.1: feature, binary feature, WSK 4: binary feature
name of the author	[AMN]
literature position	<p>📖 BACON, F. [1623] De dignitate et augmentis scientiarum. In: MAYER, P. [Hg. 1829] De dignitate et augmentis scientiarum libri IX. Nürnberg ■ CHOMSKY, N./ HALLE, M. [1968] The Sound Pattern of English. New York ■ JAKOBSON, R./ HALLE, M. [1956] Fundamentals of Language (JanLing-Minor-H 1). Den Haag ■ JAKOBSON, R./ FANT, G./ HALLE, M. [1969] Preliminaries to Speech Analysis. The Distinctive Features and their Correlates. Cambridge, MA [etc.].</p>

In the following, further information is given about the field “Further explanation and discussion”. This field consists of a text item comprising at least one full sentence, and is separated from the definition by a line-initial bold-faced bullet (•) serving as a microstructural indicator. The series editors will provide only very general guidelines to the format of this text field. The most important guideline is: lemmas that are similar in conceptual type should all have the same type of text items. A “type” of a text item can, for example, be specified as a grid which is filled by the author of the article. Groupings of similar lemmas into types would include e.g.:

- antonyms, synonyms, hyponyms, hyperonyms, meronyms, etc.
- sound, syllable, word, sentence
- grapheme, phoneme, moneme, morpheme, lexeme, etc.
- Ablaut, Umlaut, breaking, etc.
- compounding, derivation, affixation, etc.
- Dependency Grammar, Functional Grammar, Generative Grammar, etc.
- verb, noun (substantive), adjective, adverb, particle, etc.

The grid will present those features and concepts that article authors should take into consideration when treating lemmas of a particular type. The actual textual form of the text item will be left to the individual authors, with the provision that text condensation concerning the syntax is not permitted.

The text items may include, *inter alia*, such items as: expansion of the definition, mention of alternative views and approaches, literature references, brief factual discussion, historical outline, diagrams, graphs, illustrations, linguistic examples.

I omit giving a more detailed specification of the articles' format here. Experience has shown that linguistic terms of all sorts can be accommodated in a user-friendly way using the above format.

### 3 Current Situation

The project started in 2004 and the original goal of finishing the first volumes five years later (2009) could not be met. This is due to technical and personal reasons and maybe the series editors underestimated the complexity of such a huge project. This does not concern the work of the series editors or volume editors, but in most cases the work and reliability of the single authors who were very often not able to finish their articles in time. The online version of 2013 pushed the whole project forward, because the publishing house sold it very well and thus they were (and are) able to support the ongoing work regularly. Nowadays, more and more articles are uploaded to the online version every year, so that the online version should contain about 20,000 articles in 2020. In parallel to this, the first print dictionaries are in preparation and we think there is a realistic chance to have the first three volumes on the market by 2019.

This project is necessary for the area of our special field, Linguistics and Communication Science. It was, is and will be a lot of work which depends on idealists with considerable courage and great endurance, but, luckily, there are enough people in our subject field who possess these qualities.

### References

- Haunstätter, Kerstin (2010). Glottopedia – die kostenlose Online-Enzyklopädie im Vergleich mit den Wörterbüchern zur Sprach- und Kommunikationswissenschaft. In: *Lexicographica* 26, pp. 229-247.
- Schierholz, Stefan J. (2007). Neue Fachwörterbücher für die Sprach- und Kommunikationswissenschaften. In: Di Meola, Claudio/ Gaeta, Livio/ Hornung, Antonie/ Rega, Lorenza (Hrsg.): *Perspektiven Zwei. Akten der 2. Tagung Deutsche Sprachwissenschaft in Italien*. Rom: Istituto Italiano di Studi Germanici (Italienische Studien zur deutschen Sprache 3) Roma 2007, pp. 223-234.
- Schierholz, Stefan J. (2008). Die Übersetzung linguistischer Fachtermini. Eine Studie zu den Lemmata in den WSK. In: *Akten des vierten Internationalen Kolloquiums zur Lexikographie und Wörterbuchforschung in Maribor 2006 (Germanistische Linguistik 2007)*. Hildesheim, pp. 62-81.
- Schierholz, Stefan J. (2010). Die Fachwörterbuchreihe „Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK)“. In: Jarillot Rodal, Cristina et al. (Hrsg.): *Bestandsaufnahme der Germanistik in Spanien. Kulturtransfer und methodologische Erneuerung*. Bern [etc.], pp. 113-122.
- Schierholz, Stefan J. (2015). WSK: Ein Fachinformationssystem zur Sprach- und Kommunikationswissenschaft in Deutsch und Englisch als Online- und als Print-Version. Herausgegeben von Stefan J. Schierholz und Herbert Ernst Wiegand. In: Robles i Sabater, Ferran/ Calanas-Continente, Jose-Antonio (Hrsg.): *Die Wörterbücher des Deutschen: Entwicklungen und neue Perspektiven (Spanische Akzente – Studien zur Linguistik des Deutschen 2)*. Frankfurt/Main [etc.], pp. 13-41.
- Schierholz, Stefan J./ Herbert Ernst Wiegand (2004). Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK). Eine neue Konzeption der linguistischen Fachlexikographie und ihre computergestützte Praxis. In: *Lexicographica* 20, pp. 164-264.

- Wiegand, Herbert Ernst (2001). Was eigentlich sind Wörterbuchfunktionen? Kritische Anmerkungen zur neueren und neuesten Wörterbuchforschung. In: *Lexicographica* 17, pp. 217-248.
- Wiegand, Herbert Ernst (2002). Altes und Neues zur Mediostruktur in Printwörterbüchern. In: *Lexicographica* 18, pp. 168-252.
- Wiegand, Herbert Ernst (2003). Wörterbuch zur Lexikographie und Wörterbuchforschung (WLWF). Dictionary of Lexicography and Dictionary Research. In: *Wissenschaftliche Lexikographie im deutschsprachigen Raum*. Im Auftrag der Heidelberger Akademie der Wissenschaften hrsg. v. Thomas Städtler. Heidelberg, pp. 417-437.
- Wiegand, Herbert Ernst (2004). Überlegungen zur Mediostruktur in Fachwörterbüchern. Auch am Beispiel des „Wörterbuchs zur Lexikographie und Wörterbuchforschung“. In: *Lexikalische Semantik, Phraseologie und Lexikographie. Abgründe und Brücken*. Festgabe für Regina Hessky. Hrsg. von Rita Brdar-Szabó und Elisabeth Knipf-Komlosi (Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 57). Frankfurt/Main [etc.], pp. 339-365.
- Wiegand, Herbert Ernst (2006a). Die „Wörterbücher zur Sprach- und Kommunikationswissenschaft“ (WSK) und ihre Benutzungsmöglichkeiten im Fach Deutsch als Fremdsprache. In: Dimova, Ana/ Jesenšek, Vida/ Petkov, Pavel (Hrsg.): *Zweisprachige Lexikographie und Deutsch als Fremdsprache*. Drittes Internationales Kolloquium zur Lexikographie und Wörterbuchforschung (GL 184/185). Hildesheim [etc.], pp. 1-35.
- Wiegand, Herbert Ernst (2006b). Das Lern- und Konsultationswörterbuch. Ein neuer Fachwörterbuchtup am Beispiel der „Wörterbücher zur Sprach- und Kommunikationswissenschaft“ (WSK). In: *Lexikos* 16 (AFRILEX-reeks/series 16), pp. 1-17.
- Wiegand, Herbert Ernst (2009). *Diccionario de aprendizaje – diccionario de consulta*. Los *Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK)*: un nuevo tipo de diccionario especializado. In: Fuentes Morán, María Teresa/ Model Benedikt A. (eds.): *Investigaciones sobre lexicografía bilingüe* (Colección Lexicografía 1). Granada, pp. 11-28.



# When Learners Produce Specialized L2 Texts: Specialized Lexicography between Communication and Knowledge

**Patrick Leroyer<sup>1</sup>, Henrik Köhler Simonsen<sup>2</sup>**

<sup>1</sup>Aarhus University, <sup>2</sup>Copenhagen Business School

E-mail: [pl@cc.au.dk](mailto:pl@cc.au.dk), [hks.msc@cbs.dk](mailto:hks.msc@cbs.dk)

## Abstract

This article discusses the theoretical distinction between communicative- and cognitive-oriented dictionary use situations and explores whether or not this sharp distinction is still valid at a time when users do not use dictionaries but instead online language resources, particularly in learning environments. The paper seeks to answer this research question based on empirical data from a user study conducted at Copenhagen Business School in 2017. We carried out a controlled experiment involving ten test persons and the user study produced ten screen recordings, ten specialized texts, ten self-assessments and ten teacher-assessed rubrics. On the basis of our empirical data we found that the sharp distinction between communicative and cognitive-oriented dictionary use situations does not seem to make much sense anymore when users, to an increasing extent, do not use dictionaries but instead online language resources. We found that specialized language and specialized knowledge are completely intertwined, mutually interdependent and form a dialectic relation, which in fact can be identified by analyzing the test person's information search and retrieval processes. We also found that new, modern language resources make it possible to make searches in text directly and to take full advantage of the dialectic relation between specialized language and specialized knowledge.

**Keywords:** specialized lexicography, cognitive functions, communicative functions, situation distinction, learning situations, functional interdependence

## 1 Introduction

In this paper, we will reflect upon the results of a user study we have recently conducted within the field of specialized lexicography. (Leroyer & Simonsen 2017). We have compared user behavior during L2 text production processes with assistance from online multilingual language resources on the one hand, and the corresponding text production behavior with assistance from specialized dictionaries specifically designed for the purpose of specialized text production on the other hand, cf. also Nesi (2013) and Lew (2016). Online multilingual language resources here include L1 and L2 texts (in this case company websites, LinkedIn pages, etc.) as well as assistance from Google Translate for L1>L2 translation (when students use the system to translate small text segments from L1 to L2, or simply to find or check equivalents), see also Table 1 in Section 5 for an overview of the resources and their use.

The comparison of user behavior is largely based on a metalexicographic theory, which has had a great impact for the past 25 years, not only in the Scandinavian countries, where it was devised, but also in many other places in the world.

The theory is known as the function theory (Bergenholtz & Pedersen 2017; Tarp 2009; Fuertes Oliveira & Tarp 2014) and is based on an axiomatic distinction between communicative and cognitive-oriented dictionary use situations depending on whether the user needs communicative assistance to write an L2 text, or whether the user needs knowledge of the specialized language itself or of the subject field as support for the L2 text production.



The sharp distinction between communicative and cognitive-oriented dictionary use situations (Bergenholtz & Pedersen 2017) originated in a didactic reflection upon teaching specialized language and communication, and it was a logical approach in connection with the functional design of specialized dictionaries, and particularly the selection and presentation of the relevant lexicographic data.

However, it is difficult to acknowledge this distinction in situations where users do not use dictionaries, but make use of online language resources to assist their specialized L2 text production processes. Based on the findings of our study, it seems that this behavior tends to blur and even annihilate the communicative vs. cognitive distinction, as both seem to be interdependent and dynamically embedded in each other.

## 2 Research Questions

The underlying research questions of this paper are firstly to analyze to what extent the theoretical distinction between communicative and cognitive-oriented dictionary use situations is still valid in situations where users do not use dictionaries significantly, but make an extensive use of online language resources, and secondly to reflect on the theoretical implications of our findings.

## 3 Methodology

The empirical data were collected by means of a controlled experiment during which ten students taking an MSc in International Business Communication (both male and female, and all advanced learners of English for special purposes) at the Copenhagen Business School were asked to solve a concrete L2 text production task based on a case. The experiment produced ten texts, ten screen recordings, ten self-evaluations and ten teacher-formulated rubrics. Although ten students seems as a small group, they are fully representative of students enrolled in the same MSc program. The text production task is also completely in line with the assignments they normally have to write during the term, and with their normal work conditions in an online environment.

The test persons were allowed to use all available language resources for L2 text production, including online dictionaries, online text resources as well as Google Translate, and were instructed to:

- A. Write a LinkedIn article (sales text) in Forum in Moodle
- B. Perform a self-assessment in the quiz module in Moodle
- C. Record their working process by means of an integrated screen recorder
- D. Upload the text, self-assessment and screen recording to Moodle<sup>1</sup>

Prior to the analysis of the empirical data, we formulated a number of measuring points. We wanted to understand, analyze and look further into the axiomatic difference between communicative and cognitive-oriented dictionary use situations, to be able to discuss the relevance and potential development of this so far quite useful distinction. The ten test subjects were asked to write a sales-oriented article to be published on LinkedIn, and the data gave us ample opportunity to study what really takes place in this context.

For each of the test subjects we were able to analyze a screen recording of how the student worked, which tools (s)he used and how (s)he used them, and to analyze what the student wrote in the text, what needs (s)he had as to specialized language elements and what needs (s)he had as to specialize knowledge.

<sup>1</sup> Assessment data were generated for the purpose of further research into the correlation between text production processes and output quality, but will not be used in this article, where focus is on text production processes only.

In the following section of our paper, we will outline and discuss some existing theoretical considerations on communicative and cognitive dictionary use situations related to text production situations in a learning environment.

## 4 Communicative and Cognitive Dictionary Use Situations

The underlying assumption of our paper is that specialized lexicography in our opinion so far has been based on a highly abstract distinction between two fundamentally different use situations when dictionaries can be of use and are actually used (Tarp 2006:58).

Tarp distinguishes between communicative dictionary use situations, when there is a communication-related need in connection with an ongoing or planned communication task, and cognitive dictionary use situations, when there is a knowledge-related need.

This distinction is one of the cornerstones upon which the function theory is based. The distinction is still valid now ten years later, as will be clear from the following citation from Agerbo and Bergenholtz (2017), which describes communicative and cognitive dictionary use situations as follows:

Communicative situations: the need to get information, which in a specific situation, is necessary in order to accomplish successful communication

- situation in which a person needs help to understand parts of a text
- situation in which a person needs help to formulate parts of a text
- situation in which a person needs help to translate parts of a text

Cognitive situations: the need in a specific situation to acquire knowledge, which you do not already have

- situation in which a person needs to acquire knowledge about a specific (singular) phenomenon or about a more complex theme
- situation in which a person has a goal-oriented need to learn something, either on his/her own or by being taught by someone. (Agerbo & Bergenholtz 2017:35-36)

However, Agerbo and Bergenholtz does realize that dictionary use situations are dynamic. In fact, the authors describe dictionary use situations as follows:

When we use the word situation, it refers to a time at which a certain information need occurs. Such a situation may occur as a result of another situation and thereby be a part of a course of events. Of course, such a situation is related to the preceding situations in the same course of events. (Agerbo & Bergenholtz 2017:24).

In summary, the existing understanding of dictionary use situations still seems to be based on a sharp distinction between communicative and cognitive dictionary use situations, respectively. We believe that the distinction is still very useful when it comes to designing specialized dictionaries based on the function theory, especially because the distinction enables the lexicographer to facilitate the selection and presentation of relevant lexicographic data in specific use situations.

However, the situation driven data selection and presentation does not appear to be that relevant anymore, when users instead use online language resources. According to our data, students in fact do not use dictionaries that much anymore, as noted in Kernerman (2013), which, with a reference to Frank Zappa, notes that “one could say dictionaries are not dead, they just smell funny”.

Things change, and we therefore need to take a critical look at existing theory. As will appear from the following sections of our paper, we will present some new theoretical considerations on communicative and cognitive dictionary use situations based on our empirical data.

## 5 Empirical Data

As already pointed out above, the user survey resulted in ten screen recordings of approximately one hour each, ten LinkedIn articles of approximately one A4 page, ten self-assessments and ten rubric-based teacher assessments.

Table 1 below gives an overview of the user behavior of the ten test subjects and of all resources used for producing the text and post-editing it:

Table 1: User behavior characteristics and resources used to do the assignment

Test Person	User Behavior Characteristics	Test Person	User Behavior Characteristics
1	Produces English text Post-edits English text Reads <i>About us</i> on company's website Reads company's LinkedIn page Watches video on company's website	6	Produces English text Reads case assignment Looks up words on Google Translate Post-edits English text
2	Reads about genre How to write at LinkedIn Reads additional online text resources Produces English text Looks up words on Google Translate Post-edits English text	7	Produces English text Reads case assignment Looks up words on Google Translate Post-edits English text
3	Activates online dictionaries Produces English text Looks up words on V&B, ØKON and BNC Looks up words on Google Translate Post-edits English text	8	Produces English text Used integrated Spell Checker in Word Looks up words on Google Translate Post-edits English text
4	Produces English text Used integrated Spell Checker in Google Docs Reads case assignment Post-edits English text	9	Produces English text Reads online resource about blog posts Reads similar assignment Looks up words on V&B Post-edits English text
5	Produces English text Looks up words on Google Translate Looks up words in BNC Looks up words on Wiktionary Post-edits English text	10	Reads online resource about blog posts Writes structure for LinkedIn article Produces English text Looks up words on V&B Reads similar assignment

Column two in Table 1 outlines the most important characteristics of the user behavior of each test person based on our analysis of the screen recordings, i.e. it shows how the test person did the assignment and which actions (s)he person performed.

As will appear from Table 1 above, the ten test persons seemed to have three different approaches to the focal task.

One group (test persons 3, 5, 7 and 8) seemed to have a clear communicative approach. In fact, this group predominantly used online dictionaries, online language resources and Google Translate, and seemed to focus on producing the English text.

Another group (test persons 4 and 6) seemed to have a mixed approach with both communicative and cognitive use situations, and they also used both Google Translate and different online resources to learn about the structure of a LinkedIn articles and to produce the English text.

Finally, the third group (test persons 1, 2, 9 and 10) predominantly had a cognitive approach focusing on learning as much as possible about the genre, the product, the company, the market and the language used to produce a LinkedIn article.

The first major observation that we made during our analysis of the screen recordings was the very limited use of dictionaries. It was interesting to learn that our test persons almost did not use dictionaries. In fact, we have ourselves observed the increasingly limited use of dictionaries by our own students. However, it did come as a surprise to us that the use of dictionaries was so limited, because the test persons were in fact introduced to a number of online dictionaries (V&B, FAG, ØKON, OED) as well as to a number of online resources, such as Wiktionary, Google Translate and the British National Corpus (BNC) prior to the user study.

This observation supports similar ones described in Leroyer and Simonsen (2017) in a discussion of whether language resources can replace dictionaries, and whether there is a measurable quality of the texts translated and produced by means of dictionaries or language resources. A similar conclusion can be found in Bundgaard (2017), which concludes that even professional translators seem to be abandoning dictionaries, in this case in-house term banks.

The second major observation that we made during our analysis of the screen recordings was that our test persons seemed to have a divergent approach to actual dictionary use situations, in contrast with those situations provided by the theory. The screen recordings showed that our test persons did not seem to behave differently with regard to communicative and cognitive dictionary use situations. In fact, our test persons seemed to use online resources to assist them in both contexts, and did not at all seem to behave differently. Our test persons simply seemed to use online resources to be able to communicate in order to fulfil the aim of their learning assignment, and thereby to be able to obtain knowledge of genre, market, company and medium at the same time and by means of the same online resource. In other words, our test persons seemed to behave the same in both dictionary use situations, as they are interdependent and dynamically embedded in each other. Producing the text seems to coincide with learning how to produce the text.

The test persons also seemed to use the online language resources in connection with both communicative and cognitive-oriented questions, even when the question was a communicative-related one. We can thus confirm our findings in Leroyer and Simonsen (2017), because our test persons also asked Google cognitive-oriented questions like "What is Immersion?" or "What is a LMS?" in connection with writing about the Learning Management System Immersion.

These findings thus question the way functional theory has traditionally treated communicative and cognitive dictionary use situations, and ultimately the way lexicographers should design dictionaries when it comes to text production situations.

The third major observation that we made during our analysis was that specialized dictionaries and online resources for text production are not only used by students to communicate, as one would normally expect, but also to learn how to communicate. Our test persons seemed to use the specialized dictionaries and online resources to learn how to write, and thus these resources support learning situations.

When analyzing the screen recordings, we discovered that when students are in a learning situation then text production is both characterized as cognition and metacognition. Text production is subject to reflections by the students on their own learning process, and we found out that communication

(the actual writing process) serves as a sort of learning tool. Our test persons are in a learning situation dictated by learning outcomes and they learn about the market, companies, genres and corporate communication by writing to a specific market, on behalf of a specific company, about a specific product, and using relevant genre conventions by writing a LinkedIn article.

Overall, we discovered that text production, as a learning situation, is cognition, a fact that finding fundamentally challenges the existing theoretical understanding of dictionary use situations in so far as text production is concerned, and changes our understanding of the axiomatic distinction between communicative and cognitive-oriented dictionary use situations.

Finally, the analysis revealed something we also identified in our user survey in Leroy and Simonsen (2017). The general IT competences of the test persons do play an important role in how they performed and how quickly they worked. The analysis also clearly showed that all test persons relied heavily on the integrated spell checker and the integrated synonym dictionary in Word.

The next section will focus on a discussion of our findings, and we will outline three overall theses on what we describe as “special text production situations”.

## 6 Analysis and Discussion

As already pointed out above, we propose the expression special text production situations when learners produce texts. In fact, a somewhat similar approach is suggested by Fuertes-Olivera (2018:275), who writes that the specialized dictionary can include “Texts for improving users’ writing skills and hyperlinks to open data sources”. In our opinion, the word “improving” opens up a discussion of learning situations and learning processes.

Based on our analysis, we argue that learning processes in concrete text production situations are dependent on the internal, autonomous knowledge creation processes and metacognitive reflection processes during the actual application of knowledge. All this takes place when the student produces a text. The learning processes that take place during the internal and autonomous text production processes are crucial to remember, and in our opinion play a major role in the selection, access, use and design of lexicographic resources.

We would argue that what we have here are extra-lexicographic “special text production processes”. They are not communicative use situations, but instead purely cognitive use situations of a special type and as we already pointed out above, the existing theoretical literature has so far not discussed this special situation. Our test persons were in a learning situation, and that dictated their knowledge acquisition processes, their knowledge application processes and their information retrieval processes in a constantly reciprocal interaction with the actual L2 text production processes.

Our empirical data showed that while producing the L2 texts, the test persons

- Learn about the product (in this case a Learning Management System) in L2
- Learn about sender, receiver, message (in this case LMS specialists, university partners, market, customers) in L2
- Learn about media and channels (in this case blogs and LinkedIn) in L2
- Learn about genre and text conventions (in this case sales blogs and LinkedIn articles) in L2
- Learn about rhetoric, argumentation and sales phrases (in this case pathos, AIDA etc.) in L2
- Learn about special terminology (in this case about learning software) in L2
- Learn about IT tools and their ability to assist specialised text production in a foreign language.



The goal of the test students was not to communicate (the communicative situation of the assignment is also a mock situation), but to learn and become better at producing texts and communicating. So the cognitive process of learning takes place during the actual text production phases. In other words, the knowledge acquisition processes and the knowledge application processes dictate all aspects of the text production process, including the metacognitive processes before, during and after the text production process.

We argue that text production in a learning situation is both cognition and metacognition – and that communication as such is a learning tool catalyst. Text production is the arena upon which the user learns about product, market, genre and so on by writing about these things, and this very fact challenges the existing understanding of lexicographic processes and the way lexicographic resources are designed.

Test persons 1 and 2, for example, do not produce texts to communicate. They clearly produce a text to learn how to communicate. They do not look up words to write, but to learn how to write. As will appear from the overview in Table 1, test persons 1 and 2 spent the majority of their time learning about the product, the company, the market, the genre, the channel, the medium and the language used. So their information search and text production processes were dictated by the cognitive learning situation.

If we take the course from which our test persons were recruited as an example, the learning objectives are described as follows:

At the end of the course, students are expected to be able to

- Analyze, reflect on and be able to produce a professional solution to a given communication challenge
- Be able to reflect on and enter into a professional dialogue with stakeholders about the appropriateness of a given solution to a communication challenge relative to the strategic aims of the organization
- Use grammatically correct and situationally and culturally adequate, professional language to communicate the information in question across the relevant media to the relevant target groups in writing as well as in spoken language (CBS 2018).

Based on the learning objectives for this module, the teacher decided to realize these learning objectives by especially focusing on the following study activities (cf. Laurillard 2012).

- Acquisition
- Production
- Practice

Based on these three study activities, the teacher planned and facilitated a number of learning activities including the case assignment, which the test persons was asked to solve, i.e. to produce a LinkedIn article about a specific type of learning management system for a specific type of target group.

So based on these learning objectives, study activities and the learning activity in question, a model for a learning-supporting, specialized lexicographic resource for L2 corporate communication might look like Figure 1, below.

As can be seen in the figure, the basis of everything is the learning situation.

The five lower boxes describe what the students need to learn during the defined learning activities, while the upper boxes describe what kind of lexicographically selected and ordered information and data they need to learn about the product, market and competitors, for example.

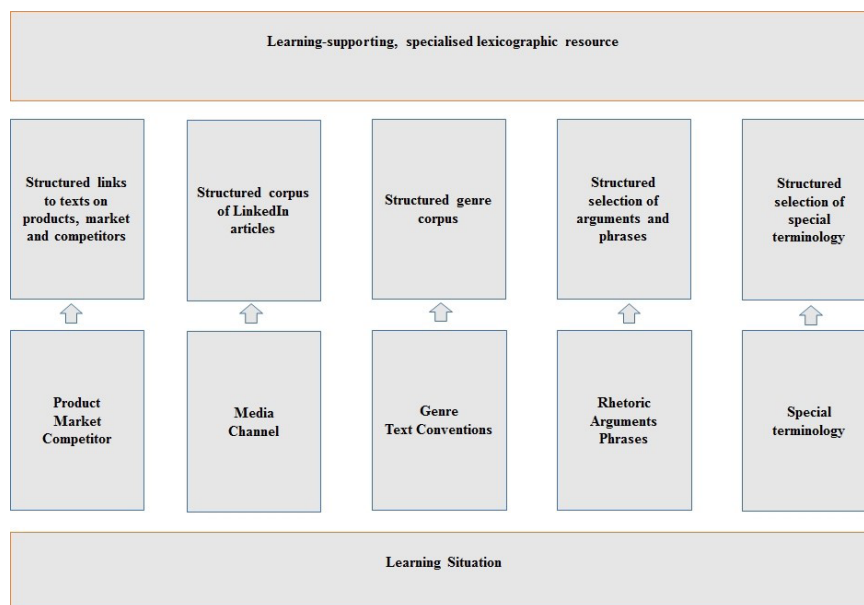


Figure 1: Learning-supporting, specialized lexicographic resource

In summary, it is argued that specialized lexicography could and should play an integrated role in learning. Not just as a tool that can be deselected by the user, but as an integrated and indispensable part of the learning process.

In fact, the very lexicographic method of selecting, structuring and ensuring easy access to data would be a huge benefit for everybody who wants to learn new competencies, just as modern technology in fact already makes it possible to integrate learning-supporting elements in already existing learning platforms.

Specialized dictionaries, including specialized and structured corpora or linked data, are as we see it 'predetermined' to support learning, and based on our findings and our theoretical considerations we have formulated three theses, as follows:

#### Thesis 1

Specialized dictionaries for text production are predetermined to develop into specific learning instruments in specific learning situations.

#### Thesis 2

Specialized language and specialized knowledge are related to each other through a dialectic relation, which is constitutive of information search and retrieval processes. So far, this distinction has been useful, as it has been aimed to speed up search and data retrieval processes through the design of monofunctional dictionaries, and to reduce search results and data accumulation. Yet, the cost has been that such lexicographic design structures have led to a fragmentation of the lexicographic data at the expense of text data.

#### Thesis 3

New information technology, and particularly linked data formats, have completely altered the basic grounds for human data access and data processing. Specialized online language resources that make use of linked data formats allow users to make searches in texts directly, and to take advantage of the dialectic relation. It may be trivial to point out, but we seldom read words in isolation nor do we write them in isolation, but in texts. New technology now makes it possible to build up word nets from dictionary resources and expand relations between words.

## 7 Conclusion

In this article, we have reflected on the sharp distinction between communicative and cognitive-oriented dictionary use situations based on empirical data from ten test persons.

To our first research question about whether we should still maintain a distinction between communicative and cognitive-oriented dictionary use situations in L2 text production contexts, we found that the distinction does not make much sense anymore when users make extensive use of online language resources and are immersed in a learning situation.

To our second question about the potential theoretical implications of our findings, we first found that this changed behavior calls for new theoretical considerations on dictionary use situations. We found that it does not make much sense talking about two separate situations, when users see them as mutually interdependent and embedded in each other. We also found that the axiomatic distinction between communicative and cognitive dictionary use situations does not make much sense in learning situations. Our empirical data showed that a per se communicative-oriented situation (writing a text) in reality is cognitive-oriented, because it is about learning how to produce a text instead of simply producing a text.

It is now up to specialized lexicography to formulate new theories that are based on a dialectic and mutually interdependent relation between communicative and cognitive-oriented dictionary use situations for learning purposes.

Specialized lexicography needs to formulate new theories and use these novel insights to design information search systems, which on one hand have inherited the user orientation and determination of functional lexicography, and which on the other have individualized needs-adapted access to relevant information in online text resources. Also, special attention should be given to specific information needs arising during the course of post-editing processes, so users can keep a critical distance to the relevance of the assistance they get and even question its authority, as needed.

## References

- Agerbo, H. & Bergenholtz, H. (2017): Types of Lexicographical Information Needs and their Relevance for Information Science. *Journal of Information Science Theory and Practice* 5(3), 15-30.
- Bergenholtz, H. & Pedersen, H. (2017). Types of lexicographical information needs and their relevance for information science. *Journal of Information Science Theory and Practice*. 5, 2, 23-38.
- BNC = British National Corpus. London: Accessed at <https://corpus.byu.edu/bnc/> [01/032018]
- CBS (2018): KAN-CICOO1008U Communication Management. Accessed at: <http://kursuskatalog.cbs.dk/2017-2018/KAN-CICOO1008U.aspx?lang=en-GB> [01/03/2018]
- Bundgaard, K. (2017): (Post-editing) A workplace study of translator-computer interaction at Textminded Danmark A/S. PhD dissertation. Aarhus: Aarhus Universitet, Department of Management.
- FAG = Thomas Arentoft Nielsen, Charlotte Langkilde, Jørgen Høedt: Gyldendals Røde Ordbøger – Fagordbog, København: Gyldendal. Accessed at: <http://ordbog.gyldendal.dk/esc-web.lib.cbs.dk> [01/03/2018]
- Fuertes-Olivera, P. A. (2018). Dictionaries for text production. In: Pedro A. Fuertes-Olivera (Ed.) *The Routledge Handbook of Lexicography*. Oxford/New York: Routledge, 267-283.
- Fuertes-Olivera, P. A. and Tarp, S. (2014). Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminology. *Lexicographica*, Series Maior 146. Walter de Gruyter. Germany.
- Kernerman, I. (2013). Kernerman News, Number 21, July 2013, 1-2.
- Laurillard, D. (2012). *Teaching as a Design Science: Building Pedagogical Patterns for Learning and Technology*. New York and London: Routledge.
- Leroyer, P. & Simonsen, H. K. (2017): Sprogressourcer kontra ordbøger: hvad er bedst? In: *Nordiske Studier i Leksikografi* 14, 2017, Rapport fra 14. Konference om Leksikografi i Norden – Island 30. maj–2. juni 2017.

- Lew, R. (2016). 'Can a Dictionary Help you Write Better? A user Study of an Active Bilingual Dictionary for Polish Learners of English'. *International Journal of Lexicography* 29 (3): 353-366.
- Nesi, H. (2013). 'Researching users and uses of dictionaries' in H. Jackson (Ed). *The Bloomsbury Companion To Lexicography*, 62-74. London: Bloomsbury.
- OED = OED Oxford English Dictionary. London: Accessed at: <http://www.oed.com.esc-web.lib.cbs.dk/> [01/03/2018].
- Simonsen, H. K. (2009): User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. In: *Hermes – Journal of Language and Communication Studies* 46(2011), 75–101.
- Simonsen, H. K. (2014): Mobile Lexicography: A Survey of the Mobile User Situation. In: Abel, Andrea, Chiara Vettori & Natascia Ralli (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014, Bolzano/Bozen. Bolzano/Bozen: EURAC research. Volume I, 249-261.
- Tarp, S. (2009). 'Reflections on lexicographical user research'. *Lexikos* 19, 275-296.
- V&B = Vinterberg & Bodelsen: *Gyldendals Røde Ordbøger – Vinterberg & Bodelsen*, København: Gyldendal: Accessed at: <http://ordbog.gyldendal.dk.esc-web.lib.cbs.dk> [01/03/2018].
- ØKON = Annemette Lyng Svensson: *Økonomisk Ordbog Engelsk-Dansk/ Dansk-Engelsk*. Samfundslitteratur. Accessed at: <http://ordbog.gyldendal.dk.esc-web.lib.cbs.dk> [01/03/2018].

# New Platform for Georgian Online Terminological Dictionaries and Multilingual Dictionary Management System

***Tinatin Margalitadze***

*Ivane Javakhishvili Tbilisi State University*

*E-mail: tinatin.margalitadze@tsu.ge*

## Abstract

The *English-Russian-Georgian Technical Online Dictionary* is the first ‘digitally born’ online dictionary of Georgian, created in a Multilingual Dictionary Management System (MDMS), specially developed for this project. The *Technical Dictionary* is the third specialized dictionary created by the same lexicographic team since 2009 (after the *English-Georgian Biology Online Dictionary* and *English-Georgian Military Online Dictionary*). Work on specialized vocabulary of different domains has revealed that terminology has evolved, particularly during the last 10 – 15 years. The traditional, standard requirements for monosemy and mononymy are not always observed in actual terminological work. There are numerous instances of terminological synonymy, many terms are polysemous, frequently developed as a result of metaphorical change of the primary meaning; there are many multiword terms consisting of two, three or even more words, giving rise to numerous terminological abbreviations; synonymous terms may belong to different stylistic registers, which requires the introduction of some stylistic labels in terminological entries. Rapid development of science and technology in the 21<sup>st</sup> century caused the appearance of an abundance of new concepts and consequently new terms. The resulting influx of new terminology into the Georgian language dictates the necessity to provide definitions of such terms alongside their Georgian equivalents. Introduction of collocations and examples of usage of terms is another issue that comes to the foreground of lexicographic description of terms.

These observations about modern terminology, which is discussed in the first part of the paper, became the basis for the development of a new platform for English-Georgian online bilingual terminological dictionaries and MDMS, as outlined in this paper.

**Keywords:** structural and semantic characteristics of modern terminology, Multilingual Dictionary Management System, a platform for English-Georgian online bilingual terminological dictionaries

## 1 Introduction

The end of the 20<sup>th</sup> century and the beginning of the 21<sup>st</sup> century has been marked by great changes and rapid developments in science and technology. Advanced technologies have penetrated into and drastically changed practically all aspects of our everyday lives. Moreover, the rate of technical progress is constantly increasing, leading to the introduction into our routine activities of certain things which not very long ago would seem to belong to the realm of science fiction. Such rapid development of any field of science implies the spontaneous generation of new scientific terms, and the influx of such terms in nearly every field of knowledge is another characteristic feature of our era.

This great increase in the number of new terms in various domains has caused some changes in the structural and semantic characteristics of modern terminology, leading to new requirements in the presentation of information in terminological entries, including bilingual terminological entries.



The distinction between terminological items and lexical items is also increasingly blurred. The nature of the linguistic items discussed by terminologists has undoubtedly evolved over the last 10-15 years. The inclusion in specialized electronic glossaries and term bases of items such as modal auxiliaries, complete sentences, collocational or phraseological patterns, images, diagrams and pictograms is driven by the needs of the target users and the requirements of modern multilingual communication. In this respect, the road from lexicography to terminology is more a continuum, a cline, rather than a hard-and-fast dichotomy (Fontenelle 2014 : 44).

The aim of the present paper is to share some observations about developments in modern terminology accumulated during the work on three terminological dictionaries in recent years (representing around 50,000 terminological entries altogether), each dictionary comprising many domains in its turn. The *English-Georgian Military Online Dictionary* (2009 – 2010), covering such fields as tactics, operations, maneuvers, trainings, units, personnel, ranks, transportation, weapons, equipment, logistics, and so on. The *English-Georgian Biology Online Dictionary* (2012 – 2014), including terms from the following fields: zoology, botany, paleontology, anatomy, physiology, genetics, immunology, biotechnology, molecular biology, etc. *English-Russian-Georgian Technical Online Dictionary* (2014 - 2016) comprising such fields as electronic and computer technologies, machinery and spare parts, metallurgy, the automobile industry and car-making, road building, information and manufacturing technologies, the mining industry, construction engineering, and so on.

These observations about modern terminology, which will be discussed below, became the basis for the development of a new platform for English-Georgian bilingual online terminological dictionaries and MDMS.

## 2 Some Tendencies in the Evolution of Structural and Semantic Characteristics of Modern Terminology

### 2.1 Migration of Common Words into Terminology

The observation of modern terminology has revealed a significant increase in the migration of words from the common vocabulary into terminology, and thus the transformation of common vocabulary words into terms. Of course, there is nothing new in this, insofar as numerous terms have been created following this exact pattern, e.g.

- (1) 'Plate' has several terminological meanings in botany and zoology (*a thin, flat organic structure or formation*), in geology (*each of the several rigid pieces of the earth's lithosphere which together make up the earth's surface*), in electrical engineering (*a thin piece of metal that acts as an electrode in a capacitor, battery, or cell*), in biology (*a shallow glass dish on which a culture of cells or microorganisms may be grown*), etc.
- (2) 'Eye' has numerous terminological meanings ranging from agricultural (*the axillary bud; the leaf-bud of a potato*), to nautical (*the extreme forward part of a ship*), geological (*a lens-shaped inclusion in a rock*), etc.

However, the novelty is the considerable intensification of this tendency. The comparison of the terminology from relatively traditional fields with that from more recent ones clearly highlights the said tendency.

A brief overview of the terms from the fields such as immunology, biotechnologies, computer, telecommunications and information technologies, and the like will suffice to reveal that this tendency is now quite conspicuous, e.g.

- (3) 'Chaperone', as a common word has two meanings: 1. *A person who accompanies and looks after another person or group of people*; 2. (dated) *A person, esp. a married or elderly woman, who, for the sake of propriety, accompanies a young unmarried lady in public, as guide and protector*.<sup>1</sup> Later, this word acquired a terminological meaning in genetics: *'protein involved in facilitating the folding or assembly of newly synthesized proteins'*.
- (4) 'Checkpoint' as a common word has the following meaning *'a barrier or manned entrance, typically at a border, where security checks are carried out on travelers'*. In genetics this word developed the following meaning *'any point in the course of a development or intracellular process at which successful completion of the previous steps in the pathway is checked before the pathway is allowed to proceed. The term is used mostly to denote such points in the eukaryotic cell cycle'*.
- (5) 'Footprint' as a common word means *'the impression left by a foot or shoe on the ground or a surface'*, in molecular biology 'footprinting' has developed the following meaning: *'any of various techniques used to determine the sites at which proteins bind to DNA or RNA, employed especially in the study of gene expression and regulation'*.
- (6) 'Canalization, canalize' as a common word means *'convey (something) through a duct or channel'*, as a term of genetics the word acquires the following terminological meaning *'the existence of developmental pathways that lead to a standard phenotype in spite of genetic or environmental disturbances'*.
- (7) 'Surprise' as a common word means *'an unexpected or astonishing event, fact, etc'*. As a military term 'surprise' is one of the principles of war, and has the following meaning: *'any military action on the enemy force when they are not expecting it'*. All other terms representing principles of war are created on the basis of the same methodology: 'objective', 'simplicity', 'mass', 'security', 'offensive', etc.

Modern computer and telecom terminology is another good example of the metaphoric transfer of meanings of common words: 'desktop', 'mouse', 'motherboard', 'scroll', 'home', 'hosting', 'wall', 'wall paper', 'page', 'to bomb', 'to hang', 'to boot', 'memory', 'server', 'jacket', etc. Such examples can be cited *ad infinitum*.

The same tendencies can be observed in the formation of analytical terms: 'jumping gene' (genetics), 'gene gun', 'gene library' (biotechnology), 'Portuguese man-of-war', 'lion's mane jellyfish' (zoology), 'memory stick', 'touch screen' (computer terms), etc.

What is the source of such intensification of the process, or what causes the use of more and more words from the general vocabulary while creating new terms? As noted above, our era is characterized by the formation of the great number of new scientific concepts as a natural result of rapid development of science and technology, which is followed by the need to create more and more new scientific terms. Under these circumstances, the language is trying to apply the principle of linguistic economy and to make the maximum use of available linguistic resources. These available resources are found, of course, in the existing common vocabulary. Consequently, in order to convey new knowledge, the language is trying to use existing words rather than create new ones.

This question is interesting in two respects. On the one hand, it is important for the development of any terminological policy by national language authorities. In order to produce equivalents for the terms created on the basis of common vocabulary, national languages have to take this circumstance into consideration, and decide how to introduce these terms into their languages – by means of mere transliteration, or by making use of the available resources of the native languages, applying the method of semantic borrowing and assigning respective terminological meanings to the same common-vocabulary words from their languages.

<sup>1</sup> Definitions of terms are quoted from respective specialized dictionaries, see References.

On the other hand, for the inclusion in the dictionary of terms created on the basis of common vocabulary, it is not enough to simply indicate an equivalent from the target language for the term from the source language. Even in bilingual dictionaries, in such cases it becomes necessary to supply the equivalent from the target language with a brief definition. It is difficult to imagine how the terms like ‘home’, ‘hosting’, ‘wall’, ‘wall paper’, ‘checkpoint’, ‘chaperone’, ‘surprise’, ‘objective’, ‘simplicity’, and so on. could be included in a dictionary without such explanations.

Brief definitions/explanations/glosses added to the Georgian equivalents of English terms have thus become an important feature of the specialized translation dictionaries composed by our team.

## 2.2 Migration of Terms into Different Domains

In the process of working on terminology, we also witnessed the growing tendency of the migration of terms from domain to domain. This phenomenon, in our opinion, is also explainable by what we have already said above. In order to cope with the influx of large amounts of terms, the language employs all available resources, including already existing terms. The migration of terms frequently occurs within a single domain, e.g. the same term may be attested in botany, zoology, anatomy or other related domains, as shown in the following example.

- (8) ‘Clone’ may mean: 1. (botany and zoology) *group of genetically identical individuals or cells derived from a single cell by repeated asexual divisions*; 2. (biotechnology) *DNA clone*; 3. *animal or plant derived from a single somatic cell or cell nucleus*, etc.

There are also many cases, when a term migrates from one domain to another, non-related domain, e.g.

- (9) ‘Tracer’ as a military term means *‘a bullet which is designed to ignite after firing and burn in flight, so that the fall of shot can be observed’*; in biology the term means *‘a substance introduced into a biological organism or other system so that its subsequent distribution may be readily followed from its color, radioactivity, or other distinctive property’*; as a technical term it’s meaning is: *‘a device which transmits a signal and so can be located when attached to a moving vehicle or other object’*.

‘Saltation’ is a term of biology, geology, software engineering, psychology.

There are countless other examples.

The migration of terms from domain to domain leads to the polysemy of terms, a phenomenon whose existence is viewed negatively by terminological standards and by the traditional approach to the semasiological characteristics of terms. In fact, our experience of working on terminological dictionaries and terms indicates that the number of cases of term polysemy is also increasing, which fact must be adequately reflected in dictionaries.

## 2.3 Analytical Terms and Acronyms

The observation of contemporary terminology has also shown the significant increase in the number of analytical, that is, multiword terms. This increase in the number of such terms has in turn led to the increase in the number of acronyms in almost every field of science and technology. There are more than half billion abbreviated terms in the IATE terminology database (Fontenelle 2014), and their number is increasing on a daily basis. A casual overview of these terms is enough to clearly see the trend:

- (10) CPU (central processing unit), SIM (subscriber identity module or subscriber identification module) / SIM card, UPS (uninterruptible power supply), USB (universal serial bus), PDF (portable

document format), GPS (global positioning system), GSM (global system for mobile [communications]), IMEI (international mobile equipment identity), CDMA (code division multiple access), HDMI (high-definition multimedia interface), HTML (Hypertext Markup Language), HSUPA (high speed uplink packet access) and HSDPA (high speed downlink packet access), UMTS (universal mobile telecommunications system), CMOS (complementary metal-oxide semiconductor), WiMAX (worldwide interoperability for microwave access), and so on.

In our opinion, this tendency is important from the point of view of the development of the terminological policy in national languages. What we mean is that the analytical terms, on the one hand, are not succinct and economic but, on the other hand, such terms are transparent and easily understandable. While introducing them into national languages, it is crucial to retain, as far as possible, this positive aspect of multiword terms. Unfortunately in the Georgian language (and possibly in other languages as well) the analytical terms are often transliterated. For instance, the psychology term ‘residual stress pattern’ is rendered in an online *Dictionary of Social and Political Terms* as ‘rezidualuri stresis paterni’ (რეზიდუალური სტრესის პატერნი). The formation of such Georgian terms has become a rule to the detriment of the effective application of the method of structural borrowing. The growing number of such terms in our languages cannot, in our opinion, promote the development of any field of science and, on the contrary, can become an obstacle thereto.

In our dictionaries we include analytical terms as separate dictionary entries, always supplying them with acronyms, if they have any. Acronyms are also included as entry words in a dictionary and are cross-referenced to their respective full forms.

## 2.4 Definitions

As noted above, in our bilingual terminological dictionary entries we supply Georgian equivalents of English terms with brief explanatory definitions. The need to add definitions arises from a number of reasons. First of all, such addition is necessitated by the polysemy of terms, where an explanatory definition is needed for sense disambiguation purposes. Another reason for the inclusion of definitions is great number of new terms. Such definitions facilitate their correct use and their rapid establishment in this or that field of knowledge. The addition of definitions is also necessary when a term is transliterated into the target language. For instance, without providing definitions of terms, the informative value of the dictionary entries cited below would remain very low:

- (11) ‘chemoattractant’ – kemoatraktanti (ქემოატრაქტანტი);  
 ‘chemoreceptor’ – kemoretseptori (ქემორეცეპტორი);  
 ‘chemorepellent’ – kemorepelenti (ქემორეპელენტი);  
 ‘chemosensory’ – kemosensoruli (ქემოსენსორული)

Since a bilingual dictionary is not an explanatory one, our definitions are not comprehensive. Our dictionaries are not intended for narrowly specialized experts of particular fields of knowledge. Instead, they are intended for the wide spectrum of the public, including specialists in various fields, students, individuals generally interested in these fields, and so on. So we do not try to give very detailed descriptions of terms. Our approach to the definitions of terms is adequately described in the following quote from Pius ten Hacken:

For many items that belong to specialized vocabulary there is no need to delimit the concept precisely. The best approach is to treat them in the same way as a lexicographer describes a word. Such lexicographic definitions are fully adequate as long as there is no legal or scientific controversy about the concept. (ten Hacken 2010 : 925).

Very precise definitions of terms are also necessary in the cases when a very specific dictionary is being composed for specific objectives and a specific project, as described in “Experts and Terminologists:



Exchanging Roles in the Elaboration of the Terminological Dictionary of the Brenner Base Tunnel (BBT)” (Chiocchetti & Ralli 2014).

## 2.5 Terminological Variation

As we know, the terms which are monosemous (one meaning per term), with one term corresponding to one specific concept, were traditionally regarded as ideal ones. Synonymy, according to the traditional view, was not regarded as a desirable characteristic for a term either. We already addressed the issue of polysemy above; as for synonymy it also constitutes an important feature of contemporary terminology.

Although specialized language initially aspired to having one linguistic designation for each concept for greater precision, it is true that the same concept can often have many different types of linguistic designations. In the same way as in general language, there is terminological variation based on user-based parameters of geographic, temporal or social variation or usage-based parameters. (León-Araúz & Reimerink 2014 : 658).

The synonymy of terms, as one of their characteristic features, is already reflected in terminological standards.

There might be more than one designation for the same concept, i.e. there might be synonyms. (ISO 2009:704 : 7.2.4). Also term variants, e.g. abbreviated forms like clippings and acronyms, are common in specialised domains. (ISO 2009 :704 : B.2.4).

Our experience has also demonstrated that terms often have synonyms. These may belong to different stylistic registers, may represent an acronym of a term, or a term expressed by means of a symbol. The acceptance of synonymy as one of the characteristic features of terms is also important for the development of terminological policy by national language authorities. In particular, for the terms introduced in a language through transliteration, there may be created synonyms based on the resources from the native tongue / target language in order to ensure the coexistence of both terms in the language vocabulary as synonyms. We made active use of this methodology while working on our dictionaries, creating, in close collaboration with domain experts, Georgian synonyms for the international terms already established in our language. An interesting paper on this subject was presented by our colleague Enn Veldi at the XVI EURALEX International Congress in Bolzano/Bozen, Italy. The title of the paper was “Concerning the Treatment of Co-existent Synonyms in Estonian Monolingual and Bilingual Dictionaries” (Veldi, 2014).

## 2.6 Labels

In their writings, some terminologists note that while describing terms it is necessary to indicate the register of the application of the words in question.

Very much like a traditional dictionary which makes use of usage notes and a variety of labels aimed at capturing levels of formality (formal, informal, slang, taboo...), a terminological database such as IATE makes extensive use of metalinguistic labels. (Fontenelle 2014 : 35).

Our experience has also shown us that it is necessary to introduce stylistic labels. Terminological vocabulary also often shows some very high degree of creativity in the process of the production of colloquial or slang, even jocular varieties of terms. This can be seen with computer slang words such as: barebone, bare metal, bloatware, blue screen of death, crippleware, and the like; and military slang words such as: basket case, beetle-crusher, broolly hop, bull, battle bowler, and so on.

In specialized dictionaries it is also necessary to supply terms with subject-field labels in order to identify the exact field or fields of application of the term in question. Terms can be not only nouns,



but also adjectives or verbs; consequently, the dictionary entry must also be supplied with part-of-speech labels.

## 2.7 Related Words

In addition to providing definitions, in order to better highlight the meanings of terms it is important to describe them in an interrelated way. In our dictionaries we do not aim to describe the whole network of terms, but usually cross-reference them to other closely related terms. In our experience, the inclusion of related words in a dictionary entry is an important component of the description of term, e.g.

- (12) ‘Air defence’ (a military term) is cross-referenced to ‘active air defence’ and ‘passive air defence’; ‘A 1 Echelon’ (a military term) is sent for additional information to ‘A Echelon’ and ‘A 2 Echelon’; ‘Abrasive blasting’ (a technical term) is sent to ‘bead blasting’ and ‘sand blasting’; ‘Analog-to-digital converter’ (a computer and telecom term) is cross-referenced to ‘analog signal’, ‘digital signal’ and ‘digital-to-analog converter’; ‘Bitmap graphics’ (a computer term) to ‘bitmap’, ‘raster graphics’ and ‘vector graphics’; ‘Call tracing’ (a telecom term) is cross-referenced to ‘on-demand call tracing’ and ‘permanent call tracing’, etc.

As can be seen from the examples above, related words to which the main headword is cross referenced may reflect hyper-hyponymic relations between terms, or they may be co-hyponyms providing additional information about the concept, or a term may be cross referenced to antonymous or otherwise related terms which help to understand the term in question.

## 3 MDMS and a New Platform for Georgian Online Bilingual Terminological Dictionaries

### 3.1 Structure and Fields of MDMS

MDMS was specially developed for the *English-Russian-Georgian Technical Online Dictionary*, the first ‘digitally born’ online multilingual specialized dictionary of Georgian, created in MDMS. The development of this system was possible thanks to the grant provided by the Shota Rustaveli National Science Foundation of Georgia. Other dictionaries, the *English-Georgian Military Online Dictionary* and *English-Georgian Biology Online Dictionary* will be ported to this system in the near future. Planning of the MDMS was based on our views on the type and amount of information that should be present in a bilingual specialized dictionary entry, as discussed above. In MDMS, the items of dictionary information are divided into separate fields, which enables their efficient use, management and display. The basic fields of the MDMS are:

1. *Field of Headwords* which in its turn is subdivided into 1) *English Headwords*, 2) *Georgian Headwords* and 3) *Russian Headwords*. The Georgian headwords section contains a subfield of *Definitions in Georgian*;
2. *Field of Labels*, subdivided into the following components: 1) *POS Labels*, 2) *Subject Field Labels*, 3) *Stylistic Labels*;
3. *Field of Other Forms*, including the following subfields: *Plural*, *Abbreviation*, *Symbol*, *Full Form*;
4. *Field of Examples* comprises *Collocations* and example *Phrases* and *Sentences*;
5. *Field of Synonyms*;
6. *Field of Similar Words*;
7. *Field of Related Words* (see Figure 1).

In cases of polysemy, separate sections are added for each meaning of a polysemous term with the same fields, as follows: Georgian headwords, Russian headwords, Georgian definitions, POS, subject field and stylistic labels, synonyms, related words, and so on.

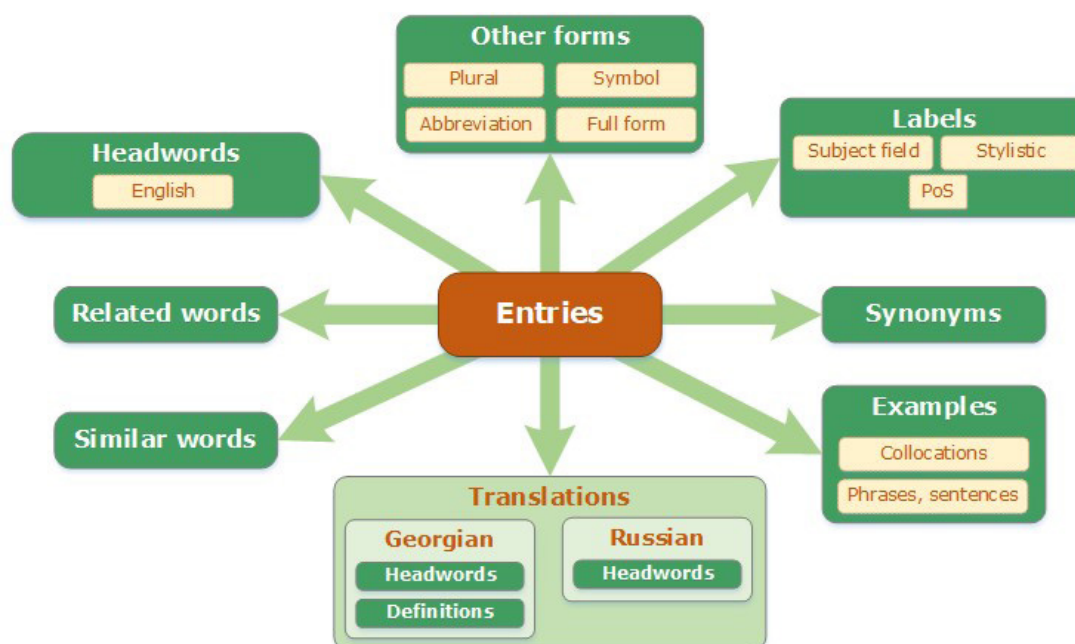


Figure 1: Fields of MDMS

### 3.2 New Platform for Georgian Online Bilingual Terminological Dictionaries

Dictionary entries are composed by placing information components in appropriate fields, thus establishing connections between them. Such an approach almost rules out the possibility of duplication, facilitates further editing and enables the generation of additional connections or backlinks.

For the online dictionary platform, raw working material from MDMS is converted into data in the appropriate format, adapted to the functionalities of the online dictionary (see Figures 2, 3). This also includes the automatic generation of additional language pairs, which are not actually present in the initial working data but are based on the existing connections. Thereby, the working data of the English-Russian-Georgian Technical Online Dictionary, which initially is English-Georgian and English-Russian, is transformed for the online dictionary into additional Georgian-English, Georgian-Russian, Russian-Georgian and Russian-English terminological sets, six altogether (see Figures 4, 5).

The *English-Russian-Georgian Technical Online Dictionary* contains: 21,232 English headwords, 24,036 Georgian headwords and 11,591 Russian headwords. As for terminological pairs, it consists of: 18,670 English-Georgian and 8,516 English-Russian term pairs. The dictionary also contains automatically generated 24,036 Georgian-English, 12,966 Georgian-Russian, 11,591 Russian-English and 11,513 Russian-Georgian term pairs, give a total of 87,292 term pairs.

On the opening of the hyperlink to a dictionary entry, on the left area of the computer screen there are displayed the contents of the fields of MDMS discussed above for each term, these are: similar and related words, synonyms, where necessary abbreviations, symbols, etc. Also shown are nearby entries, as well as the compound terms, which include the given word in their composition (see Figure 2). It should be noted that in case of Georgian-English or Georgian-Russian terminological pairs such information appears on the computer screen in Georgian (see Figure 4), while in the case of Russian-English or Russian-Georgian ones it appears in Russian.

- all fields -

capture

Q

აბგ

# ა ბ ვ გ დ ე ე ჯ ზ ი ი კ ლ მ ნ ო პ რ ს ტ უ ფ ხ ც ჭ შ წ ტ ყ რ

## Related words

- print screen
- screenshot

## Synonyms

- data capture
- screen capture
- video capture

## Nearby entries

- capsule
- CAPTCHA
- captive
- captive balloon
- captive nut
- capture I
- capture II
- capture efficiency
- car
- car2car communication
- carabiner

## You may also be interested in

- automatic identification and data capture / collection
- capture II
- capture efficiency
- data capture
- screen capture
- video capture

## capture I

noun /'kæptʃə(r)/

in Georgian | in Russian

- ფიზ. წატაცება (ატომისა, მოლეკულისა, იონისა ან სხვა ნაწილაკისა);
- კომპ. გამოსახულების დაფიქსირება, გამოსახულების გადაღება (ეკრანზე წარმოდგენილი გამოსახულების დაფიქსირება და გრაფიკულ ფაილად შენახვა; აგრ. screen capture) [იხ. აგრ. screenshot, print screen];
- კომპ. ვიდეოგამოსახულების გადაღება (ეკრანზე წარმოდგენილი გამოსახულებისა და შესრულებული მოქმედებების გადაღება და ვიდეოფაილად შენახვა; აგრ. video capture);
- კომპ. მონაცემების დაფიქსირება, მონაცემების შეტანა (მონაცემების / ინფორმაციის მიღება / დაფიქსირება ამა თუ იმ სახის ინფორმაციული ნაკადიდან, საკომუნიკაციო არხიდან, გადამღები ან წამკითხავი მოწყობილობიდან და ა.შ.; აგრ. data capture).

Figure 2: Entry of a polysemous term.

- all fields -

XML

Q

აბგ

# ა ბ ვ გ დ ე ე ჯ ზ ი ი კ ლ მ ნ ო პ რ ს ტ უ ფ ხ ც ჭ შ წ ტ ყ რ

## Related words

- markup language
- Full form
- extensible markup language

## Synonyms

No synonyms were found

## XML

noun /ˈɛksɛmˈɛl/

in Georgian | in Russian

- კომპ. (extensible markup language-ის აბრევ.) მონიშვნის / მარკირების გაფართოებადი ენა, XML-ენა (პროგრამირებაში - მონიშვნის / მარკირების ენა, რომელიც ნებისმიერი სიმბოლოს / ტეგის გამოყენების საშუალებას იძლევა) [იხ. აგრ. markup language].

Figure 3: Entry of an acronym.

TECH DICTIONARY  
DB version: 2017-05-22 01:50:44

ქართული  
About Dictionary | User Guide | Contact Us

- all fields - მონიშვნის გაფართოებადი ენა

აბვ # ა ბ ვ გ დ ე ე ჯ ზ ი კ ლ მ ნ ო პ რ ს ტ უ ფ ხ ც ჭ შ წ ტ ყ რ

**Related words**  
No related words were found

**Synonyms**  
■ XML-ენა  
■ მარკირების გაფართოებადი ენა

**Nearby entries**  
■ მონელ-ლითონი  
■ მონიკლევა  
■ მონიტორი  
■ მონიშვნა  
■ მონიშვნის გაუქმება  
■ მონიშვნის გაფართოებადი ენა  
■ მონიშვნის ენა  
■ მონიშნული  
■ მონობლოკი  
■ მონობლოკური  
■ მონობლოკური დისკი

**მონიშვნის გაფართოებადი ენა**  
noun  
[in English](#) | [in Russian](#)

კომპ. расширяемый язык разметки

Figure 4: Automatically generated Georgian-Russian term pair.

TECH DICTIONARY  
DB version: 2017-05-22 01:50:44

ქართული  
About Dictionary | User Guide | Contact Us

- all fields - расширяемый язык разметки.

აბв # ა ბ ვ გ დ ე ე ჯ ზ ი კ ლ მ ნ ო პ რ ს ტ უ ფ ხ ც ჭ შ წ ტ ყ რ

**Related words**  
No related words were found

**Synonyms**  
No synonyms were found

**Nearby entries**  
■ расширительная плата  
■ расширительное гнездо  
■ расширительный бак

**расширяемый язык разметки**  
noun  
[in English](#) | [in Georgian](#)

კომპ. მონიშვნის გაფართოებადი ენა, მარკირების გაფართოებადი ენა, XML-ენა

Figure 5: Automatically generated Russian-Georgian term pair.

## 4 Conclusion

As noted at the beginning of the present paper, the rapid development of science and technology and the generation of numerous terms in all fields of knowledge has brought about some changes to the



structural and semantic characteristics of modern terms. Languages have begun to use all of their available resources in order to denominate new concepts, including the reuse of existing terms. There is thus an increase in the number of cases when the words from general vocabulary migrate into terminology, as well as an increase in the transfer of the terms from one field to another, both related and unrelated fields of knowledge. Examples of polysemy have increased, as well as the number of acronyms, triggered by the growth in the number of analytical, multiword terms. The emergence of colloquial, slang and jocular terms has resulted in a need to supply entries with relevant stylistic labels and so on. As such, the “translations, definitions, acronyms, subject field labels, usage notes and examples [in terminological dictionaries] are similar to what can be found in monolingual or bilingual dictionaries” (Fontenelle, 2014 : 43). It is because of these very changes, in our opinion, that Hennie van der Vliet opines: “... terminology management, although a very practical undertaking, may gain great profit from theoretical findings in lexical semantics” (van der Vliet, 2006).

These changes in the field of terminology have also caused the exchange of roles between domain experts and terminologists in the process of working on a terminological dictionary. In 2014, while presenting their dictionary in Bolzano, Elena Chiocchetti and Natascia Ralli described how their roles had changed while working on their joint project. While previously the leading roles in the process of working on such dictionaries were performed by terminologists with domain experts assisting them, their joint project saw the main work performed by domain experts assisted by terminologists. In an interesting article titled “In Quest of a Profile: Portrait of a Terminologist as a Young Sublanguage Expert”, Willy Martin argues that specialized lexicography needs “domain expertise, linguistic expertise and information management expertise in order to function properly” (Martin, 2006 : 83), and concludes that “The (ideal) terminologist as an individual does not exist. The (ideal) terminologist is a team” (Martin, 2006 : 92).

In fact, who was it who worked on bilingual terminological dictionaries two decades ago? Perhaps a terminologist who was looking for equivalents from his/her native language for source language terms and included them in the dictionary. In this activity the terminologist was assisted by and consulted domain experts. But today terminology has undergone such big changes, and the amount of information to be included in every entry of terminological dictionaries is so considerable, that terminologists have to make a huge number of decisions with respect to each particular term (see, for example, Adamska-Sałaciak, 2016), along with requiring extensive text corpora in order to identify and extract collocations of terms (e.g. see Taljard, 2016), and thus successfully performing such work is now unimaginable without the involvement of a team and, most important of all, domain experts.

These changes in modern terminology have determined the inclusion of more information in a bilingual terminological entry. The development of MDMS and the new platform for bilingual online specialized dictionaries for Georgian, as outlined in this paper, was based on these new demands with regard description of modern terminology.

## References

- Adamska-Sałaciak, A. (2016). On bullying, mobbing (and harassment) in English and Polish: Foreign-language-based Lexical Innovation in a Bilingual Dictionary. In Margalitadze, Tinatin & Meladze, George (eds.). *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 758-766.
- Chiocchetti, E., Ralli, N. (2014). Experts and Terminologists: Exchanging Roles in the Elaboration of the Terminological Dictionary of the Brenner Base Tunnel (BBT). In *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Andrea Abel, Chiara Vettori & Natascia Ralli (eds.). Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp.609 – 620.



- Fontenelle, Th. (2014). From Lexicography to Terminology: a Cline, not a Dichotomy. *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Andrea Abel, Chiara Vettori & Natascia Ralli (eds.). Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp.25 – 45.
- Ghambashidze, R. (1986). *Georgian Scientific Terminology*. Tbilisi : Metsniereba (in Georgian).
- ISO 704 : 2000. Terminology Work – Principles and Methods.
- ISO 704 : 2009. Terminology Work – Principles and Methods.
- León-Araúz, P., Reimerink, A. (2014). From Term Dynamics to Concept Dynamics: Term Variation and Multidimensionality in the Psychiatric Domain. In Andrea Abel, Chiara Vettori & Natascia Ralli (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 657 – 667.
- Taljad, E. (2016). Collocational Information for Terminological Purposes. In Margalitadze, Tinatin & Meladze, George (eds.). *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 553 – 560.
- ten Hacken, P. (2010). The Tension between Definition and Reality in Terminology. In Anne Dykstra & Tanneke Schoonheim (eds.). *Proceedings of the XIV EURALEX International Congress*. 6-10 July 2010, Leeuwarden/Ljouwert: Fryske Akademy, pp. 915 – 927.
- van der Vliet, H. (2006). Combinatorics for special purposes. In Pius ten Hacken (ed.), *Terminology, Computing and Translation*. Narr Francke Attempto verlag GmbH + Co. KG, Tübingen.
- Veldi, E. (2014). Concerning the Treatment of Co-existent Synonyms in Estonian Monolingual and Bilingual Dictionaries. In Andrea Abel, Chiara Vettori & Natascia Ralli (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 829 – 835.
- Martin, W. (2006). In Quest of a Profile: Portrait of a Terminologist as a Young Sublanguage Expert. In Pius ten Hacken (ed.), *Terminology, Computing and Translation*. Narr Francke Attempto verlag GmbH + Co. KG, Tübingen, pp. 73 – 94.

## Dictionaries

- Bowyer, R. (2004). *Dictionary of Military Terms*. Macmillan, Bloomsbury 2004.
- Comprehensive English-Georgian Online Dictionary* (2010). T. Margalitadze, G. Meladze, G. Khundadze et al. Lexicographic Centre, Tbilisi State University. Accessed at: <http://dict.ge>
- Dictionary of Social and Political Terms*. Accessed at: <http://www.dictionary.css.ge/content/residual-stress-pattern> [28.03.2018]
- Dictionary of Zoology* (2009). Michael Allaby (ed). Oxford University Press.
- English-Georgian Biology Online Dictionary* (2014). T. Margalitadze, N. Porakishvili, G. Meladze, G. Khundadze et al. Lexicographic Centre, Tbilisi State University. Accessed at: <http://bio.dict.ge>
- English-Georgian Military Online Dictionary* (2009). T. Margalitadze, G. Meladze, G. Khundadze et al. Lexicographic Centre, Tbilisi State University. Accessed at: <http://mil.dict.ge>
- English-Russian-Georgian Technical Online Dictionary* (2016). T. Margalitadze, G. Meladze, G. Khundadze et al. Lexicographic Centre, Tbilisi State University. Accessed at: <http://techdict.ge>
- Henderson's Dictionary of Biology* (2008). Eleonor Lawrence (ed). Pearson, Edinburgh.
- King, R.C., Stansfield, W.D. & Mulligan, P.K. (2006). *Dictionary of Genetics*. Oxford University Press.

# Building a Portuguese Oenological Dictionary: from Corpus to Terminology via Co-occurrence Networks

**William Martinez<sup>1</sup>, Sílvia Barbosa<sup>2</sup>**

<sup>1</sup>Université Sorbonne Nouvelle Paris 3, <sup>2</sup>NOVA FCSH - CLUNL

E-mail: [wmartinez68@gmail.com](mailto:wmartinez68@gmail.com), [silviabarbosa@fcs.unl.pt](mailto:silviabarbosa@fcs.unl.pt)

## Abstract

This paper focuses on the elaboration of a dictionary of terms in the Portuguese language which describe the wine-tasting experience. We present a corpus-based analysis aimed at designing an electronic dictionary: on the basis of a compilation of approximately 21,000 wine descriptions downloaded from a dozen Portuguese websites, we estimated both by frequency analysis and lexicographical study which terms were recurrent, relevant and representative of the “hard to put into words” occupation that is oenology. From the results thus obtained, a list was made of words that describe the sensory analysis in its three main aspects: visual, olfactive and gustatory. An exhaustive co-occurrence analysis then identified those terms which contribute most to structuring the text by way of their tendency to attract other words against statistical odds. When displayed in a co-occurrence network, these anchors emerge from the mesh as the foundational lexicon for wine tasting, and can be evaluated as prime candidates for a distributional thesaurus.

**Keywords:** collocations, co-occurrences, word network, corpus linguistics, oenology, terminology

## 1 Objectives

Extracting relevant information from text in order to establish lexicographical lists of domain-specific terms, or as Kilgariff (2005) says ‘putting the corpus into the dictionary’ is a complex operation. In the field of Information Science, the hierarchy known as the DIKW pyramid (from data to information to knowledge to wisdom) describes the processing chain that transforms dispersed raw facts into organized synthesized information. In this process, the most likely transition where value can be lost lies between data and information, because this transition is essentially achieved by inference: interrogative questions like *what?*, *when?* or *how?* are answered by data invested locally with meaning for a purpose. Typically, this data is selected in context, and relevant keywords are extracted by combining two things: a large volume of domain-related text and one or more domain experts, thus aiming for the highest representativity and the greatest degree of precision.

We believe that while higher levels of the DIKW pyramid require human cognition and judgment to reach understanding, the initial stage of information gathering can be greatly improved by a statistical approach to text. Indeed, an exhaustive analysis of all operating contexts can provide exact statistics regarding repeated word coincidence, which in turn can help describe word usage. According to the principle set out by J. R. Firth (1957) – “You shall know a word by the company it keeps” – co-occurrence analysis provides a contextually validated description of operative vocabulary for a given domain.

At the heart of the distributional hypothesis (Harris 1968) is the belief that the small number of realized combinations between words in context compared to the huge theoretical combinatorics that are possible between these elements must be indicative of strong linguistic relations at work. Whether

semantically or syntactically, words are bound and operate in a systematic manner that can be measured by their organized coincidences in context.<sup>1</sup>

## 2 Building a corpus

Cellars provide a time-tested environment for the preservation of wines. Temperature, humidity and light are carefully monitored and maintained for the optimal long-term aging. However, when it comes to harvesting wine vocabulary it is best to aim for quantity, diversity and topicality. This all-around representativity is achieved by collecting domain-specific information, where it is massively and archived in huge amounts: on the internet.

The corpus consisted of texts referred to as the *notas de prova* (wine tasting note) – a succinct text – presented in specific sections of newspapers or on producer’s web pages. Those notes prototypically represent the descriptions of wines made by experts to describe the organoleptic sensations of each wine (color, aroma, type of fruit, degree of complexity, texture, final persistence, etc.) and evaluate the drink, using the descriptors available in the related repertoire, like a guide for those who read it.

By selecting a set of major websites relative to the Portuguese wine industry (with 1,480 notes of Portuguese producers and 20,011 notes collected from Portuguese online specialized wine magazines from several producers/wine companies) we were able to compile a set of over 21,000 wine tasting notes with 589,498 word tokens for 7,652 word types, among which there were 2,815 hapax legomena (Table 1).

Table 1: Corpus characteristics.

Tokens	589 498
Types	7 652
Hapax	2 815
Maximal frequency	40 548 (most frequent type: e)

Upon browsing the word-type dictionary extracted from the corpus (Table 2), it is interesting to note that there are very few tool-words among the most frequent types appearing in the text. The originality of layout in the frequency dictionary may find its explanation in the particular format of the text. Indeed, tasting notes are most often short and concise reports containing just one to three sentences.<sup>2</sup> Such a short text requires fewer tool words, which explains their absence at the top of the dictionary: connectors such as prepositions and conjunctions, anaphoric pronouns, etc.

Table 2: Most frequent word-types (frequency  $\geq 5000$  occurrences).

Rank	Type	Freq.	Rank	Type	Freq.
1	<i>e</i>	40 548	11	<i>bem</i>	7 377
2	<i>de</i>	24 747	12	<i>final</i>	7 013
3	<i>com</i>	23 712	13	<i>acidez</i>	7 006
4	<i>muito</i>	14 318	14	<i>fruta</i>	6 726
5	<i>boca</i>	11 815	15	<i>fruto</i>	6 494
6	<i>na</i>	11 091	16	<i>no</i>	5 429
7	<i>a</i>	10 887	17	<i>taninos</i>	5 401
8	<i>um</i>	10 315	18	<i>boa</i>	5 327
9	<i>aroma</i>	7 993	19	<i>mas</i>	5 315
10	<i>notas</i>	7 816			

<sup>1</sup> The possible relationship between distributional and semantic similarities has been exploited for the generation of automatic thesauri in previous works, notably Lin 1998 and Curran & Moens 2002.

<sup>2</sup> Based on three declared signs of punctuation (.,?), 39 939 sentences are identified in the corpus with an average length of 14.76 words for a standard deviation of 5.10.

### 3 Type frequency vs type valency

Word frequency is commonly used as a topic indicator because it is an obvious and efficient measure of themes developed in a corpus: among the most frequent word-types of a text are to be found the recurrent nouns and verbs<sup>3</sup> (and hopefully the subjects and predicates of the topic being discussed). However, as bricks are mixed with mortar, hierarchization by frequency mixes content words with tool words among the most frequently occurring types. Indeed, a text follows an inevitable organization of words according to their frequency. Word frequencies are conditioned by Zipf's Law (Zipf 1965) whereby a small number of words have a very high frequency and a large number of words have a very low frequency.

From this predictable structure, Luhn (1958) – whilst working on automatic summarization – derived a method for locating valuable information-loaded words in between too frequent and too rare elements in the dictionary. Contrary to the structure proposed by Luhn, as described by Deghani (2016), the very frequent word types (function words) and the very rare words (hapax) contain few if any information-loaded words. As consequence, automatic cut-off limits can usually be defined to isolate significant words based only on their frequency. Clearly, in the case of our corpus this filter cannot be automated. Moreover, many experiments prove that dictionary hierarchy is no measure of significance in context.<sup>4</sup>

Another strong indicator of lexical behavior in context is lexical valency<sup>5</sup> i.e. the propensity for a given keyword to attract other words in context. Typically, word valency is measured according to the number of co-occurrences<sup>6</sup> which are detected around a given keyword in a defined context. For example, in our corpus, around the 11,815 occurrences of the word *boca* (mouth) and within the limits of each of the 11,688 sentences (222 540 tokens or 37% of the corpus) where the word occurs, we tallied every occurrence of every other word appearing in these contexts.

Beyond the mere co-frequencies of these words alongside the keyword *boca*, the results we obtained can be normalized to take into account the volume of the corpus, the volume of the sample (all phrasal contexts of *boca*) and the global frequencies of the co-occurring words. These four parameters (noted  $T, t, F, f$  for global and local text volumes and frequencies) are the input data for a great number of statistical models. Our choice of method for calculating a probabilistic score is the Hypergeometric Model<sup>7</sup>, because of its easy-to-read result which measures the degree of surprise when confronting

3 Another assumption is that these inevitable words are evenly distributed in the corpus even if overall high frequency is sometimes due to particularly loaded sections of a corpus.

4 More recently the suitability of word frequency as a criterion for vocabulary selection has been questioned in language teaching by Okamoto (2015).

5 Valency is a notion borrowed from chemistry, where it denominates the combining power or affinity of an element, especially as measured by the number of hydrogen atoms it can displace or combine with, all depending on the electrons present in the outermost shell.

6 Unlike the term 'collocation', which implies a number of two adjacent collocates, 'co-occurrence' alludes to attractions between words in a broad sense, without imposing constraints of contiguity, orientation or distance. As a consequence, the phenomena detected are numerous and varied, thus reflecting the richness of lexical activity in the corpus.

7 The Hypergeometric Model determines the probability for an observed word frequency ( $x$  occurrences of word  $w$  in the vicinity of keyword  $k$ ) based on four parameters:

$T$ : number of tokens in the corpus

$t$ : number of tokens in keyword contexts

$F$ : frequency of co-occurring word in corpus

$f$ : frequency of co-occurring word in keyword contexts

$$P[X = f] = \frac{\binom{F}{f} \binom{T-F}{t-f}}{\binom{T}{t}}$$

A numerical specificity summarizes the deviation between the theoretical value and observed value, which can be null, positive or negative. If the observed frequency is more or less what is expected in theory, then there is no specificity. If the observed frequency is higher than expected, then the specificity is positive. Inversely, if the observed frequency is lower than expected, then the specificity is negative. The value indicates the degree of probability of the occurrence, for example: +3 indicates a positive specificity (more occurrences than expected) and a likelihood of 1/1000 (3 zeroes). A negative specificity of -10 would indicate a negative co-occurrence between words (less coincidences in context than expected) with a probability of 1/10000000000 (10 zeroes).

the observed co-frequency with the expected co-frequency of a given word. From a pragmatic point a view, this model, albeit complex, yields a result that is very easy to interpret (also beyond the binary co-occurrence / no co-occurrence it sometimes indicates a negative result, which signals anti-co-occurrence or repulsion between words).

Our list of co-occurrences is then re-ordered according to this probabilistic measure. We counted a total of 4,538 statistically specific co-occurrences, of which 2,259 were positive (words over-represented around *boca*) and 2,279 negative (under-represented). In Table 3 an excerpt from our results shows that the strongest co-occurrences reach very high degrees of statistical unlikelihood. For example, the number of encounters between the keyword ‘mouth’ and co-occurring terms ‘acidity’, ‘tannins’, ‘body’ and ‘soft’ are given a specificity of +100. This means that the odds of these coincidences are 1 over 1 plus 100 zeroes ( $1^{-100}$ ), so very unlikely to occur in context and thus worthy of our interest.

Table 3: Main positive and negative co-occurrences around *boca* (mouth).

Rank	Type	Positive		Rank	Type	Negative
1	<i>acidez</i> (acidity)	+100		1	<i>fruta</i> (fruit)	-100
2	<i>final</i> (end)	+100		2	<i>aroma</i>	-100
3	<i>taninos</i> (tannins)	+100		3	<i>madura</i> (mature)	-100
4	<i>corpo</i> (body)	+100		4	<i>frutos</i> (fruits)	-100
5	<i>bom</i> (good)	+100		5	<i>especiarias</i> (spices)	-100
6	<i>macio</i> (soft)	+100		6	<i>cor</i> (color)	-100
7	<i>volume</i>	+100		7	<i>minerais</i> (minerals)	-100
8	<i>redondo</i> (round)	+82		8	<i>flores</i> (flowers)	-100
9	<i>frescura</i> (freshness)	+68		9	<i>preta</i> (black)	-100
10	<i>doçura</i> (sweetness)	+63		10	<i>vermelha</i> (red)	-87
11	<i>estrutura</i> (structure)	+62		11	<i>floral</i>	-86
12	<i>sabor</i> (taste, flavor)	+60		12	<i>citrinos</i> (citruses)	-85
13	<i>textura</i> (texture)	+52		13	<i>vegetais</i> (vegetables)	-84
14	<i>secura</i> (dryness)	+48		14	<i>nariz</i> (nose)	-73
15	<i>viva</i> (bright)	+47		15	<i>barrica</i> (barrel)	-72
16	<i>longo</i> (long)	+46		16	<i>tostados</i> (toasted)	-70
17	<i>mediano</i> (average)	+46		17	<i>aromática</i> (aromatic)	-65
18	<i>equilíbrio</i> (balance)	+40		18	<i>silvestres</i> (wild, sylvan)	-64
19	<i>sabroso</i> (tasty)	+35		19	<i>folhas</i> (leaves)	-60
20	<i>cheia</i> (full)	+32		20	<i>chocolate</i>	-57

With these results we are able to build an understanding of the buccal experience in wine tasting: acidity, tannins, body, soft, round, freshness and sweetness are part of the tasting experience. Inversely, other words are given negative specificities to indicate their absence in the contexts of ‘mouth’: fruit, mature, spices, flowers, citruses and vegetables or chocolate for example do not come to the mind of wine tasters when describing their ‘in mouth’ appreciation.

Regarding these probability scores, it should be noted that they reward high degrees of coincidence in context, regardless of individual word frequency. This means that low-frequency terms can be promoted to the top of our statistical ranking dictionary to provide an order of importance quite different to that in our initial frequency dictionary. The results in Table 4 show how words ranked according to their valency reveal an unexpected order. From a linguistic perspective, keywords with a high valency can be interpreted as elements of disruption. Indeed, as statistical measures point out, whenever a high-valency word appears in context it seems to trigger the appearance of one or several other words which, according to the laws of probability, were not expected to arrive at this moment in the flow



of text. Therefore, we can consider high-valency words to play a particular role in context as they contribute strongly to structuring the entire text.

Table 4: Highest valency types.

Rank	Type	Valency	Rank	Type	Valency
1	<i>fruto</i>	49	11	<i>longo</i>	30
2	<i>notas</i>	41	12	<i>corpo</i>	30
3	<i>frutos</i>	39	13	<i>acidez</i>	29
4	<i>cor</i>	38	14	<i>final</i>	28
5	<i>leve</i>	37	15	<i>fruta</i>	27
6	<i>taninos</i>	36	16	<i>médio</i>	25
7	<i>especiarias</i>	35	17	<i>boca</i>	24
8	<i>preta</i>	33	18	<i>vermelha</i>	24
9	<i>citrinos</i>	33	19	<i>flores</i>	23
10	<i>aroma</i>	31	20	<i>encorpado</i>	23

To fully appreciate this lexical dynamic, we extend co-occurrence to all word-types<sup>8</sup> with a frequency  $\geq 10$ , which yields a table of 2,170 x 2,170 words (excerpt in Table 5) identifying all co-occurring types. This huge table requires some form of filtering to reduce its size to the essential statistical data; for example: words retained after analysis should co-occur at least  $x$  times at no greater distance than  $y$  words. A more stringent criterion would be mutual co-occurrence, whereby  $A$  is in co-occurrence with  $B$  if and only if  $B$  is in co-occurrence with  $A$  (indeed depending on the method for measuring, a co-occurrence relation may not be reciprocal<sup>9</sup>).

Table 5: Table of co-occurrences (excerpt).

Keyword	Co-oc. 1	Co-oc. 2	Co-oc. 3	Co-oc. 4	Co-oc. 5
boca	<i>prova</i>	<i>final</i>	<i>na</i>	<i>acidez</i>	<i>fácil</i>
na	<i>final</i>	<i>boca</i>	<i>acidez</i>	<i>firme</i>	<i>bom</i>
acidez	<i>final</i>	<i>na</i>	<i>boca</i>	<i>viva</i>	<i>cremoso</i>
final	<i>na</i>	<i>boca</i>	<i>acidez</i>	<i>longo</i>	<i>boa</i>
taninos	<i>final</i>	<i>na</i>	<i>boca</i>	<i>firme</i>	<i>cheio</i>
corpo	<i>prova</i>	<i>final</i>	<i>boca</i>	<i>acidez</i>	<i>viva</i>
prova	<i>boca</i>	<i>uma</i>	<i>fácil</i>	<i>corpo</i>	<i>ampla</i>
bom	<i>na</i>	<i>acidez</i>	<i>conjunto</i>	<i>nervo</i>	<i>corpo</i>
macio	<i>prova</i>	<i>e</i>	<i>acidez</i>	<i>fácil</i>	<i>corpo</i>
volume	<i>bom</i>				
tanino	<i>algum</i>	<i>maduro</i>	<i>alguma</i>	<i>meio</i>	<i>secura</i>

The description provided on the Table 5 is total: all statistically remarkable encounters in context are recorded. Because it describes all relations of attraction between words in the corpus, this ‘adjacency matrix’ is akin to a lexical mesh that supports the entire corpus. Every single binary co-occurrence contributes to building a complex network and forming, link by link, the lexical backbone of the text. However, once this information is ordered in tabular form, the problem is to interpret it, or at least read it. This is why, given their density, such matrices are usually visualized in the form of a co-occurrence network.

<sup>8</sup> Textual statistics requires a minimal number of phenomena to rule on the over or under-representation of words in a given context. For this reason, hapax legomena and low frequency words are typically excluded from analysis. A common threshold would be  $F \geq 5$  for an average sized corpus or  $F \geq 10$  or higher for bigger textual compilations.

<sup>9</sup> With the TtFf parameters for the Hypergeometric Model, measuring co-occurrence from  $A$  to  $B$  can provide a different score than from  $B$  to  $A$ .

## 4 Co-occurrence networks

Because of their logic of construction (in context, every word co-occurs – directly or not – with every other word) co-occurrence graphs can build up exponentially and produce unreadable results. Choices must therefore be made with a view to filtering out less important word attractions: minimal frequency and minimal specificity are the basic filters available. Even at co-frequency and specificity thresholds of 50 and 25, a total of 25 networks are detected in the corpus.<sup>10</sup> Most of these have two or three components: [*concentrado, rico*] (concentrated, rich), [*como, aperitivo*] (as, aperitif), [*lote, castas*] (lot, varieties), [*sempre, presente*] (always, present), [*são, os*] (are, the), [*sauvignon, blanc, cabernet*], [*framboesa, groselha, morango*] (raspberry, currant, strawberry), [*muita, bela, frescura*] (very, beautiful, freshness). Others, more elaborate describe semantic fields: [*região, apesar, marca, da, as, onde, aqui*] (region, despite, brand, of the, the, where, here).

Of the 25 networks detected in our corpus, one represents the major structure in a text with over 230 lexical components. Despite the strict thresholds for network extraction, the graph hereafter (Figure 1) is huge and as a result does not lend itself to close-up analysis but rather calls for observation from a distance. What should be noted in the following network is the gross organization of nodes: some are strongly connected to the graph by several links, others are attached by one relation. This topological view helps identify nodes of great importance whose absence from the graph would considerably alter its structure whilst others could be removed from the graph without any consequence. From a linguistic perspective, these opposite profiles correspond to words that either have enormous consequences on their contexts whenever they appear, or very little influence on their surroundings. Some words trigger a cohort of co-occurring words, others entail little, if anything, in context.

On closer inspection of the graph, the salience of over a dozen nodes corresponds to the identification by statistics of the essential vocabulary in wine tasting. Some words are outstandingly magnetic and form constellations around themselves: *taninos, fruta, fruto, frutos, leve, notas, cor, boa, corpo, longo, boca, final, acidez* and *preta* (*tanins, fruit(s), light, notes, color, good, body, long, mouth, end, acidity and black*). By altering the statistical thresholds, the graph can be reduced to under 150 word-types, which makes the figure easier to read close-up (see appendix, Figure 2).

This first glance at the figure shows that from a topological angle the underlying structure of the co-occurrence network is clear and meaningful: co-occurrence activity is uneven among the lexical components of our corpus. Some essential words contribute fundamentally to structuring the text. How does this contextual prominence translate to lexical importance?

## 5 From map to dictionary

Once the co-occurrence network is extracted from the corpus, what (type of) information does it provide?

Reading the graph in Figure 2 is an unpredictable experience. Any node can be a point of entry into the mesh. Any group of words can be read to form a meaningful set in one's mind which corresponds – or not – to an attested sequence in context. Here lies the complexity of the virtual network: all associations are presented simultaneously. Whilst multi-word expressions, grammatical constructions and all kinds of dependencies are suggested in the graph, only some really exist in the corpus. Consider the aforementioned example [*muita, bela, frescura*] (very, beautiful, freshness). Several constructs

<sup>10</sup> This implies that only words which coincide in context at least 50 times with a hypergeometric specificity of +25 will qualify for the network. Depending on the volume of the corpus, only words with the strongest power of co-occurrence should appear on the graph.

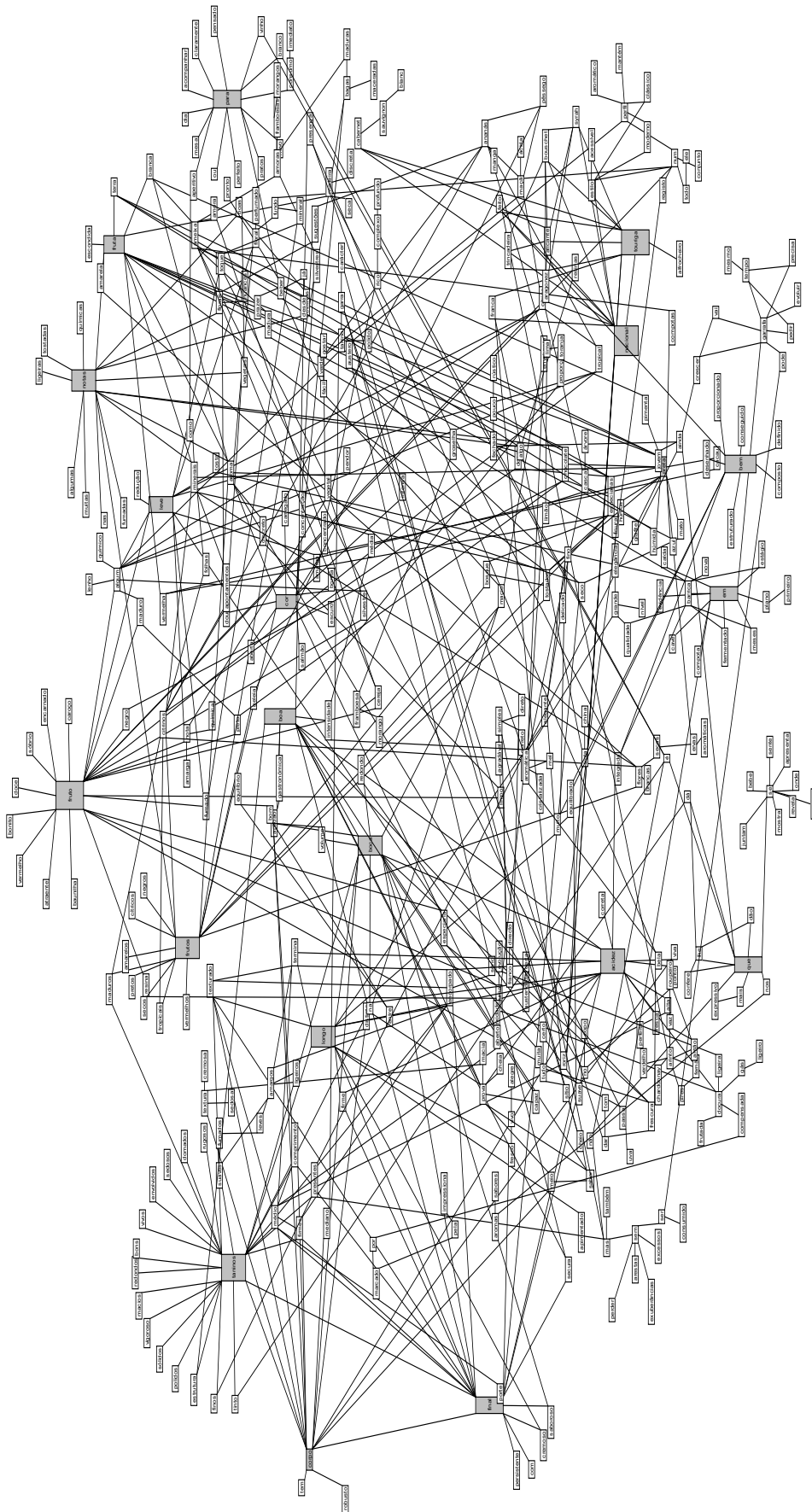


Figure 1: Abstract view of a co-occurrence network graph.

are possible in theory (“*muita bela*”, “*muita bela frescura*”, “*frescura muita bela*”) yet this network actualizes in the corpus in only two forms: “*muita frescura*” 222 occ. and “*bela frescura*” 71 occ.

Is this misleading? Not if one accepts that the graph shows words according to their statistical importance. If we bear this in mind, our interpretation of the network is cautious: all co-occurrences are significant but not necessarily meaningful. Working on the map as a foundation, a lexicographer is shown the entire backbone of the *corpus*. In this hierarchized schema, he can consciously choose which nodes to investigate and which to ignore.

The relation between an object and its representation is often described by the semiotician Alfred Korzybski’s famous words “the map is not the territory”, which he extended to a more domain-specific “the word is not the thing”. The general idea is that perception always intercedes between the observed and observer. This inevitable distortion rears its head in any human-based enterprise, including lexicography.

Let’s consider WordNet (Miller 1991), a (mostly) handwritten lexical database that was started by psycholinguists in the 1980s and is now emulated in many different languages. Its 1,7000 synonymy sets are interlinked by conceptual relations all determined by lexicographers. As Maziarz (2013) points out, these synsets are *de facto* the building blocks of the thesaurus, not words. Thus, pre-imposed synonymy becomes the norm and poses a problem of circularity in database construction and maintenance: WordNet presents words as synonyms because someone upstream deemed them to be.

When we set out to build a Portuguese terminology for wine we wanted to avoid, as much as possible, all initially built-in flaws. Therefore, our main objective was to implement an unsupervised method and apply it to a representative corpus. We consider the extracted data – filtered by strict statistical thresholds – to be exhaustive and objective, thus presenting a representative view of lexical phenomena in our wine-tasting compilation.

However, while typical problems do exist in our database they appear under a new light. Take circularity, for example. Whatever the thresholds we set for contextual exploration, small subsets of words mutually defining each other in very closed systems continue to emerge from co-occurrence network analysis. Yet, when these isolates appear beside major word networks, we are able to visually appreciate their importance vis-à-vis the main co-occurrence structures. Indeed, in network theory, these cliques are a well-documented phenomena, and their integration to the general structure is a matter of graph algebra not a decision made by the lexicographer.

Our main preoccupation is to preserve, as honestly as possible, the structure of our corpus, both syntactically and semantically. Indeed, as Korzybski underlined, the single most important item of information is the structure: a map *is not* the territory it represents, but, if correct, it has a *similar structure* to the territory, which accounts for its usefulness – Korzybski (1933).

After accumulating knowledge produced by total co-occurrence analysis for all word-types in the corpus provides, we consider this to be a statistically validated basis for the production of a distributional thesaurus. Indeed, distributed representations of words learned from text have proven to be successful in various Natural Language Processing tasks, such as word sense disambiguation, information retrieval or document summarization. In recent applications, distributional similarity has successfully been exploited as an approximation to semantic similarity. Kilgarriff and Rychly (2007) present an automatically produced thesaurus which identifies words which occur in similar contexts as the keyword, and draws on the hypothesis of distributional semantics. Ziai *et al.* (2016) use distributional semantics to support qualitative insights into the data and identify phenomena at the lexical level. Notably, Maziarz *et al.* (2013) recenter their Polish WordNet on lexical units in order to automatically construct synsets out of words with similar connectivity.



## 6 Conclusion

In this paper we have shown how co-occurrence analysis can be applied in the process of information retrieval for deriving lexicons from a domain-specific corpus. The results of generalized co-occurrence analysis for all word-types show circumscribed lexical systems that operate in the text. These structures display semantic homogeneity and can be interpreted as sense clusters, where linked words all serve to complete and precise their meaning in context.

After looking at a selection of 21,000 wine tasting notes, our experiment made it possible to extract from a word-type dictionary of over 7,500 terms a list of 300 words with high co-occurrence activity and establish evidence for repeated meaning building in context by association or dissociation of words as measured by positive and negative co-occurrence. Where typical distributional *thesauri* identify only words that occur in similar contexts to the keyword to posit their synonymy, our co-occurrence network provides data for the production of a more precise thesaurus. Following Greffentette (1994), the analysis of second order co-occurrences (co-occurrences of co-occurrences) can identify sets of synonyms (words that appear in the same type of context as the keyword). An extension of this logic would be to exploit negative co-occurrences for a given keyword and detect their second order co-occurrences so as to build sets of antonyms within a domain vocabulary.

Future work involves expanding the corpus to provide a larger image of the list of words and co-occurrences used to describe the main aspects: visual, olfactive and gustatory, in order to understand, in more detail, the apparent non-complex verbalization of the wine-tasting experience.

## References

- Abhik J. & Pawan G. (2018). Can Network Embedding of Distributional Thesaurus be Combined with Word Vectors for Better Representation? In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*. New Orleans. June 2018.
- Čermák F. (2006). Collocations, Collocability and Dictionary. In *Proceedings of the 12th Euralex International Congress*, 2006. Turin, publisher Edizioni dell'Orso, pp 929-937, 2006.
- Curran J. & Moens M. (2002). Improvements in Automatic Thesaurus Extraction. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX): Unsupervised Lexical Acquisition*. pp 59-66. 2002.
- Deghani M. et al. (2016). Luhn Revisited. Significant Words Language Models. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. Indiana. pp 1301-1310, 2016.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Special Volume of the Philological Society*. Oxford, Oxford University Press. 1957.
- Grefenstette G. (1994). Corpus-derived first-, second- and third-order word affinities. In *Proceedings of Euralex*, pp 279-290, Amsterdam, Holland. 1994.
- Guthrie D., Allison B., Liu W., Guthrie L., Wilks Y. (2006). A closer look at skip-gram modelling, In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy, 2006.
- Harris Z. S. (1968). Mathematical Structures of Language. John Wiley, New York. 1968.
- Kilgariff A. (2005). Putting the Corpus into the Dictionary, In *Proceedings of Meaning Workshop*, Trento.
- Kilgariff A. & Rychly P. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague. June 2007, pp. 41- 44. 2007.
- Kim W., Wilbur J. (2000). Corpus-based Statistical Screening for Phrase Identification, In *Journal of the American Medical Informatics Association*, Volume 7 Number 5 Sep / Oct 2000.
- Korzybski A. (1995). Science and Sanity: an introduction to non-Aristotelian systems and general semantics, Institute of General Semantics; 5th edition, (1<sup>st</sup> edition 1933) 1995.



- Lin D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING98)*. pp 768-774. Montreal, Canada. 1998.
- Luhn H. P. (1958). *A Business Intelligence System*. IBM Journal. October 1958.
- Luísa & Pereira A., Santos & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications.
- Marziaz, M., Piasecki M., Szpakowicz S. (2013). The Chicken and Egg Problem in WordNet Design: Synonymy, Synsets and Constitutive Relations, In *Language Resources and Evaluations*. Volume 47, Issue 3, September 2013.
- Miller, G., Beckwith R., Fellbaum, C., Gross D., Miller K. (1991). Introduction to WordNet: An On-line Lexical Database, In *International Journal of Lexicography*. Volume 3, January 1991.
- Okamoto M. (2015). Is corpus word frequency a good yardstick for selecting words to teach? Threshold levels for vocabulary selection. *System*. Volume 51, July 2015.
- Periñán-Pascual C. (2015). The underpinnings of a composite measure for automatic term extraction. The case of SRC. In *Terminology across Languages and Domains*. Special Issue of Terminology. Volume 21. Issue 2. Edited by Drouin P., Grabar N., Hamon T. & Kaguera K. pp 151-179. 2015.
- Sinclair J. (1991). *Corpus, Concordance, Collocation*, OUP, Oxford. 1991.
- Tsegaye R., Wartena C., Drumond L. & Schmidt-Thieme L. (2016). Learning Thesaurus Relations from Distributional Features. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. pp 442–446, Portorož, Slovenia. 2016.
- Verdaguer I. & González E. (2004). A lexical database of collocations in scientific English, In *Proceedings of the 11th Euralex International Congress*, 2004.
- Wanner L., Bohnet B., Giereth M., (2006). What is beyond collocations? Insights from Machine Learning experiments, In *Proceedings of the 12th Euralex International Congress*, 2006, Turin, publisher Edizioni dell'Orso.
- Wanner L., Bohnet B., Giereth M., (2006). Making sense of collocations. *Computer Speech & Language*, volume 20, number 4, pp. 609-624. 2006.
- Weeds J. & Weir D. (2005). Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *N Journal of Computational Linguistics*. Volume 31. Issue 4. December 2005. MIT Cambridge, MA, USA. 2005.
- Ziai R., De Kuthy K. & Meurers D. (2016). Approximating Givenness in Content Assessment through Distributional Semantics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. pp 209–218, Portorož, Slovenia. 2016.
- Zipf G. K. (1965). *The Psycho-Biology of Language. An Introduction to Dynamic Philology*. The MIT Press. Cambridge, MA, USA. 1965.

## Acknowledgements

The second author gratefully acknowledges the financial support of the Fundação para a Ciência e Tecnologia PhD grant (PD/BD/52261/2013) and NOVA FCSH - CLUNL for this research.





# Using Diachronic Corpora of Scientific Journal Articles for Complementing English Corpus-based Dictionaries and Lexicographical Resources for Specialized Languages

**Katrin Menzel**

*Dept. of Language Science and Technology, Saarland University*

*E-mail: k.menzel@mx.uni-saarland.de*

## Abstract

As technology and science permeate nearly all areas of life in modern times, there is a certain trend for standard dictionaries to bolster their technical and scientific vocabulary and to identify more components, for instance more combining forms, in technical terms and terminological phrases. In this paper it is argued that recently built diachronic corpora of scientific journal articles with robust linguistic and metadata-based features are important resources for complementing English corpus-based dictionaries and lexicographical resources for specialized languages. The Royal Society Corpus (RSC, ca. 9,800 digitized texts, 32 million tokens) in combination with the Scientific Text Corpus (SciTex, ca. 5,000 documents, 39 million tokens), as two recently created corpus resources, offer the possibility to provide a fuller picture of the development of specialized vocabulary and of the number of meanings that general and technical terms have accumulated during their history. They facilitate the systematic identification of lexemes with specific linguistic characteristics or from selected disciplines and fields, and allow us to gain a better understanding of the development of academic writing in English scientific periodicals across several centuries, from their beginnings to the present day.

**Keywords:** English diachronic corpora, Graeco-Latin combining forms, scientific journal articles, scientific vocabulary, corpus-based dictionaries, lexicographical resources for specialized languages

## 1 Introduction

The value of large corpora for the complementation of lexicographical resources has been demonstrated by various scholars, e.g. Podhajecka (2010, 2011) using Google Books. Although this huge book corpus posed some challenges due to partly incorrect metadata (cf. James & Weiss 2012) and poor OCR quality, it proved useful, e.g. for antedating various OED headwords. This paper will argue that recently created, large specialized corpora representing specific text types, technical areas, academic vocabulary and domain-specific words from different stages of the English language have not yet been exploited to their full potential. Such resources will enhance the systematic identification of lexemes and expressions with predefined linguistic features which otherwise may be found only rarely or randomly in general corpora as broad samples of language use that may lack fine-grained annotations. Numerous citations that illustrate the natural usage of words in current dictionaries, such as the *Oxford English Dictionary* – the “definitive record of the English language” – have been drawn from general language texts and from various canonical and culturally important types of sources, such as literary, theological, historical and philosophical works of renowned authors and particular time periods. Nowadays, such lexicographical resources are updated constantly with examples from Present-Day English that are often drawn from the internet, from digital archives and electronic corpora, e.g. from newspapers, magazines or blogs. It is worth investigating whether a more differentiated and complete view of the English language can be achieved by integrating more information retrieved from enriched domain-specific, specialized electronic corpora that are becoming more widely

available and searchable for a variety of linguistic patterns. An important source for lexicographical research are scientific periodicals serving as vehicles for the communication of new discoveries and ideas and as repositories of knowledge. In this paper, it will be shown how two large scientific journal corpora can be used as linguistic resources for lexicographers. If we take, for instance, items containing combining forms (CFs), these data offer valuable insights into the general linguistic and terminological developments in scientific writing, particularly with regard to topics related to natural phenomena and the study of the material world.

## 2 Resources

The analyses presented in this paper are based on two English scientific journal corpora that can be queried via the Saarbrücken CQPweb (Corpus Query Processor) interface: the Royal Society Corpus (RSC; Kermes et al. 2016) and the Scientific Text Corpus (SciTex; Degaetano-Ortlieb et al. 2013).<sup>1</sup> These corpora cover scientific articles from the early stages of the first scholarly journals published in English from the middle of the 17<sup>th</sup> century onwards to contemporary scientific publications. The latest release of the RSC (V3.7.0) consists of around 9,800 digitized texts (ca. 32 million tokens) published between 1665 and 1869 in the *Philosophical Transactions* (*Philos. Trans.*) and the *Proceedings* (*Proc.*) of the Royal Society of London – publications that figure among the longest-running English scientific journals in the world. These journals used to cover all branches of science of the time. Later, when the number of scientific journals grew considerably, they became increasingly specialized. Each of them split into separate series (series A covering mathematics and the physical sciences, and B focusing on the life sciences). The authors were British fellows of the Royal Society and other outstanding scientists of the time. Early research articles resembled letters addressed to a wider audience in many ways. We see a development in this genre from a more involved to a more informational style, and the development of research community standards with regard to terminological consistency and typical phrases and collocations. Linguistic precision, objective style and the features of plain English became more dominant textual features over time. We are currently extending the RSC corpus in the framework of the project *Information Density and Scientific Literacy in English – Synchronic and Diachronic Perspectives*<sup>2</sup> by integrating all texts from these publications, including their A and B series, from five additional decades into the corpus. (1870-1920). This resource will therefore cover the entire Late Modern English period (LModE, ca. 1700-1900) as well as the transition periods at the beginning and the end of LModE, i.e. the end of the Early Modern English period and the beginning of Present-Day English. The ongoing corpus extension will also increase the corpus size substantially as these journals, particularly *Proc.*, grew in size over time.<sup>3</sup>

The RSC contains annotations at different linguistic levels, for instance, part-of-speech and lemma annotations and information on historical spelling variants. It provides a number of query and graphical

1 The RSC has been made available for free download and online query from the CLARIN-D center at Saarland University under the persistent identifier <http://hdl.handle.net/11858/00-246C-0000-0023-8D1C-0>, cf. also <http://corpora.clarin-d.uni-saarland.de/cqpweb/>. Access to SciTex can be granted to researchers upon request. As it includes texts from the more recent past, there are stricter copyright restrictions than with the RSC and users interested in these data for research purposes cannot easily be given unrestricted access to the full corpus texts.

2 [http://www.sfb1102.uni-saarland.de/?page\\_id=275](http://www.sfb1102.uni-saarland.de/?page_id=275)

3 One might argue – given that the *Philosophical Transactions* of the Royal Society of London already are among the top ten sources of the OED with almost 16,000 quotations – that there is no further need to analyse more material from this journal or similar resources, at least not for this dictionary. Indeed, numerous quotations in the OED come from scientific journals, but a considerably much higher number has been drawn from other specific sources representing entirely different text types such as newspapers, poets and literary prose writers (e.g. ca. 42,000 quotations from the Times, ca. 33,000 from Shakespeare's works, more than 5,000 from Milton's *Paradise Lost* etc.).



visualization options.<sup>4</sup> Various types of metadata that were collected or generated are encoded for each text, e.g. information on the author, publication date, text type (e.g. abstract, book review, full article etc.) and text topics (based on probabilistic topic modelling, cf. Fankhauser, Knappen & Teich 2016) so that the texts can be sorted according to subjects such as optics, thermodynamics, botany, metallurgy, medicine, chemistry, etc. To obtain more variables for analysis, we are currently enriching the corpus with more fine-grained syntactic and morphological annotations and more types of high-quality metadata, some of which we recently obtained from the newly indexed and digitized Royal Society Journal Collection. Language models were built to estimate predictability and information density at the level of words and certain word-internal elements. Surprisal, an information-theoretic operationalization, has been calculated for various units (e.g. tokens and various word-internal elements) and has been annotated. Our current results confirm an increasing specialization of scholarly work reflected in the growing number of informationally dense expressions and structures. Additionally, an ongoing conventionalization of various linguistic structures can be observed over time.<sup>5</sup> The RSC is complemented with SciTex, an equally well-annotated large corpus of more contemporary texts from the 1970/80s and the early 2000s. It consists of English scientific journal articles from several disciplines, such as computer science, linguistics, biology, and mechanical engineering. The current version contains approximately 5,000 documents and 39 million tokens. Like the RSC, it is also tokenized, lemmatized and part-of-speech tagged, and contains rich metadata, e.g. information on the author(s), discipline/topic and year of publication.

Working with data from these specialized corpora by taking full advantage of their linguistic and metadata annotations and their sophisticated query options has substantial potential for making a unique contribution to lexicographical knowledge. The data offer possibilities for compiling or enhancing discipline-specific reference sources, e.g. terminology databases, technical dictionaries and encyclopaedias, and for assessing the vocabulary coverage of general dictionaries. Early scientific and technical terms are a particularly important part of English vocabulary. They can give us essential insights into co-developments between language and society. Many early terms have changed with the evolution of science and technology, some have dropped out of use, and others have become common in non-specialized use or in even more specific contexts than in the past. Large historical dictionaries like the OED may benefit from the possibility of incorporating newly retrieved information resulting from the systematic search in the RSC and SciTex for certain types of lexemes and word-formation patterns, as well as for typical collocations and multiword expressions that function as technical terms. If we take, for instance, the adjective *ethereal*, general dictionaries emphasize its chiefly literary or poetic usage or its function as synonym for *beautiful*, *delicate* or *heavenly* in formal contexts. The OED illustrates this with quotes from various literary and other sources. The word is quoted there less frequently in precise technical meanings from scientific contexts where it occurs in specific, term-like adjective-noun-patterns for which we find many historical and contemporary examples in our corpus data. The nouns in this pattern become more and more specific over time in the corpora. Many can be found in the 1740s (*ethereal fire* / *matter* / *medium* etc.), in the 1850/60s (*ethereal solution* / *odour* / *filtrate* / *extract* etc.) and in the 20<sup>th</sup> century (e.g. *ethereal diazomethane*).

The RSC in combination with SciTex as two recently created corpus resources that offer the possibility to provide a fuller picture of the development of specialized vocabulary and of the number of meanings that general and technical terms have accumulated during their history. They facilitate the systematic identification of lexemes with specific linguistic characteristics or from selected disciplines

4 Cf. also Knappen et al. (2017) for annotation quality, OCR correction methods and the evaluation and improvement of part-of-speech tagging in the RSC.

5 Surprisal is an information-theoretic notion measuring the probability of a linguistic unit to occur in a given textual context. It has the advantage of accounting for probabilities conditioned on a context, which cannot be achieved by considering mere frequencies (see also Degaetano-Ortlieb and Teich (2016) for a comparison of surprisal and type-token ratio to investigate productivity).

and fields, and allow us to gain a better understanding of the development of academic writing in English scientific periodicals across several centuries, from their beginnings to the present day.

### 3 Case Study on Combining Forms (CFs)

The following section reports some illustrative findings from a case study on different morpheme types in our corpora which underline the relevance of such resources for lexicographical and terminographical purposes. The main focus of this chapter will be on Graeco-Latinate combining forms<sup>6</sup> in English derived from classical nouns, verbs and adjectives, a morpheme type playing an important role in numerous scientific formations in combination with other elements that are also often of Latinate or Greek origin. One example of these CFs is *-lysis* and its variants, which can be traced back to Greek roots (λύειν ‘to loosen’ or λύσις ‘loosening’). It exemplifies the complexity of this morpheme group as it plays a role in a variety of lexeme-formation processes. This CF co-occurs with various root morphemes and affixes in our data (e.g. *photo+lysis*, *para+lytic+al*, *dia+lys+er*, *hydro+geno+lysis*). In earlier texts, it occurs in various neoclassical compounds and derivations. In texts from the more recent past, new lexemes also reflect current trends in word-formation processes, such as clipping and blending involving CFs. There are also backformations with word-class changes and newly coined independent lexemes that have evolved from initially bound CFs into free morphemes (e.g. *lysis* as noun and *lyse* as verb) as well as some hybrid forms with neoclassical and native English elements (e.g. *LysoTracker*, a trademark for a lysosome tracker). Term formation in the early stages was frequently based on adjective-noun patterns with at least one CF involved. The methods and insights from the results on CFs can be generalized to other morpheme types (e.g. pre- and suffixes), specific word-classes, word-formation processes or part-of-speech patterns if they are characterized by formal similarities so that they can be identified via corpus queries.

Various monolingual English dictionaries apply the category of CFs to initial and final bound lexical elements in the description of complex words and their components, but the information available on CFs in dictionaries can only give us a rough estimation of how productive these elements are and how many different words have been coined with them in English as a whole, or in particular time periods or registers. As many words with such forms tend to be rare in average texts and general corpora of modern English, they are sometimes assumed to play only a marginal role in the language as a whole. However, in scientific and technical English their components are very productive. In diachronic English corpora, such as our specialized corpora of English academic writing, they are related to important register-specific word-formation patterns. They are also very interesting from a cross-linguistic perspective, as they have often been incorporated into English as lexical or phraseological borrowings from other European languages or vice versa, and in many cases serve as scientific internationalisms. As argued by ten Hacken and Panocová (2014), it is useful to include entries with such neoclassical formatives in lexicographical resources and to link them in electronic dictionaries with the entire group of words they appear in (for a summary of the treatment of CFs in lexicographic and lexicological resources and studies and the role of combining forms in scientific discourse cf. also Menzel and Degaetano-Ortlieb (2017: 187-209)). McCauley (2006) gave an overview on some aspects of the revision of etymology sections in CFs entries in the 3<sup>rd</sup> edition of the OED, and explained that dictionary editors hope to take advantage of newly available data to provide more consistent, transparent and informative entries for CFs.

6 Lexicographers nowadays also include other word-formation elements of native and non-native origins in the list of combining forms, e.g. forms derived from prepositions (e.g. *by-*) or affixes (*-ene*), truncated words or clipped word fragments (*splinters*) such as *-gate* in the sense of *scandal* and pseudo-morphemes that have undergone semantic and structural reanalysis by processes of analogy (e.g., *-aholic*, *-(a)-thon*) which are productive sources for novel blends in contemporary creative language use, advertisements, media discourse and quasi-technical jargon.

This group of morphemes presents various challenges. As a potentially open-class category due to their lexeme-like semantics, CFs do not share specific semantic features and cannot be identified automatically as a group through a certain length or specific combinations of letters. Around 2,300 elements are currently identified as CFs in the OED, but there seem to be many more that have been used or are still used productively in word-formation processes in English. Other resources, e.g. those edited by Quinion (2005) or Sheehan (2000), focus specifically on lists of word-initial and word-final elements. Such elements are typically presented with no strict dividing line between affixes and CFs, and are labelled with more general terms in their description, e.g. *word parts*, *vocabulary elements* or *word beginnings* and *endings*. In addition to alphabetical lists, word parts are sometimes presented in thematic lists, for instance on anatomy and physiology. In Sheenan (2000: 187) one example for a thematic list are morphemes related to ‘breath’ such as *afflat-* (*afflatus*); *-hale* (*exhale*); *halit-* (*halitosis*); *ozostom-* (*ozostomia*); *-pnea* (*apnea*); *pneo-* (*pneograph*); *pneumato-* (*pneumatometer*); *pneumo-* (*pneumobacillus*); *pneumono-* (*pneumonophorous*); *pneusio-* (*pneusiobiognosis*); *pulmo-* (*pulmonary*); *spiro-* (*spiograph*). The information on CFs obtained from such existing specialized lexicographic resources and general corpus-based dictionaries can be taken as a starting point for the search for domain-specific or semantically related CFs in order to verify and complement our current knowledge through the analysis of our corpora. This will allow us to identify and record new words, specific meanings and semantic shifts and to check whether first attestations in established dictionaries can be antedated.

If we search, for instance, the morpheme group *pneo-*, *pneu-*, *pulmo-* *spiro-* with a CQP query in the RSC and SciTex, all lexemes and collocational patterns with these components can be retrieved in these annotated corpora of scientific texts from Early Modern English to Present-Day English, currently comprising 71 million tokens. These CFs occur around 1,700 times in about 420 different texts, and are slightly more frequent in the older data from the RSC than in SciTex.<sup>7</sup> Among the earliest words with these components in the data are *pneumatic* and *pneumatical* (e.g. in *pneumatic / pneumatical engine*), and *pneumatiques* (as in *science of pneumatiques*) in the 17<sup>th</sup> century, indicating some lexical, terminological and spelling variation in scientific writing of this time. Some early terms and collocations were already relatively complex patterns, e.g. *hydraulopneumatical fountain* (1660s/70s). The most frequent result found with this query in both the earlier and the more contemporary corpus is the adjective *pulmonary* (typically in adjective-noun patterns such as *pulmonary artery / vein / vessels* etc., again with some similarly structured near-synonyms, e.g. *pneumonique vessels / arteries* etc.). Newer words with one of these CFs (e.g. *electropneumatic* or *pneumoencephalography*) reflect scientific developments and specialization processes. Many lexemes with these elements become established and conventionalized; and increasingly more derivatives and combinations with native forms can be found in the data, e.g. *spirometrically-measured* in a text on bioinformatics from the 1970s. It is possible to focus on selected n-gram types and part-of-speech patterns to find out which are most typical for particular time periods, e.g. trigrams such as adjective-adjective-noun patterns that function as collocations or compounds (as in *pneumatic mercurial apparatus*, *mercurial pneumatic apparatus* (1800s/1810s), *great pneumogastric nerve*, *pulmonic capillary vessels*, *pulmonary semilunar valves* (1830s-1860s)). In earlier texts, there is a noticeably higher lexical variability in our data, later texts are characterized by fewer general-language words and more specific adjectives, more standardized terminological usage and adjective-adjective-noun patterns embedded in more complex nominal structures (e.g. *the anterior aortic and pulmonary semilunar valve-rudiments* (1869)).

Another example is the CF *iso-* and its Latin-based equivalent *equi-* (‘equal’).<sup>8</sup> They often occur in multimorphemic adjectives. Related lexemes, derivations and frequencies of variants can easily

7 A few lexemes are filtered out by more precise queries out as *spiro-* (from Latin *spīrāre* – ‘to breathe’) can also have a different sense (from Latin *spīra* ‘coil, spiral’) as in some names of spirally shaped bacteria.

8 Variants are *is-* (possible before vowels, but does not occur in our dataset) and *aequi-* (in older texts, e.g. *aequivocal*).

be found and extracted (e.g. *isochronous*, *isochronal*, *isochronic*, *isochron*, *isochronously*, *isochronism*). Additionally, the data may be sorted and filtered according to selected categories of metadata to obtain relevant quotations from specific time periods or types of authors if they appear to be underrepresented in the dictionary quotation database. It is possible to select only those quotations, for instance, from early female scientists such as Caroline Herschel (who used this CF in the term *isosceles triangle*), from American scientists from the 18<sup>th</sup> century such as Benjamin Franklin (using for instance the noun *equinoctial*), authors that may be slightly less well-known today, but published far more texts than others, e.g. Everard Home (using *equivocal* as a predicative adjective) or from specific decades or other time slices (e.g. some words seem to be under-represented in 18<sup>th</sup> century quotations in dictionaries).<sup>9</sup> Newer texts from SciTex include few coinages with the Latin-based *equi-*, but several lexemes that were coined from the end of the 19<sup>th</sup> century onwards with the CF *iso-* (e.g. *isochore*, *isologous*, *isoleucine*, *isocyanate* and *isooctane*). It is possible to sort the results according to topics and disciplines that are specified in the document metadata. Texts on the solar system in the RSC, for instance, frequently include these CF in words such as *equinoctial* (or sometimes *equinoxial*), *equinox*, *equidistant* / *equi-distant* and in occasionally occurring words such as *equilateral*, *equiangular*, *isosceles* and *isochronal*. Texts from similar time periods in the RSC on chemistry have more types and tokens with these CFs in total. Among the most frequently used lemmata with these forms in chemistry texts are *equivalent*, *isomer*, *isomeric*, *isomerism*, *isomorphous*, *isomorphism*, *isoprene*, *isopropyl*, *equi-diffusive* / *equidiffusive* and some adjectives related to geometry such as *equidistant*, *isosceles* and *equilateral*. Texts with mathematics as the primary topic in the RSC include frequent occurrences of *equivalent*, *equilibrium*, *equidistant*, *equilateral*, *isotropic* and some occurrences of lexemes such as *isodynamic*, *equipotential*, *isoperimetrical*, *equimultiples*, *equiponderant*, *equiradial* and *isobarism*.

CFs like these examples are distributed differently across research articles from different academic disciplines, but there is also variation across different text types. The two morpheme groups *pneo-* / *pneu-* / *pulmo-* / *spiro-* and *iso-* / *equi-* are most frequently found in abstracts (ca. 63 and 493 occurrences per million words (pmw), respectively) while the lowest frequencies can be found in review papers (ca. 35 and 92 occurrences pmw). Text types with high frequencies of these initial CFs are not only characterized by a large number of low-frequency forms indicating a high productivity rate of these morphemes, but also by a certain use of lexical repetition and high frequency terms. Final CFs, particularly those with less specific or more abstract meanings such as *-(o)logy* and *-(o)graphy*, tend to play the most prominent role in review papers in the data. These two forms taken together occur ca. 214 times pmw in reviews compared to 133 in abstracts and 70 in full research papers. They steadily increase in frequency over time in all corpus registers. Earlier corpus texts contain a relatively low number of hybrid forms. Newer coinages more freely combine neo-classical elements of Greek and Latin origin with each other. In the more contemporary corpus texts, we also find more independent lexemes that have evolved from bound CFs (e.g. *graph*, *photo*, *bio*) and more combinations of CFs with native elements and fully anglicized lexemes of various origins (e.g. *cell-biology*, *mailgram*, *ultrasoundcardiography*, *polarography*), with proper names and acronyms (*roentgenology*, *galvanoscopic*, *DUV-lithography* etc.). Several low-frequency items, particularly in newer texts, contain more than two CFs within one lexeme indicating increasingly specialized discussions requiring more specific and complex terms (e.g. *psychobiology*, *photolithography*, *electrocardiography*, *cinephoto-macrography* and *electrophoresis-homochromatography*).

<sup>9</sup> Texts from female scientists are rare in the time span currently covered by the RSC as the Royal Society did not elect any female fellows until the 1940s. American (colonial) scientists supplied relatively few contributions until the end of the 18<sup>th</sup> century, but there are a number of articles written by such authors in this period on astronomy and electricity. The person who wrote most articles in the journals included our corpora is Everard Home. He published more research papers in *Philos. Trans.* than anyone else but he is not quoted in the OED from this source.



Highly frequent lexemes with CFs that are also used in non-scientific discourse and general language sometimes become semantically bleached over time towards a less literal usage or lose their compositionality (e.g. *apology*, *equivalent*), while many morphologically complex terms with CFs remain relatively transparent with regard to their component morphemes. This does not necessarily mean that the exact meaning of compounds or multiword expressions with CF can easily be derived from the literal meaning of their constituents without knowing a proper definition or the context in which they are used, but they give strong hints as to the meaning of technical and scientific terms. Classically educated scholars and physicists of the 19<sup>th</sup> century who were familiar, for instance, with the concept of the *luminiferous* (= light-bearing / light-producing) *ether* and with word-formation patterns with combining forms in general probably had no difficulties in decomposing similarly structured complex adjectives in phrases such as *sanguiferous vessels*, *carboniferous limestone*, *mortiferous diseases* or *odoriferous flowers*.

The corpus data may also serve to provide additional information to that contained in dictionaries. *Luminiferous*, for example, is quoted twice in the OED from *Philos. Trans.*, but it is not straightforward to see there from the information given with regard to date and author whether these sentences

(1) and (2) are from the same or from different text:

- (1) 1801 Young in *Philos. Trans.* (Royal Soc.) 92 22 The actual velocity of the particles of the luminiferous ether.<sup>10</sup>
- (2) 1802 T. Young in *Philos. Trans.* (Royal Soc.) 92 14 A luminiferous Ether pervades the Universe, rare and elastic in a high degree.

Queries for these passages in the RSC will enable users to browse detailed lists of document metadata to obtain information on the title of the paper, the author, text topics, and so on, and to find the full text by clicking on a DOI link. Then it can easily be seen that the text from which these two quotes were taken was read before the Royal Society in 1801 as the one of Thomas Young's Bakerian Lectures and published in 1802. Our corpus data indicate in which time periods certain CFs played a particularly important role (e.g. *-scope* was used especially productively in the RSC between 1750 and 1800), and we can find words that are morphologically related and not attested in the OED yet (e.g. *colloidoscope*), or where the OED records quotations only from a specific time period. *Laryngoscope*, for instance, is quoted in the OED only from texts between 1860 and 1880, but has been assigned to Frequency Band 4 indicating a moderately high frequency in current use.<sup>11</sup> In the RSC, it occurs a few times in the 1860s and in Scitex once in the 1970s. Preferred general, domain-specific or time-specific spelling variants can be detected in the data, e.g. in the 19<sup>th</sup> century the adjective *hypovanadous* was sometimes used in texts on chemistry, most frequently in the collocation *hypovanadous salt*.<sup>12</sup> In the OED, only the differently spelt *hypovanadious* is recorded which does not occur in our dataset.

If new terms are introduced, various types of recurring multiword sequences can be queried in the corpora as indicators for lexical innovation and neologistic forms (e.g. 'propose to call' as in 'I have constructed the instrument which I propose to call the Stereomonoscope' or 'Mr. Faraday proposes to call the acid [...] Sulpho-naphthalic Acid.'). Additionally, the data indicate how and by whom scientific terms and collocations were popularized in English in certain time periods, and in which domains they used to play a dominant role. The adjective-noun collocation *galvanoscopic frog(s)* is

10 cf. <http://www.oed.com/view/Entry/111121> and <http://www.oed.com/view/Entry/64728>

11 <http://public.oed.com/how-to-use-the-oed/key-to-frequency/>

12 *Hypovanadous salt* in the corpus data refers to another concept than *hypovanadic salt*, which is also indicated by the fact that both terms typically co-occur in the same corpus texts. Moreover, they have cognates in French texts on chemistry from similar time periods (*des sels hypovanadeux / hypovanadiques*). It should be kept in mind that lexemes that may look structurally very similar and could appear to be synonyms or term variants (as in the examples *pneumonic / pulmonary* and *pneumatic / pneumatical* discussed above) might refer to quite different scientific concepts.



an example that was relatively frequent in our data in texts from the 1840s and 1850s and also occurred a few times in texts on experiments, electromagnetism and physiology from later time periods. It was used repeatedly in various articles in *Philos. Trans.* by Carlo Matteucci, an Italian neurophysiologist. As already mentioned in Section 2, surprisal values as information-theoretic operationalization for measuring information density have been annotated in the data. The information gained from this probability value reveals an additional aspect to observed unconditioned frequencies and type-token ratios of words as indicators of productivity. Surprisal provides us with additional information on the probability of a linguistic unit to occur in a given textual context. It is defined as the negative logarithm of the probability of a unit (e.g. a word) in context (e.g. its preceding words), and it is measured in bits. The value indicates, for instance, whether a given word is more *surprising* after a sequence of preceding words and hence provides more information (see also Degaetano-Ortlieb and Teich (2016) for more details and a comparison of surprisal and type-token ratio). Surprisal values can be downloaded for the corpus query results or visualized with the tool implemented in CQPweb (Fischer, Fankhauser & Teich 2017).

Figure 1 shows a visualization of surprisal values for an extract from a corpus text from 1845. This figure shows, for instance, that in our corpus data, the adjective *galvanoscopic* most typically occurs before the word *frog*, sometimes also before words such as *effects*, *nerve*, *limb* or *leg*. The most typical contextual pattern in which this adjective is used is the noun phrase *the nerve of the galvanoscopic frog*. *Galvanoscopic* in general is not a frequently used adjective in the whole corpus. However, it has different surprisal values in different syntactic and lexical contexts. The probability of this adjective occurring after the word sequence *nerve of the* is relatively high, and therefore the surprisal value of this word in such contexts is low, indicated by the small font size in the visualization of Figure 1. After other preceding contexts it is less expected and thus visualized with a larger font size. However, the word *frog* can be expected with an increasingly high probability after this adjective.



Figure 1: Visualization of higher and lower surprisal values by the use of larger and smaller font sizes in a corpus excerpt from the Royal Society Corpus.

The examples presented in this section demonstrate how the RSC and SciTex can be used to effectively to search and identify lexemes, multi-word expressions and collocations in which combining forms play a role. They illustrate the insights which can be gained through a detailed analysis of the corpus resources.

## 4 Conclusion

The linguistic evidence found in our specialized diachronic corpora can complement English corpus-based dictionaries and lexicographical resources for specialized languages in several ways. The convenient query and visualization options in the Royal Society Corpus and Scientific Text Corpus provide various research routes for lexicographical and lexicological purposes. They allow us to browse an enormous number of specialized texts on a broad spectrum of topics with fine-grained annotations and rich metadata. This enables us to find linguistic structures with certain components, used by specific authors, in particular time spans or with respect to particular topics in a systematic way. This has substantial potential to lead to a more nuanced understanding of the developments and dynamics of language use in specialized registers over time. The examples given in this paper have particularly emphasized the usefulness of diachronic corpora for updating and improving historical dictionaries, but the methods described can also be applied to dictionaries of contemporary language. The information obtained from these corpora can be used to improve various corpus-based general and specialized lexicographical reference works.

## References

- Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E. and E. Teich (2013). SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Bennett, P. et al. (eds.). *New Methods in Historical Corpus Linguistics. Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP*, 3. Tübingen: Narr, pp. 93-104.
- Degaetano-Ortlieb, S. & Teich, E. (2016). Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of Language Technologies for the Socio-economic Sciences and Humanities (LATECH'16), Association for Computational Linguistics (ACL), 7-12 August 2016*, Berlin, Germany, pp. 165-173.
- Fankhauser, P., Knappen, J. and Teich, E. (2016). Topical Diversification over Time in the Royal Society Corpus. In *Digital Humanities 2016. Conference Abstracts. 11-16 July 2016*. Jagiellonian University and Pedagogical University, Kraków, Poland., pp. 496-500.
- Fischer, S., Fankhauser, P. and E. Teich. 2017. Visualization of Corpus Frequencies at Text Level. In *Proceedings of the Corpus Linguistics Conference, CL2017, Poster Session, 25-28 July 2017*, University of Birmingham, UK.
- James, R. and A. Weiss (2012). An Assessment of Google Books' Metadata, In: *Journal of Library Metadata*, 12(1), pp. 15-22.
- Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and E. Teich (2016). The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC'16, 23-28 May 2016*. Portorož, Slovenia, pp. 1928-1931.
- Knappen, J., Fischer S., Kermes, H., Teich, E. & Fankhauser, P. (2017). The Making of the Royal Society Corpus. In: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, 22 May 2017*. Gothenburg, Sweden, pp. 7-11.
- McCauley, J. (2006). Technical Combining Forms in the Third Edition of the OED: Word formation in a Historical Dictionary. In *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX), 17-19 March 2005*, Helsinki, Finland, pp. 95-104.
- Menzel, K. and S. Degaetano-Ortlieb (2017). The Diachronic Development of Combining Forms in Scientific Writing. In: *Lege Artis. Language Yesterday, Today, Tomorrow, The Journal of University of SS Cyril and Methodius in Trnava*. Warsaw: De Gruyter Open, vol. 2 (2) 2017, pp. 185-249.
- Oxford English Dictionary*. (2000- 3rd ed.), Oxford: Clarendon Press. Accessed at: <http://www.oed.com/> [28/03/2018]
- Podhajecka, M. (2010). Antedating Headwords in the Third Edition of the OED: Findings and Problems. In *Proceedings of the XIV EURALEX International Congress, 6-10 July 2010*. Fryske Akademy, Leeuwarden, Netherlands, pp. 1044-1064.

- Podhajecka, M. (2011). Research in Historical Lexicography: Can Google Books Collection Complement Traditional corpora? In S. Góźdz-Roszkowski (ed.), *PALC 2009: Explorations Across Languages and Corpora*. Frankfurt am Main: Peter Lang, pp. 529-546.
- Quinion, M. (ed.), (2005). *Ologies and isms: A dictionary of word beginnings and endings*. Oxford: Oxford University Press.
- Sheehan, M.J. (ed.). (2000). *Word parts dictionary. Standard and reverse listings of prefixes, suffixes, roots and combining forms*. 2nd ed. Jefferson, NC & London: MCFarland.
- ten Hacken, P. and R. Panocová (2014). Neoclassical Formatives in Dictionaries. In *Proceedings of the 16th EURALEX International Congress, 15-19 July 2014*. Bolzano, Italy, pp. 1059-1072.

## Acknowledgements

This study contributes to the project Information Density and Scientific Literacy in English – Synchronic and Diachronic Perspectives in the framework of the Collaborative Research Center 1102 with the title Information Density and Linguistic Encoding funded by the German Research Foundation (DFG).

# ELeFyS: A Greek Illustrated Science Dictionary for School

**Maria Mitsiaki<sup>1</sup>, Ioannis Lefkos<sup>2</sup>**

<sup>1</sup>*Democritus University of Thrace*, <sup>2</sup>*University of Macedonia*

E-mail: [mmitsiaki@helit.duth.gr](mailto:mmitsiaki@helit.duth.gr), [lefkos@uom.edu.gr](mailto:lefkos@uom.edu.gr)

## Abstract

This paper reports on the design and compilation of ELeFyS (Εικονογραφημένο Λεξικό Φυσικής για το Σχολείο, ΕΛεΦυΣ), a Greek specialized school dictionary of science. Since its conception ELeFyS has been intended as a reference tool for the parallel development of scientific and linguistic literacy in a school context. To fulfil such an objective, generic entries include scientific terms that fall within the school subject of physics and are likely to be encountered in the upper grades of primary and lower grades of secondary school; however, the dictionary coverage is not restricted to terminology, but is also expanded to the terms/headwords' respective general sense(s) and use(s). Moreover, encyclopedic and cultural material is given as further stimuli for critical thinking. Under this scope, ELeFyS works both as a lexicographic product and a multi-functional teaching resource. In sum, it constitutes a novel endeavor of combining pedagogy and specialization in order to meet the complex linguistic and cognitive/scientific needs of school children in the late primary and the early secondary school grades. Such a complex aim of determining both communication- and knowledge-oriented lexicographic functions is being realized thanks to the enduring collaboration of a linguist and a science expert, well-rooted in long teaching experience. In what follows, we focus on the policy decisions made at the outset of the lexicographic project and the entry-building process.

**Keywords:** Greek science dictionary, macro- and microstructure, content-based learning/instruction

## 1 Introduction

It is only recently that the Greek school community has embraced dictionary use, setting the basis for the establishment of a dictionary culture. The first pedagogical dictionaries were introduced into the Greek educational system as schoolbooks about 15 years ago (Antypa et al. 2006; Gavriliadou et al. 2008, among others), despite the fact that linguists had emphasized the need to initiate a school practice of dictionary use since the 1990s (Anastassiadis-Symeonidis 1997; Iordanidou & Mantzari 2004). Our lexicographic endeavor attempts to resituate the pedagogical dimension of dictionary use, by providing intensive opportunities to integrate a specialized dictionary into the school learning process.<sup>1</sup>

ELeFyS innovates in several aspects. Namely, it is:

- the first specialized science dictionary that has been compiled in Greece in order to foster content-based learning/instruction both in L1 and in L2, thus promoting reception and production of scientific terms and their respective use in general language
- a pedagogical dictionary intended to cover the specific cognitive, encyclopedic, linguistic and cultural needs of school children with respect to science, as they arise in various types of learning situations
- a monolingual dictionary, which establishes interlingual equivalence of scientific terms in five languages, thus being a useful reference tool for L2 learning in academic contexts
- an illustrated dictionary, as it provides visual tools represented by images and animation with

<sup>1</sup> Such an integration is more urgent nowadays, since most primary education curricula aim at (1) raising awareness of the relevance of science with regard to environmental and social concerns, and (2) promoting learning through inquiry (Harlen & Qualter 2014).



sound effects for the scientific terms and processes they entail, but also for the general word meanings, and finally

- an electronic dictionary freely accessible on the Internet ([www.elefys.gr](http://www.elefys.gr)), as it delivers the dictionary data via the use of digital media, thus circumventing the common dictionary conventions in terms of space limitation, and making imaginative use of new technology in order to ensure flexibility, user-friendliness, and a pedagogy-oriented format.

## 2 Theoretical Grounding: Lexicography and Scientific Literacy

ELeFyS draws on a wide range of theoretical inputs, so as reliance on intuition is kept to a minimum. As a monolingual dictionary it is firmly-grounded on lexicographic theory, taking into consideration the seminal works of Mel'čuk (1996), Rosch (1973), Lakoff (1987), and Geeraerts (1990) among others for lexical functions, prototypes and definition writing respectively; it is also informed by the latest trends in lexicographic practice (Rundell 2006; Atkins & Rundell 2008; Rundell 2012), especially regarding the use of digital media for delivering lexicographic data. As a pedagogical dictionary (Tarp & Gouws 2012) it takes into consideration the perceived cognitive and linguistic –academic and communicative – needs of first and second language learners in the late primary and early secondary school grades.<sup>2</sup> As a pedagogical specialized dictionary (see also Tarp 2005), it attempts to initiate young learners into the academic language of science, following though a more broad-brush treatment of the different senses and uses, considered to be suitable for the targeted user group (school children who are 10-13 years old). No further reference to theoretical lexicographic insights will be made at this point, as the team's decisions are justified on a theoretical basis both for macro- and microstructure in Section 3.

However, it should be mentioned that ELeFyS is also consistent with the well-established body of theory (see e.g. Driver et al. 1996; De Boer 2000; Osborne 2002; Plakitsi 2010) underpinning the importance of scientific literacy as a transferable outcome of science education. Although many educators advocate the naive belief that science is equivalent to empirical work in the laboratory and that scientific language is simple and unambiguous (Lemke 1990), scientific literacy is a far more complex concept, closely related to knowledge, linguistic performance, argumentation, cultural identity, and so on, presupposing that school children are able to communicate science and the language of science. In most cases, the learners' difficulties in understanding science are attributed to the complexity of its terms or concepts (Shayer & Adey 1981), a consideration which is partly valid, as it holds true for polysemy, an inherent property of language. Polysemy undoubtedly imposes an additional conceptual load, since terms are very often used with different meanings in general language, e.g. *κύκλωμα* “circuit”, *δύναμη* “force”, *ενέργεια* “energy”. Nevertheless, the difficulties encountered by learners in scientific language should also be traced to:

- contextual parameters
- the multi-semiotic practice of science
- the nature of its genres (Halliday 2004)
- the structural features of scientific discourse.

In particular, a word's precise meaning can only be determined by examining the context of its use, a process which requires learners to acquire skills of recontextualization, e.g. the term *ηλεκτρισμός* “electricity” could refer to various interconnected but different concepts, such as *ηλεκτρικό φορτίο* “electric charge”, *ηλεκτρική τάση* “electric voltage” or *ηλεκτρικό ρεύμα* “electric current” (Osborne 2002: 209). In addition, the multi-semiotic nature of science may hinder progress with regard to

<sup>2</sup> The term “pedagogical dictionary” is used in its broader sense, being targeted both to native and second/foreign language learners (Dolezal & McCreary 1999).



achievement in this subject. For instance, energy can be multi-dimensionally represented as a symbol (E), a diagram or a mathematical equation, a complex definition and so on. Scientific language is also impersonal, objective and distant, reflecting the description of physical phenomena through the eyes of an independent observer, making use of inquiry, report, explanation and argumentation. Moreover, it is cumulative, since each argumentation in any given scientific domain builds on ones that have gone before (Osborne 2002). Finally, it exhibits complex characteristics and structures, such as:

- lexical density, e.g. *άτομο* “atom”, *μόριο* “molecule”, *χημική ένωση* “chemical compound”
- high or +learned register, e.g. *ασκώ έλξη* “pull/attract” vs *νιώθω έλξη* “feel attracted”
- passive constructions, e.g. *το φως διαθλάται* “the light is refracted”
- extended use of subordinate clauses, e.g. *μαγνήτης είναι το σώμα που έχει την ιδιότητα να έλκει αντικείμενα από σίδηρο και ορισμένα άλλα μέταλλα* “a magnet is an object that has the property of attracting iron-containing objects and other metals”
- taxonomies, e.g. *υποατομικά σωματίδια: ηλεκτρόνια, πρωτόνια, νετρόνια* “subatomic particles: electrons, protons, neutrons”
- abstraction, e.g. *ύλη* “matter”, *ενέργεια* “energy”, *δύναμη* “force”
- nominalization, e.g. *η διάθλαση του φωτός* “the refraction of light”, etc. (see also Arapopoulou & Giannouloupoulou 2001; Anastasiadis-Symeonidis et al. 2014).

In this respect, “every science lesson is a language lesson” (Wellington & Osborne 2001: 2).

ELeFyS attempts to capture, codify and resolve the aforementioned inherent difficulties of the language of science by combining cognitive, linguistic, encyclopedic and usage information. Furthermore, its structure contributes to the term/general word recontextualization by exposing learners to numerous illustrative and illustrated examples. Finally, from a critical point of view, the information and prompts included function as stimuli for experimentation, reasoning and argumentation. This way, it serves a dual function as a dictionary and as an educational resource, which can be utilized for the application of innovative teaching approaches to science, making scientific content comprehensible to native, second language or foreign learners and portraying the similarities and differences between the scientific and general use of words. Needless to say that the comparative presentation of the words’ scientific and everyday meanings and uses facilitates the interconnectedness of scientific and linguistic literacy, which, in turn, opens the way for interactive approaches that “safeguard the subject being taught whilst promoting language as a medium for learning” (Coyle in Marsh 2002: 27), such as CLIL and content-based instruction.

### 3 ELeFyS: Design Principles and Description

#### 3.1 Macrostructure: Policy Decisions

##### 3.1.1 User Profile, Headword Selection and Resources for Entry-building

The user profile crucially affects the selection and presentation of the lexicographic information included in the dictionary, and drives the specific editorial decisions made with respect to its content and form (Atkins & Rundell 2008). ELeFyS exhibits the following range of potential users and uses. Firstly, it addresses the needs of children – native Greek speakers or second/foreign language speakers – in the school setting. Secondly, it caters for several types of uses that target the school children’s receptive and productive skills, such as:

- studying the science school subject (i.e. understanding the scientific discourse, producing oral or written argumentation, reports, essays, etc., preparing for a written or oral exam)
- general reference purposes (understanding unfamiliar lexical items, distinguishing words in

- general language from terms, finding information on word families, grammar, usage, etc.), and
- learning the Greek language and acquiring not only communicative but also academic skills.

As a result, the way that information is selected and presented is largely determined by what we know about the users' skills and knowledge. More specifically, 10-13 year-old school children – both native speakers and second language learners – are 'quasi-proficient' in the school language, as they have to face the discontinuity between the conversational focus of the primary Greek curriculum and the academic focus of Greek as a medium of instruction in secondary school (see also Cummins & Yee-Fun 2007). Such a difference between 'surface fluency' and cognitively-related skills calls for lexicographic decisions that aim at both communicative and academic competence or/and performance. This means that students in the late primary and early secondary school grades are in need of a different configuration of lexicographic facts, which makes use of pedagogical prompts establishing a learning environment of creativity and enjoyment, as well as of academically-related stimuli enhancing cognitive achievement and motivation. Under this scope, in compiling ELeFyS we opted for the more broad-brush treatment of terminology and the incorporation of a limited amount of scientific information, without, of course, misrepresenting scientific theory or violating its principles. Additionally, we adopted numerous pedagogy-oriented strategies, such as:

- substitution of metalanguage and abbreviations by lexicographic symbols
- illustration of terms and general-use words
- alternative wording of definitions
- a wealth of examples
- prioritization of word meanings
- appropriate grammatical and usage information
- translation of terms in five languages
- recorded pronunciation, etc.

On the other hand, in order to initiate students into academic discourse, we favored components such as:

- gradation of term definitions in terms of difficulty
- etymological information
- a manageable number of informative examples showing the lemma in use in its various meanings and patterns (e.g. nominalization, passive construction, subordination, etc.) and grounding the scientific theory to direct experience of physical phenomena, and
- notes/prompts for experimentation, further scientific study and critical thinking.

In order to decide about ELeFyS's coverage, we based our work on the following resources:

- the Glossary that accompanies each section of the Greek science school textbooks
- the online 2-million-word Greek School Textbook Corpus (from the Center for the Greek Language website), and
- the headword list included in equivalent dictionaries of other languages (e.g. the *Oxford Primary Illustrated Science Dictionary* and the *Oxford Student's Science Dictionary*).

Recourse to such resources is justified by the targeted user group and its needs; since the dictionary entries reflect the combinatorial behavior of scientific terms and their respective use in general language, school language corpora should provide the basis for headword selection.<sup>3</sup> Moreover, the terms' conceptual opaqueness and their semantic inclusiveness were introduced as additional lemmatization criteria. For instance, despite its centrality and wide coverage in the language of Science, the meaning of the term *ὕλη* "matter" is rather fuzzy and thus not easily conceivable by young learners;

3 Unfortunately, no extensive corpora of school discourse are available at the moment for the Greek language. The sample of classroom interaction data that is included in the Corpus of Spoken Greek is not representative of the language of science.

therefore, the term was included in the headword list. In such a way, less-frequent words denoting peripheral scientific concepts were excluded from the lemmatization process. For instance, the term *εξάχνωση* “sublimation” exhibits only five occurrences in the corpus (mostly in high school science textbooks), whereas the term *εξάτμιση* “evaporation” exhibits 65 occurrences (displaying an equal distribution in primary, secondary and high school textbooks). In sum, ELeFyS attempts to reflect and ‘ease’ the phenomenon of nominalization apparent in the language of science by lemmatizing noun terms characterized by abstraction or/and denoting processes, e.g. *άπωση* “repulsion” instead of *απωθώ* “to repel”.

### 3.1.2 Layout, Constituent Parts & Medium

The dictionary entries are arranged according to their semantic interconnectedness (see also Bowker 2003). Such an arrangement led to the grouping of terms into broader thematic fields, i.e. Properties of Matter & Atomic Structure, Heat & Temperature, Electricity, Energy, etc., which is also supported by the Greek science textbooks.<sup>4</sup> However, all lemmas that fall within a specific subfield are arranged in alphabetical order, e.g. *άτομο* “atom”, *μάζα* “mass”, *μόριο* “molecule” in the subfield “Properties of Matter & Atomic Structure”.

As a specialized dictionary of science, ELeFyS contains compiled information which is partly scientific, but also partly linguistic for the same lexical unit.<sup>5</sup> That is why each dictionary page corresponds to a distinct multi-lemma (see Figure 1), i.e. an entry that encompasses other morphologically and semantically-related lexical units.

**αγωγός (ο) (άγω)**

Η λέξη **αγωγός** προέρχεται από το αρχαιοελληνικό ρήμα *άγω* που σημαίνει «μεταφέρω». Στην αρχαία ελληνική σημαίνει «οδηγός, καθοδηγητής». Τη σημερινή επιστημονική σημασία, όμως, τη δανειστήκαμε από τη γαλλική (*conducteur*).

Όλα τα μέταλλα είναι **(καλοί) αγωγοί** της θερμότητας. Αυτό οφείλεται στη δομή των ατόμων τους. Στα άτομα των μετάλλων υπάρχουν κάποια ηλεκτρόνια που μπορούν εύκολα να μετακινούνται. Έτσι, μεταφέρεται η θερμότητα από το ένα μέρος του υλικού στο άλλο.

«Πρόσεξε, Αιλάν, θα καεί!» Η μεταλλική κουτάλα είναι **(καλός) αγωγός** της θερμότητας. Πάρε καλύτερα την ξύλινη, για να σε βίρσεις τη σούπα, είτε η μαμά.

Ο αέρας είναι **κακός αγωγός (μονωτής)** της θερμότητας. Γι' αυτό, στην κατασκευή των σπιτιών χρησιμοποιούμε υλικά που περιέχουν αέρα, όπως τα τούβλα και το φελλό.

Οι κούπες του καφέ έχουν χοντρά τοιχώματα που εμποδίζουν την **αγωγή** της θερμότητας. Έτσι, ο καφές διατηρείται ζεστός για περισσότερη ώρα.

Άμαξα, εισαγωγή, παραγωγή, στρατηγός, αγωνιάτης

Για να έχουμε νερό στα σπίτια μας στις πόλεις, το μεταφέρουμε με μεγάλους **αγωγούς** από τις πηγές των ποταμών.

Η Ιβάνα έχει πάρει καλή **αγωγή** από τους γονείς της και είναι πάντα ευγενική με τους ηλικιωμένους.

Ο αέρας είναι **κακός αγωγός** της θερμότητας. Πώς μπορεί τότε να ζεσταίνεται ένα δωμάτιο από το καλοριφέρ; Μήπως υπάρχει και άλλος τρόπος μετάδοσης της θερμότητας εκτός από την **αγωγή**? Βρες σχετικές πληροφορίες εδώ.

Αγγλική	Αραβική	Ρωσική	Τουρκική	Κινεζική
conductor	موصل حراري	проводник	kondüktör	导体

Figure 1. A dictionary page, the lemma *αγωγός* “conductor”.

The dictionary front matter contains a foreword (for the teachers but also for the students) and an explanation of labels/symbols used in the text. The Teacher’s Foreword is based on ELeFys’s Style Guide, the document that sets out in detail the way in which the dictionary entries are written. The

<sup>4</sup> As teachers, we consider such an arrangement to be helpful for the integration of a specialized dictionary into classroom practice.

<sup>5</sup> Of course, the boundary between terms and words is not always clear-cut; concepts that may have once been part of a highly specialized domain may filter down into our everyday lives, e.g. (*φυσικό*) *αέριο* “(natural) gas” (de-terminologization, Meyer & Mackintosh 2000 from Bowker 2003).

use of lexicographic symbols instead of abbreviations and metalanguage ensures the pedagogical role of ELeFyS; each symbol was selected on the grounds of its pictorial transparency, i.e. a penguin family is used to denote the word family notes and a two-finger hand gesture is used for multi-word compounds. All of the lexicographic symbols are explained immediately after the Teacher's and Student's Forewords.

Its digital format frees ELeFyS from the constraints of traditional printed dictionaries, such as space limitation and navigational difficulties. In terms of space, though, we do not plan to 'swamp' the user with lexicographic data just because we can (see also de Schryver 2003). On the contrary, following Lew (2012) we give special emphasis to the 'presentation space', by taking into account how much information the targeted user can process; thus, one of our main design principles was to maintain a double-column page-like layout for every lemma, in an effort to accomplish a holistic single-glance overview of the information. Navigation through the dictionary can be achieved in various ways. Primarily, it is facilitated by an interactive alphabet marker on the left side of the lemma page, graphically resembling a printed dictionary alphabet marker. Every letter is linked to the corresponding alphabetized index page at the end of the dictionary. Additionally, there are hypertext links which allow cross-references to other lemmas and navigation buttons for quick transition to neighboring pages or scientific subfields.

The interactive web-based format leads to further benefits, i.e. ease of access from anywhere through an internet connection inside and outside the school classroom, ease of navigation, multimedia content, such as animated illustrations and lemma pronunciation, around-the-clock debugging and updating, page printing, etc. From a technical point of view, ELeFyS was compiled on specialized software for creating multimedia e-books with built-in interactivity, offering the capability of publication to a number of different digital formats, like web-based (as in our case) or e-reader (epub3) format or mobile phone app.

So far the first edition of ELeFyS contains about 200 lemmas and sub-lemmas, a figure not to be 'sneezed at', considering its pedagogical dimension and specialization, and its intended user group and purpose. However, the compilation of an electronic dictionary is a dynamic process; thus, more lemmas are about to be added in the future and any potential technical problems will be resolved in the next edition.

### ***3.1.3 Entry Components and Lexicographic Information Distribution***

Within the broad scope of an entry, there are three principal components that carry additional information, related, though, to the main lemma: (1) sub-lemmas, (2) multi-word expressions and (3) run-ons. To start with, decisions were made concerning both the assignment of a main lemma status to the various items and the distinction between main lemmas and sub-lemmas. Both the entry components and distribution of lexicographic information analyzed below is portrayed in the template entry (Figure 2). Specifically, lexical items that are related to the entry either morphologically (derivatives, compounds or multi-word compounds) or semantically (hyponyms, meronyms) are entered as sub-lemmas under a particular headword. In any case, sub-lemmas are not granted their status arbitrarily, but they are directly associated with the main lemma (either a term or a general-use word). In such a way, multi-lemmas are formed combining interrelated linguistic information. For instance, the main entry *μαγνήτης* "magnet" includes the sub-lemmas *μαγνητικός* "magnetic" (derivative), *μαγνητικό πεδίο* "magnetic field" (multi-word compound), and *ηλεκτρομαγνήτης* "electromagnet" (compound), whereas the main entry *άτομο* "atom" includes the sub-lemma *πυρήνας* "nucleus" (meronym). Multi-word items are a central part of the Greek scientific and general-use language (Anastassiadis-Symeonidis 1986; Anastassiadis & Efthymiou 2007, Tantos et al. 2016). Despite their



fluid boundaries and the acknowledged difficulty in establishing robust criteria for their lemmatization (Cowie 1994; Mel'čuk 1998), multi-word items are given a specific treatment by being classified into two broad categories: (1) (semi)fixed phrases and (2) multi-word compounds.

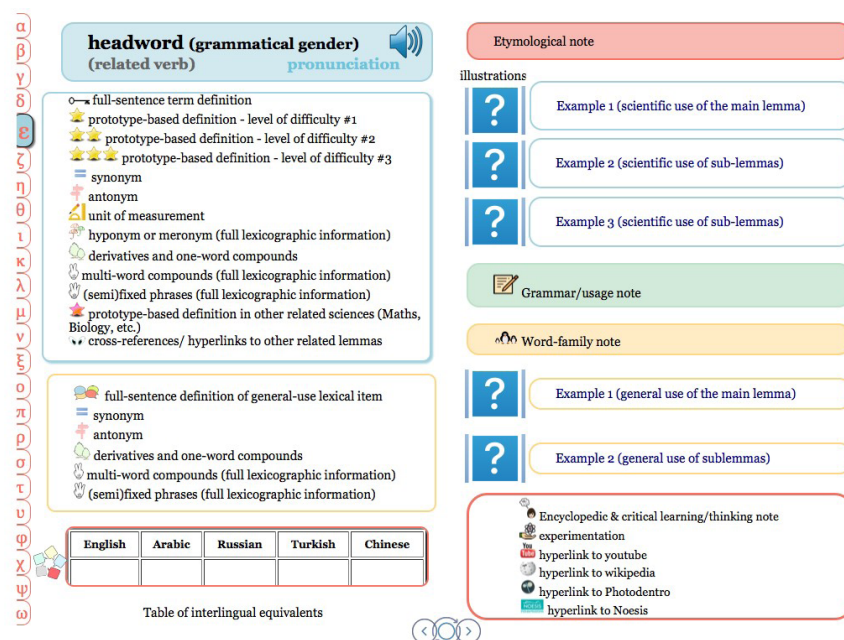


Figure 2. Representative template entry of ELeFyS.

(Semi)fixed phrases are in most cases collocations, such as *έπεσε η (ηλεκτρική)<sup>6</sup> ασφάλεια* “the (electric) fuse has blown” (terminology) or *γεμίζω τις μπαταρίες μου* “to recharge my batteries” (general use).

Multi-word compounds abide by the following lemmatization criteria: they are systematically entered as sub-lemmas under the main noun entry, which may constitute either the head or non-head of the compound, e.g. under the headword *μαγνήτης* “magnet” the multi-word compounds *φυσικοί* vs *τεχνητοί μαγνήτες* “natural vs artificial magnets”, under the headword *επαφή* “contact” the multi-word compound *φακοί επαφής* “contact lenses” (general use).<sup>7</sup> Such a decision has been made on the grounds that learners will benefit from the contrastive or combinatorial behavior of scientific and general-use lexical items. For instance, under the main lemma *ρεύμα* “current” they can find lexicographic information for electric current, river current and current as “opinion” or “tendency”, being thus exposed to the unifying and differentiating features that underlie the item’s senses, constructions and uses in scientific and everyday language. For all of the aforementioned entry components a full range of lexicographic information is supplied. In contrast, morphologically related derivative adjectives, e.g. the adjective *μοριακός* “molecular” under the main lemma *μόριο* “molecule”, are entered as run-ons, without any definition, but selectively used in examples (mostly those that are more frequent in school language corpora). Moreover, taking into account that word family information is meaningful as an indication of lexical richness and breadth (Laufer & Nation 1995), ELeFyS subsumes

6 It is very common for several scientific terms to be used both as multi-word compounds and as single-word nouns through the process of ellipsis, i.e. *ηλεκτρικό ρεύμα-ρεύμα* “electrical current”-“current”. Actually, in some cases the elliptical noun head form is more frequent than the compound; that is why the non-head part of the compound (adjective) is put in parentheses.

7 Only one deviation can be found from this systematic approach, that is in cases where abstract nouns construct multi-word compounds that fall within different subfields of Physics. For instance, the word *ενέργεια* “energy” constitutes the head of several multi-word compounds that fall within the subfield of Properties Matter and Atomic Structure, i.e. *ατομική ενέργεια* “atomic energy”, *πυρηνική ενέργεια* “nuclear energy” or the subfield of Heat & Temperature, i.e. *θερμική ενέργεια* “thermal energy”. In order to abide by the thematic arrangement of the dictionary, such multi-word compounds are treated as main lemmas, entered at the appropriate section of ELeFyS.



general-use words of the same family into word family notes, e.g. the word family note under the lemma *κύκλωμα* “circuit” includes items such as *κύκλος* “circle”, *κυκλώνω* “to circle”, *κυκλικός* “circular”, *ανακύκλωση* “recycling”.

As such, not only is a thorough understanding of the scientific concept obtained, but the learners’ consciousness on the productivity and polysemy of the Greek language is also raised.<sup>8</sup> Furthermore, since the chief macro-structural criterion must be user-friendliness, such a distribution facilitates the user’s search: the noun, i.e. the most frequent grammatical category (Anastassiadis-Symeonidis 1986) is the systematic search unit. Lastly, equal weight is given to all dictionary-relevant components, such as grammar, style and register, pragmatic features, relationships of synonymy and antonymy, etymology, etc., to help learners replace the apparent linguistic randomness with systematicity, and thus learnability (see Section 3.2.4). The orthographic conventions suggested by the *Dictionary of Standard Modern Greek* (1998) were used in ELeFyS.

## 3.2 Microstructure

### 3.2.1 Definitions & Senses

Definitions are fine-grained particularly in the case of polysemous words. Word senses of scientific terms are promoted to appear at the top left side of the entry in a light blue frame, whereas their corresponding senses in general vocabulary follow in a yellow frame. Besides conventional defining formulae (such as prototype and genus/differentiae-based definitions), contextual defining formats are used, such as full-sentence definitions (Rundell 2006), embedded in a rich microstructure. The scientific definitions are of gradual difficulty following a ranking from the simplest (suggested for a primary observation/understanding of the phenomenon) to the most complex (leading to academic wording). For instance, the main lemma (*ηλεκτρική επαφή* “(electric) contact” is firstly defined in a more pedagogical contextual format:<sup>9</sup>

- (1) *Ο διακόπτης μοιάζει με κουμπί που μπορείς να το ανεβάζεις ή να το κατεβάζεις με το δάχτυλό σου. Σε βοηθά να ανοίγεις και να κλείνεις μια ηλεκτρική συσκευή ή να ανάβεις και να σβήνεις το φως στο δωμάτιό σου.* “A switch looks like a button which you can push up or down with your fingers. It helps you to turn on/off an electric device or the lights in your room.”

Subsequently, more complex definitions are displayed in order to cover the different needs of learners in accordance with their age, cognitive state and linguistic competence, e.g. a definition of medium difficulty:

- (2) *μηχανισμός που ‘σταματά’ ή ‘ξεκινά’ τη σύνδεση σε ένα ηλεκτρικό κύκλωμα* “A device that ‘starts’ or ‘breaks’ the connection in an electric circuit.”

and a definition of great difficulty:

- (3) *στοιχείο ενός ηλεκτρικού κυκλώματος, με το οποίο μπορούμε να διακόπτουμε τη ροή του ηλεκτρικού ρεύματος* “A component of an electric circuit that interrupts the flow of electric current.”

Gradation of conventional definitions is flagged by one to three stars, so that difficulty in content or form can be marked. Of course, not every lemma exhibits the four suggested stages of gradual definitions; it depends on the conceptual difficulty or abstraction of the term, and the linguistic means

<sup>8</sup> Although we are aware of the counter-argument that sub-lemmas and run-ons are not favoured in pedagogical dictionaries, we proceeded with such a distribution of lexicographic information into multi-lemmas, thus prioritizing the learners’ needs to establish thematic and taxonomic conceptual relations (Mirman et al. 2017) in science and everyday life.

<sup>9</sup> Full-sentence scientific definitions placed in a complete microstructure are flagged by a key symbol.

available for its wording.<sup>10</sup> The same gradual pattern is also followed for the full-sentence definitions of the corresponding general-use words, this time from the literal to the figurative meaning.

We consulted several resources in order to choose the definition wordings that best fit the needs of our targeted user group, such as other specialized or general dictionaries (i.e. *Oxford Science Dictionaries*, *Cambridge and MacMillan Dictionaries*, *Dictionary of Standard Modern Greek*, etc.), Wikipedia, and the school textbooks. It should be mentioned, though, that since ELeFyS is multi-functional it does not aspire to catalogue senses in exhaustive detail; therefore, we opted for those general-use meanings that are highly frequent in the school textbook corpus. For instance, we omitted the meaning “tumor” from the lemma *όγκος* “volume” and the meaning “to have intercourse with somebody” for the fixed expression *έρχομαι σε επαφή με κάποιον*.

In specific cases, glosses in parentheses are used for a more informal explanation of the definition’s wording, in order to facilitate learners’ understanding, e.g. in the lemma *άτομο* “το μικρότερο συστατικό (=κομμάτι) της ύλης” “the smallest component (=piece) of matter” or in order to provide clarifying remarks, e.g. “οι ασφάλειες είναι διακόπτες που σταματούν αυτόματα τη ροή του ηλεκτρικού ρεύματος σε περίπτωση βλάβης (π.χ. βραχυκύκλωμα)” “electric fuses are switches that automatically interrupt the flow of electric current in cases of device breakdown (e.g. short circuit)”.

### 3.2.2 Examples

A broad spectrum of examples is offered in light blue (scientific terms) and yellow frames (general use) at the right side of the page, so that the lemma’s syntactic, collocational and pragmatic behavior is fully illustrated. Both authentic and lexicographer-made examples (Laufer 1992) are used, in order to reveal the words’ patterning and preferences; however, the pedagogical intent of ELeFyS is to provide meaningful examples, tailored to the communicative and academic needs of primary and secondary school children. Thus, in most cases we customized the authentic corpus-based examples by rewording or/and simplifying them.<sup>11</sup> The examples written for the main senses (scientific and general-use) of the polysemous Greek word *ασφάλεια*, “electric fuse” but also “safety/security”, are portrayed below:

(4) *Μόλις ο ηλεκτρολόγος μπήκε στο διαμέρισμα, βρήκε τον πίνακα με τις (ηλεκτρικές) ασφάλειες πίσω από την πόρτα. Κάποια από αυτές ήταν καμένη. Την άλλαξε και άναψαν τα φώτα.* “When the electrician entered the apartment, he found the switchboard behind the door. One of the fuses had blown. He changed it, and the lights went on.”

(5) *Η Ιβάνα δεν φοβάται τίποτα και νιώθει ασφάλεια στην αγκαλιά της μαμάς της.* “Ivana is not afraid of anything and she feels safe in her mother’s arms”

Pedagogy is also reinforced by the ‘continuous presence’ of four children-protagonists, i.e. Timos, Zoe, Ivana and Aylan, throughout the dictionary pages.

### 3.2.3 Interlingual Equivalents

Given its design principles, content, and form ELeFyS can be also appropriate for L2 learning. Towards this end, a table of interlingual equivalents in five languages (English, Arabic, Russian, Turkish

10 To ensure intelligibility, a ‘controlled defining vocabulary’ was used in the low-difficulty definitions, consisting of high-frequency words which the learner is expected to know sufficiently well. However, due to the conceptual complexity of special terms, definitions of great difficulty may include more demanding vocabulary.

11 It should be noted that in specific cases of hyponyms, meronyms or morphologically-related sub-lemmas, illustrative examples substituted definitions, in order to avoid dense information and cognitive burden from the definition section.

and Chinese) contributes to the thorough understanding of terminology.<sup>12</sup> These five languages were selected on the grounds of their criticalness at this specific socio-political juncture for Greece. At the same time, it helps students make interlingual and intercultural associations. In translating the terms, we consulted bilingual (electronic and print dictionaries).<sup>13</sup>

### 3.2.4 Grammatical, Etymological & Usage Notes

ELeFyS routinely provides micro-structural information on the form and use of each lemma. Word-formational indications are given as run-ons (derivative and compound words not lemmatized separately) and word family notes (see section 3.1.3), thus enabling the use of the word-part analysis (Oxford 2016) or morphological segmentation strategy (Anastassiadis-Symeonidis & Mitsiaki 2010) during the learning process. Additionally, etymological notes make a vital contribution to the new vocabulary's reception and use (see also Chatzisavvas 2005), as students benefit from the discovery of the word's history. For instance, they realize the etymological connection of electricity to the Ancient Greek *ἤλεκτρον* "amber", and get informed of the early observation of Ancient Greeks that amber exhibits electric properties. However, they also find out that the scientific term was formed much later in French, being based, however, on the Ancient Greek *ἤλεκτρον*. Therefore, they become familiar with the word's origin and the different historical paths through which terminology arose, while at the same time they detect the similarities between the first and second language in the case of internationalisms, e.g. *electricity* (English), *elektrik* (Turkish) *электричество* (Russian). In verifying the lemma etymology, we consulted the *Dictionary of Standard Modern Greek* (Petrounias 1998) as well as the digital version of the *Dictionary of Ancient Greek* (from the Center for the Greek Language website).

Furthermore, explicit grammatical and pragmatic guidance is offered through grammar and usage notes, which include information on the lemmas' structural features (see Section 2), i.e. spelling, pronunciation, inflection, syntax, +learned register, etc. Likewise, some of the headword's inherent grammatical properties are included, such as the grammatical gender (denoted by the definite article), the absence of plural form, lack of inflection for loanwords, related verbs, and so on. Finally, it should be noted that ELeFyS has dispensed with phonetic transcriptions, since learners are able to hear what a lexical unit sounds like.

### 3.2.5 Encyclopedic & Critical Learning/Thinking Notes

For every single article thought-provoking material is provided in the following forms:

- notes raising issues or questions that expand encyclopedic knowledge, e.g. the note included in the lemma *ηλεκτρισμός* "electricity" makes reference to the early observation made by Thales of Miletus who discovered that if he rubbed amber (*ἤλεκτρον*) with a piece of fur it could attract lightweight objects
- suggestions for experimentation that enhance critical thinking or/and intercultural sensitivity, e.g. the note included in the lemma *έλξη* "attraction" suggests the following experiment: "Rub a plastic pen on a wool sweater. Then put it near paper cut in small pieces. What do you observe?" or the note under the lemma (*καλός*) *αγωγός (θερμότητας)* "thermal conductor" makes reference to igloos, the double-walled ice shelters made by Eskimos, in order to prevent heat conduction
- hyperlinks to Wikipedia for a deeper understanding of physical phenomena and their history
- hyperlinks to videos in YouTube

12 Since the primary objective of such a specialized dictionary is to facilitate academic vocabulary/language learning, only the term interlingual equivalents were given.

13 At the moment, the translated terms are also checked by native speakers of the five aforementioned languages.

- hyperlinks to Digital Educational Resources from Photodentro – The Greek National Aggregator of Educational Content), and
- hyperlinks to multimedia available at Noesis (Thessaloniki Science Center & Technology Museum).

### 3.2.6 Pictorial Illustrations

Pictorial illustrations aid reception and production, thus complementing verbal explanations, leaving little room for misinterpretations, and promoting retention. The illustrations we have employed were selected from repositories that allow re-use and attribution under standardized licenses (Creative Commons). To serve the pedagogical role of ELeFyS, the illustrations were selected on the basis of their target-group age and cultural background (Ilson 1987; Biesaga 2016).

## 4 Conclusion

ELeFyS is an innovative, specialized Greek Science Dictionary intended for school use and an open educational resource that promotes learner autonomy through inquiry-based, strategy-based and cross-disciplinary learning. As a joint effort and a product of interdisciplinary collaboration between experts in the areas of applied linguistics and science education, it aims at bridging the gap of the parallel teaching/learning of science and language, greatly supporting the development of (meta) cognitive learning strategies. The dictionary micro- and macro- structure, the digital modality and the linguistic, conceptual or cultural stimuli provided render ELeFyS a valuable resource in the context of interactive learning in a school setting.

The compilation of a digital dictionary is a dynamic process, which means that it should be constantly revised and updated with new lemmas; therefore, our team welcomes feedback from academics, teachers and students. To this end, some small-scale pilot studies have already been conducted and partially reported (Mitsiaki & Lefkos 2017), providing positive feedback about the usability of ELeFyS. Additionally, a large-scale implementation is scheduled as the next step. Finally, in order to facilitate teachers and students in using such a dictionary creatively, ELeFyS will be complemented by a student's workbook, which is under development at the moment.

## References

- Anastassiadis-Symeonidis, A. (1997). Lexicography in Education (in Greek). In C. Tsolakis (Ed.), *Teaching of the Greek language* (pp. 149–176). Thessaloniki: Kodikas.
- Anastassiadis-Symeonidis, A. (1986). Neology in Standard Modern Greek (in Greek). Thessaloniki: AUTH.
- Anastassiadis-Symeonidis, A., & Efthymiou, A. (2007). Fixed phrases and teaching Modern Greek as a second language (in Greek). Athens: Patakis.
- Anastassiadis-Symeonidis, A., & Mitsiaki, M. (2010). Morphological segmentation: A strategy for teaching vocabulary in Greek as a second/foreign language (in Greek). In A. Psaltou-Joycey & M. Mattheoudakis (Eds.), *Proceedings of the 14th International Conference “Advances in Research on Language Acquisition and Teaching: Selected Papers* (pp. 65–77). Thessaloniki: Monochromia Publications. Accessed at: <http://www.enl.auth.gr/gala/14th/Papers/Greek%20papers/Anastasiadi-Symeonidi&Mitsiaki.pdf> [29/01/2018].
- Anastassiadis-Symeonidis, A., Vletsis, E., Mitsiaki, M., Oikonomou, S., & Aleksandri, K. (2014). Seeing the world from a different perspective: light, colours. A Science and Language teaching approach to foreign students (in Greek). In A. Psaltou-Joycey, E. Agathopoulou, & M. Mattheoudaki (Eds.), *Proceedings of the 15th International Conference of Applied Linguistics, “Cross-curricular Approaches to Language Education.”* Thessaloniki: Greek Applied Linguistics Association.



- Antypa, J., Efthymiou, A., & Mitsiaki, M. (2006). Mon Premier Dictionnaire Illustré: La rédaction d'un dictionnaire scolaire Grec. In *XII Euralex International Congress* (p. 383–393). Torino: Edizioni dell'Orso.
- Arapopoulou M., & Giannouloupoulou, G. (2001). Encyclopedic Guide. The use of language in non-linguistic school subjects: The scientific discourse (in Greek). Accessed at: [http://www.komvos.edu.gr/glwssa/odigos/thema\\_e7/e\\_6\\_thema.htm](http://www.komvos.edu.gr/glwssa/odigos/thema_e7/e_6_thema.htm) [10.11.2017].
- Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.
- Biesaga, M. (2016). Pictorial illustration in dictionaries. The state of theoretical art. In T. Margalitadze & G. Meladze (Eds.), *Proceedings of the XVII EURALEX International Congress* (pp. 99–108). Ivane Javakhishvili Tbilisi University Press.
- Bowker, L. (2003). Specialized lexicography and specialized dictionaries. In P. van Sterkenburg (Ed.), *A practical guide to lexicography, Terminology and Lexicography Research and Practice* (Vol. 6, pp. 154–164). Amsterdam, Philadelphia: John Benjamins Publishing.
- Chatzisavvas, K. (2005). The benefits of etymology in the vocabulary development of greek esl students. *ESL, Kean University, NJ, USA*.
- Cowie, A. P. (1994). Phraseology. The Encyclopedia of Language and Linguistics. In R. E. Asher & J. M. Simpson (Eds.) (pp. 3168–3171). Oxford: Pergamon Press. Reprinted in Fontenelle (2008).
- Cummins, J., & Yee-Fun, E. M. (2007). Academic Language. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 797–810). Boston, MA: Springer US.
- DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationships to science education reform. *Journal of Research in Science Teaching*, 37(6), 582–601.
- de Schryver, G. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2), 143–199.
- Dictionary of the Ancient Greek Language, Center for the Greek Language. Accessed at: [http://www.greek-language.gr/digitalResources/ancient\\_greek/tools/liddell-scott/index.html](http://www.greek-language.gr/digitalResources/ancient_greek/tools/liddell-scott/index.html) [30/01/2018].
- Dictionary of Standard Modern Greek (1998), Institute of Modern Greek Studies, Manolis Triandafyllidis Foundation, Thessaloniki.
- Dolezal, F. T., & McCreary, D. R. (1999). *Pedagogical lexicography today: a critical bibliography on learners' dictionaries with special emphasis on language learners and dictionary users*. Tübingen: Max Niemeyer Verlag.
- Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Buckingham: Open University Press.
- Gavrilidou, M., Giouli, V., & Labropoulou, P. (2008). The Greek High School Dictionary: Description and issues. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 515–524). Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.
- Geeraerts, D. (1990). The lexicographical treatment of prototypical polysemy. In S. L. Tsohatzidis (Ed.), *Meanings and Prototypes Studies in linguistic categorization* (pp. 195–210). London & New York: Routledge.
- Harlen, W., & Qualter, A. (2014). *The Teaching of Science in Primary Schools*. London: Routledge.
- Ilson, R. (1985). Illustrations in dictionaries. In A. Cowie (Ed.), *The dictionary and the language learner Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985* (pp. 193–212). Tübingen: Max Niemeyer Verlag.
- Institute of Modern Greek Studies, (Manolis Triantaphyllidis Foundation). (n.d.). Corpus of Spoken Greek (CSG). Accessed at: <http://corpus-ins.lit.auth.gr/corpus/index.html> [29/01/2018].
- Iordanidou, A., & Mantzari, E. (2004). Suggestions for the design of pedagogical dictionaries (in Greek). In *Proceedings of the 6th International Conference of Greek Linguistics* (pp. 1–12). Rethymno: University of Crete, Linguistics Lab. Accessed at: <http://www.philology.uoc.gr/conferences/6thICGL/ebook/f/iordanidou&mantzari.pdf> [03.01.2018].
- Julia, A., Efthymiou, A., & Mitsiaki, M. (2006). Mon Premier Dictionnaire Illustré: La rédaction d'un dictionnaire scolaire Grec. In *Proceedings of the XII Euralex International Congress* (p. 383–393). Torino: Edizioni dell'Orso.
- Lakoff, G. (1987). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*. (pp. 63–100). New York, NY, US: Cambridge University Press.
- Laufer, B. (1992). Corpus-based versus lexicographer examples in comprehension and production of new words. In J. S. Hannu Tammola, Krista Varantola, Tarja Salmi-Tolonen (Ed.), *EURALEX Congress* (pp. 71–76). Tampere, Finland: Studia Translatologia, University of Tampere.



- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322.
- Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex Publishing Corporation.
- Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 343–362). Oxford University Press.
- Marsh, D. (2002). CLIL/EMILE-The European dimension: Actions, trends and foresight potential.
- Mel'čuk, I. A. (1998). Collocations and lexical functions. In A. P. Cowie (Ed.), *Phraseology. Theory, analysis, and applications* (pp. 23–53). London: Oxford: Clarendon Press.
- Mel'čuk, I. A. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37–102). Amsterdam/Philadelphia: John Benjamins Publishing.
- Meyer, I., & Mackintosh, K. (2000). When terms move into our everyday lives: An overview of de-terminologization. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 6(1), 111–138.
- Mitsiaki, M., & Lefkos, I. (2017). ELeFyS - Pilot study of usage from Primary School Students (in Greek). In *Proceedings of the 1st National Conference on Didactics, Paedagogy And ICT*. Kavala: (to be published).
- Osborne, J. (2002). Science Without Literacy: A ship without a sail? *Cambridge Journal of Education*, 32(2), 203–218.
- Oxford Primary Illustrated Science Dictionary (Oxford Dictionary)*. (2013). OUP Oxford.
- Oxford Student's Science Dictionary (Oxford Dictionary)*. (2013). OUP Oxford.
- Oxford, R. L. (2016). *Teaching and Researching Language Learning Strategies: Self-Regulation in Context 2nd Edition (Applied Linguistics in Action)*. New York: Routledge.
- Plakitsi, K. (2010). Collective curriculum design as a tool for rethinking scientific literacy. *Cultural Studies of Science Education*, 5(3), 577–590.
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3), 328–350.
- Rundell, M. (2006). More than one way to skin a cat: Why full-sentence definitions have not been universally adopted. In E. Corino, C. Marelllo, & C. Onesti (Eds.), *Proceedings of the XII EURALEX International Congress* (pp. 323–337). Torino, Italy: Edizioni dell'Orso.
- Rundell, M. (2012). "It works in practice but will it work in theory?" The uneasy relationship between lexicography and matters theoretical. In *Proceedings of the 15th Euralex* (pp. 47–92).
- School Textbook Corpora, Portal for the Greek Language, Centre for the Greek Language. (n.d.) Accessed at [http://www.greek-language.gr/greekLang/modern\\_greek/tools/corpora/pi/content.html](http://www.greek-language.gr/greekLang/modern_greek/tools/corpora/pi/content.html) [29.01.2018].
- Shayer, M., & Adey, P. (1981). *Towards a science of science teaching: cognitive development and curriculum demand*. London; Exeter, N.H: Heinemann Educational Books.
- Tantos, A., Markantonatou, S., Anastassiadis-Symeonidis, A., & Kyriakopoulou, P. (2016). *Computational Linguistics* (in Greek). Athens:Hellenic Academic Libraries Link. Accessed at <http://hdl.handle.net/11419/2205> [20.11.2017].
- Tarp, S. (2005). The pedagogical dimension of the well-conveived specialised dictionary. *Ibérica*, 10, 7–21.
- Tarp, S., & Gouws, R. H. (2012). School dictionaries for first-language learners. *Lexikos*.
- Wellington, J., & Osborne, J. (2001). Language and Literacy in Science Education. In *McGraw-Hill Education (UK)* (pp. 1–8). London: Open University Press.

## Acknowledgements

We would like to thank Anna Anastassiadis-Symeonidis, Professor Emerita of Linguistics, AUTH, for her constant linguistic and lexicographic support and insightful comments.



# Terms Embraced by the General Public: How to Cope with Determinologization in the Dictionary?

**Jana Nová**

*Czech Language Institute of the Czech Academy of Sciences*

*E-mail: novaj@volny.cz*

## Abstract

The determinologization can be described as a process when specialized words (scientific terms) move into general vocabulary, followed by changes in their meaning. Czech and Slovak linguists have described two types or subsequent stages of determinologization. Firstly, the term is widely used outside scientific communication and it remains connected to the background scientific concept, but its meaning becomes less accurate in the use of laypeople. Secondly, the connection to the original concept is lost and a new figurative meaning of the word develops. The lexicographical approach to the language material reveals there is another, transitional type of determinologization where the word is used by laypeople in a blurred meaning, e.g. wider or narrower than if the term was used by domain specialists. The treatment of selected determinologized words in four dictionaries (Czech, Slovak and two English ones) is compared in this paper and various ways to mark the determinologized usage are presented, including a separate paragraph in the dictionary entry, an example sentence with an additional explanation or an additional note to the entry.

**Keywords:** Czech, determinologization, dictionary definition, general dictionary, lexicography, Slovak, terminology

## 1 Introduction

The accelerating development of science and technology since the beginning of the 20<sup>th</sup> century, the increasing amount of terminology and the growing interest of the public and mass media in scientific and technological matters (Cabr  1999: 4; Fontenelle 2014: 33; Becker 2016: 398) led to what we could call “terminologization” of general vocabulary. Terminology or widely specialized vocabulary – because in domains like computer technology or sport it is often difficult to distinguish between stabilized, “official” terminology and slang expressions – is the fastest growing part of the lexicon (Buz ssyov  2000: 513). Laypeople meet specialized words daily: seeing their doctor, reading newspapers and adverts, watching TV. Our personal vocabulary, at least its passive part (words we understand when we meet them), is being filled with specialized words. However, an inevitable effect of this process is a lesser or bigger change in the meaning of these words. This is what we call *determinologization*: a scientific term, during its way from a field specialist to a layperson, loses its accuracy, gets new connotations, and the word can be even moved to refer to a completely different thing.

General dictionaries must keep in touch with this development of lexicon and include more and more specialized headwords (or specialized meanings of words) from various domains in order to make their description of the contemporary vocabulary apt and to satisfy dictionary users who search for these words (cf. Sochov  1973: 196-198; Landau 1974: 241; Buz ssyov  2001: 19). Comparing a series of academic dictionaries of Czech since 1930, we found a steadily increasing amount of headwords from the field of medicine (M ourkov  et al. 2017), and we have no doubt we would find the same trend in most scientific, scholarly and technical domains. The lexicographic team of the currently compiled Academic Dictionary of Contemporary Czech (*Akademick  slovn k sou asn   e tiny*,

ASSC; Kochová & Opavská 2016) spends lots of time and effort on specialized entries. Determinologization is one of the issues we meet, cope and often struggle with in our work. This paper aims to present some practical solutions from the dictionary point of view rather than a thorough theoretical study on the topic.

The second part of the paper summarizes extremely briefly Czech and Slovak concepts of determinologization in comparison with the influential model presented in Meyer and Mackintosh (2000). Part three brings a draft classification of types/steps of determinologization, based on their impact on processing of the particular word in ASSC and compared with entries in general dictionaries of Slovak and English. All examples are taken primarily from Czech, and although I tried to select words belonging to the international vocabulary when possible, I am not sure whether their terminological or determinologized status in English is fully comparable to the situation in the Czech language material. My translations of Czech and Slovak expressions and dictionary entries into English are given in square brackets.

*Terminological note:* in the English literature the key term of this paper is often spelled *de-terminologization*. I spell it without the hyphen, following the tradition in Czech and Slovak literature and stressing the stabilized status of the term.

## 2 Linguistic Concepts of Determinologization

The concept of determinologization has been long established in Slovak and Czech linguistics. The term *determinologization* was used as early as in Horecký (1956: 36): “veľmi “odborné” termíny sa dostávajú do bežného používania a tým sa akosi “determinologizujú”. [very “specialized” terms get into common usage and therefore become somewhat “determinologized”.]” Several authors (Kuchař & Roudný 1965: 139; Sochová 1965: 199-200; Sochová 1973: 197) noted the move of terms from specialized communication to everyday communication and the general vocabulary of laypeople, together with stylistic and semantic changes of the words, leading sometimes to the establishing of a new separate meaning; this process was named *neutralization (of scientific terms)* by these authors. The term *determinologization* appeared again in Jedlička et al. (1970: 65-66) together with a clear description of the process: when a scientific term is used by laypeople, the meaning of the term loses its original accuracy and unambiguity, with words like *parameter*, *trend*, *model*, and *stereotype* given as examples.

More elaborated concepts of determinologization were presented by Poštolková (1977; 1980) and Horecký et al. (1989: 259-267; followed by Masár 1991: 150). Determinologization begins when a scientific term is widely used outside scientific communication, mostly in popularization and media texts, and this frequently happens to loanwords. While in its original scientific domain the term remains precisely defined, connected to the background scientific concept and it keeps systemic relationships to other terms in the domain, the word loses these attributes when used by laypeople. Its meaning becomes wider, less accurate; the word is newly connected into lexical relationships in the general vocabulary. As a subsequent stage of the determinologization process, a new separate meaning of the original term can be established, e.g. *climate* in natural science → *societal climate*, *spiritual climate*, meaning ‘milieu’. Poštolková (1984: 106) named the initial stage as *determinologization sensu lato*, the subsequent stage as *determinologization sensu stricto*.

Other authors mostly repeated the concepts of Poštolková and Horecký paraphrased above (e.g. Buzássyová 1983; Filipec & Čermák 1985; Bozděchová 2009). Holubová (2001) and Voborská (2001) introduced examples from more recent vocabulary. *Determinologization sensu lato* would be illustrated by popular words from medicine (*homeopathic*, *hospice*), economy (*sponsor*, *management*,

*leasing*), politics (*referendum*), ecology (*biotope*, *ozone hole*) and so on used frequently in the media and thus becoming familiar to laypeople. *Determinologization sensu stricto*, a rarer case where a new separate meaning of the term is developed or the word becomes a part of idioms, would be exemplified like this: *hyperinflation* → *hyperinflation of building*; *explosion* → *information explosion*; *blood group* → *to be of the same blood group as someone* (= to share someone's attitudes); etc.

Janovec (2007: 64) added the concept of *deslangization*, moving of slang expressions into general vocabulary, e.g. *dýza* [*disco*], *smažka* [*druggie*], *špek* [*joint of marijuana*], *čiro* [*mohawk hairstyle*]. Orgoňová and Bohunická (2011: 167-168) noted we should distinguish between *determinologization* as a spontaneous gradual language process and *transposition* as an intentional one-time pragmatic use of a scientific term in non-scientific communication in order to make a humorous effect, gain prestige, and so on.

The model of determinologization<sup>1</sup> in Meyer and Mackintosh (2000) is very similar to the model in Poštolková (1984) introduced above. Meyer and Mackintosh describe two types of determinologization:

- *retention of fundamental domain sense* (→ compare *determinologization sensu lato*) – the laypeople using the term refer to the same concept as the experts, but with shallower understanding. Connotations can be added to the meaning, e.g. “in the case of *peroxide*, laypersons tend to associate this chemical with hair coloring”;
- *dilution of original domain sense* (→ compare *determinologization sensu stricto*) – the word does not refer to the original scientific concept any more, or the laypeople do not intend to, although the linkage to the original domain is not lost completely; e.g. the medical term *anorexic* is used with the meaning ‘weak’ in collocations like *anorexic plot*, *anorexic dollar*. The authors note this type of determinologization leads to a new meaning of the word in general dictionaries.

To conclude the theoretical part, not only whole words but also parts of words can undergo determinologization. Dury (2008) focuses on prefixes *bio-* and *eco-* in English and French, widely used to form words with the meaning ‘environmental(ly friendly etc.)’; in Czech and Slovak the usage of these prefixes is comparable. Horecký et al. (1989: 267) highlight the suffix *-itida* [*-itis*], originally in medical terms like *hepatitida* [*hepatitis*], moved to form names of any exaggerated affection: *genitivitida*, *atestitida*. Holubová (2000) adds the example of *-mánie* [*-mania*] in words like *beatlemánie* [*beatlemania*].

### 3 Determinologization and the Dictionary

We will now see the situation from the dictionary point of view. Three remarkable types (or subsequent steps) of the determinologization process can be distinguished, each of them having different consequences for the lexicographical treatment of the affected words. I am bound to say first there are no strict borderlines between these three types, as there are hardly any strict borderlines between lexical categories of any kind. The larger and more varied the language material we study, the more likely we find all three types of determinologization for a given word, however it often takes thorough lexicographical consideration as to whether the blurred meaning (case B, 3.2) or the separate figurative meaning (case C, 3.3) is distinct and stabilized enough to be included in a dictionary.

Example dictionary entries are taken from the *ASSC, Dictionary of Contemporary Slovak Language* (*Slovník súčasného slovenského jazyka, SSSJ*), *Oxford English Dictionary* (*OED*) and *Merriam-Webster Dictionary* (*M-W*). The entries are transformed into a linear text; information about grammar and

<sup>1</sup> Cabré (1998) notes several times that specialized words can move into the common language, yet she does not use the term *determinologization*. Meyer and Mackintosh themselves quote Mazière (1981) and her term *dé-spécialisation*.



so on is left out, as it is not in the focus of this paper. The Czech examples from *ASSC* come from the manuscript of the dictionary, and the version cited here may not be the final one.

### 3.1 Determinologization A: Denotation Stays, Connotations Change

Type A, or *determinologization sensu lato* according to the model of Poštolková (1984), means using a specialized term outside its original domain with the background concept staying the same. Names of diseases like *arthrosis*, *diabetes*, and *encephalitis* used by doctors or their patients, names of space objects like *comet*, *galaxy*, and *white dwarf* used by astronomers or visitors to an observatory, paleontological terms like *ice age* or *secondary era* used in fiction would always refer to the same denotation. However, there is a difference in the depth of understanding: the layperson's view of the concept is always simplified and often focused on different aspects of the reality compared to the view of a field specialist. When an observatory visitor pictures a *white dwarf* like 'a small star shining white', it is true but not the most important thing for an astronomer, who would rather speak of the final stage of a star's life and extremely high density of the electron-degenerated matter. If the popular image of an *ice age* is 'a cold period with glaciers all around' then it is not very accurate, as colder and warmer periods alternated during every glacial period, and there were no glaciers in the Czech inland, for example. The names of chemical substances often have negative connotations for laypeople: going through the newspaper part of corpora when analyzing headwords like *benzene* or *biphenyl*, the general impression would be 'something chemical, smelly and toxic' (why was it ever legal to produce something that horrible?); on the other hand, a *vitamin* is always 'great for your health, eat it more' although an overdose of vitamin A, for example, can damage your health permanently.

Most terminological entries in *ASSC* represent this type; according to the conception of our dictionary (Kochová & Opavská 2016: 177; see also Machač 1964: 67), we do not include headwords which did not move from specialized communication to general communication (the decision in particular cases is based on the proportion of specialized and non-specialized sources for the word in the language corpus): strictly said, all our terminological headwords are determinologized *sensu lato*. Their definitions are based on the original scientific definition but adapted for laypeople, which means stressing the most important features relevant both from the scientific and laypeople's point of view. When the popular image of something is too biased, we strive to be more objective. For example, with the chemical substances mentioned above, the *ASSC* entry would include information like 'smelly' and 'toxic' when it is true, but it would also explain what the substance was or is used for; if it would be too long or too detailed for the definition, we include this kind of information into the exemplification part of the entry.

*Botulin* will be our dictionary example. Before the current era of botox-smooth faces, this chemical compound was known to laypeople as a toxin present in rotten canned food; the Slovak dictionary, published in 2006, expressed it well, as follows:

- (1) *SSSJ*: *botulotoxín* [...] chem., lek. ► prudko jedovatá látka vytváraná mikróboom *Clostridium botulinum* v pokazených al. nedokonale sterilizovaných potravinách, klobásový jed: *otrava botulotoxínom*; *liečba botulotoxínom* [chemistry, medicine ► severely toxic substance produced by the microbe *Clostridium botulinum* in rotten or insufficiently sterilized food, the sausage toxin: *poisoning by botulin*; *treatment by botulin*]
- (2) *OED*: *botulin* [...] The bacterial toxin involved in botulism.

The British definition in (2) is very brief, what is in fact uncommon for the *OED* when treating scientific terms. However, the *OED*'s example sentences (visible only after clicking on the "Example sentences" button) include collocations like "*botulin* toxin is not officially permitted for plastic surgery" and "the treatment, which involves an injection of *botulin* toxin type A in the forehead to smooth

out lines”, indicating using the substance in plastic surgery. Note that the Slovak entry (1) also only mentioned the medicinal use in the exemplification part.

The American dictionary *M-W* gives more information (and more practical information) to the user:

- (3) *M-W: botulinum toxin* [...] a neurotoxin formed by botulinum that causes botulism and that is injected in a purified form for therapeutic and cosmetic purposes (as to treat blepharospasm and reduce wrinkles)

Both example sentences in *M-W* refer to the medicinal use of the substance.

Finally, the Czech dictionary *ASSC* presents *botulin* like this:

- (4) *ASSC: botulin* [...] chem., biol., farmac. ► prudce jedovatá organická sloučenina (polypeptid) produkovaná bakterií *Clostridium botulinum*, způsobující ochrnutí svalů, užívaná též jako léčivo [...]: otrava botulinem; biologické zbraně s použitím botulinu; botulin je obsažený ve zkažených uzenářských výrobcích; nechal si aplikovat botulin proti vráskám [chemistry, biology, pharmacy ► severely toxic organic compound (a polypeptide) produced by the bacterium *Clostridium botulinum*, causing muscle paralysis, used also as a medicine [...]: poisoning by botulin; biological weapons using botulin; botulin is present in rotten smoked meat products; he got applied botulin against wrinkles]

The *ASSC* solution is probably most complex of all four variants compared here. It includes a chemical classification (“organic compound (a polypeptide)”), both to be just for scientists and to neutralize the definition somewhat, because an expression like “severe toxin used also as a medicine” would be too alarmist for ordinary dictionary users in our opinion. The effect of the toxin is described briefly in the definition, while information of the typical occurrence in rotten sausages is included in the exemplification part – or we could do it the other way round, to keep the definition brief.

The amount and complexity of information are usually rather high in the *OED*’s and *M-W*’s terminological entries, see further examples (10, 17, 19). Czech and Slovak dictionaries strive to balance the scientific point of view (verified information, keeping the scientific system of classification, etc.) with the needs of the dictionary users who are laypeople (what it is, where I can find it, what its practical importance is). Using domain labels is also a rather central-European tradition (Nová & Mžourková 2017), and a way to indicate “this word with this described meaning belongs to specialized communication”, while dictionaries of English only rarely label headwords for domains.

### 3.2 Determinologization B: Blurring of the Meaning

Type B forms a transition between types A and C. The background concept of the term is changed in the popular usage; the word would not (or not always) refer to the same denotation as in the specialized domain. There is a remarkable change of the meaning, but it is not yet enough for a new separate meaning to be determined and lexicographically described, or there is a wide overlap between the original (accurate, scientific) meaning and the new (determinologized) one, and it is often difficult to decide over a particular context where it belongs.

Most often the original meaning of the term is blurred<sup>2</sup> in the popular usage, the word refers to a wider range of similar denotations where the field specialist would not use the same word. We could name it *generalization of the meaning*, in other words. For example, *depression* in psychiatry or psychology is a serious mental condition with a set of characteristic symptoms, often accompanying

2 Meyer and Mackintosh (2000) use the word *dilution*; however, their statement “when laypersons use words in this category, they do not intend to designate the original domain sense” and their examples do not match the delimitation of determinologization type B in this paper, rather following type C (see 3.3), where the new meaning is clearly distinguishable.

mental disorders such as the bipolar disorder, while laypeople often use the word to name any (isolated and not so serious) episode of feeling sad. *Bakelite* was the name of the first commercially produced synthetic plastic, and it became so famous the word is sometimes used to name other plastics or even plastics in general; thus the popular statement “a Trabant (a car from former East Germany) is made of bakelite” although it is in fact made of duroplast. Names – originally trademarks – of well-known medicines are often used to name similar medicines as well; for example *brufen* can refer to any painkiller in Czech and *penicillin* to any antibiotic.

Another case consists in laypeople using the term with a narrower meaning, in other words *specification of the meaning*. In chemical nomenclature, words like *acetate* (in Czech *acetát* for organics or *octan* for inorganics) refer to a group of chemical compounds with similar structure, not just to a particular compound. However, common-speech collocations “ušít šaty z acetátu” [“make a dress of acetate”] and “dát na vymknutý kotník obklad s octanem” [“treat a sprained ankle with an acetate compress”] do not refer to *any acetate* but to a fabric of *cellulose acetate* (Czech: *acetát celulózy*) in the dress case and to *aluminium acetate* (Czech: *octan hlinitý*) in the compress case, respectively. Czech biological nomenclature is binomial as well, so when biologists speak of *bolševník* [hogweed], they refer to the whole *Heracleum* genus or any species in the genus – while for laypeople, *bolševník* is particularly *bolševník velkolepý* (*Heracleum mantegazzianum*, *giant hogweed*) and many do not know there even exist any other *bolševník* species than this infamous pest. Another biological term, *anabolic* – ‘constructing molecules from smaller units in metabolism’, underwent specification due to the popularity of bodybuilding and drugs called *anabolic steroids*, thus for laypeople *anabolic* means ‘building muscles’.

Sometimes the scientific concept changed in time, therefore the scientific usage of the term changed too, but laypeople still use the word in its dated meaning.<sup>3</sup> Czech word *bacil* is popularly used to name any bacterium, although the scientific genus *Bacillus* is now much narrower; the word is even used to name any microbes causing diseases in collocations like “chřipkový bacil”, “bacil neštovic” although chřipka [influenza] and plané neštovice [smallpox] are caused by viruses.

Determinologization type B is less frequent than types A and C, but treating these words is trickiest for lexicographers. Primarily, it is difficult to recognize whether the concrete use of the word in the sentences taken from fiction or media discourse is still (relatively) terminological or blurred. Many examples can belong to either: for collocations like “a telephone made of bakelite” or “there grows hogweed on the meadow” we cannot tell for sure unless examining the particular telephone or the plants on the particular meadow, and that is clearly no task for lexicographers. Then it is hard to decide whether the determinologized meaning is clear enough to make a separate paragraph within the dictionary entry, and how to formulate the definition to distinguish it clearly from the original terminological meaning.

The treatment of *bakelite* in selected dictionaries follows, sorted from the most to the least terminological approach:

- (5) *OED*: *Bakelite* [...] trademark An early form of brittle plastic, typically dark brown, made from formaldehyde and phenol, used chiefly for electrical equipment.
- (6) *SSSJ*: *bakelit* [...] chem. ► látka zo syntetickej živice používaná v elektrotechnike, staviteľstve a pod. [...]: výrobky z bakelitu; *b. patril k prvým materiálom, ktoré sa dali tvarovať* [chemistry ► substance of synthetic bitumen used in electrical engineering, construction, etc. [...]: *products of bakelite*; *b. was among the first materials that could be shaped*]

3 Compare thus with ten Hacken's (2010: 924) example: “technological advances leading to the discovery of more objects in the solar system have led to a blurring of the concept of *planet* followed by a tightening of the definition.” For laypeople Pluto is still a *planet*, although astronomers determined a special box called *dwarf planets* for this object and similar ones.

- (7) *ASSC: bakelit* [...] ► plast odolný tepelně a chemicky, používaný k výrobě různých předmětů, v elektrotechnice, stavitelství ap., syntetický polymer: *lisovna bakelitu; telefon, zásuvky, hračky z bakelitu; bakelit byl prvním uměle vytvořeným plastem* [► chemical-proof and thermo-resistant plastic used to make various items, in electrical engineering, construction etc., a synthetic polymer: *press shop for bakelite; telephone, sockets, toys of bakelite; bakelite was the first synthetically produced plastic*]
- (8) *M-W: Bakelite* – used for any of various synthetic resins and plastics

The *OED* (5) treats the word in the narrowest possible sense, as a trademark. For *SSSJ* (6) it is a chemical term as indicated by the domain label, but the definition is rather vague; this is not surprising, as plastics are extremely alike for laypeople when we do not want to describe their chemical composition (as the *OED* did to some extent in (5)). The non-specific example collocation “výrobky z bakelitu” [“products of bakelite”] can be applied to any plastic as well. Although we are not sure whether the Slovak lexicographers intended this, we did in *ASSC* (7). We were aware the word *bakelite* is used by laypeople to refer to any plastic, quite like *M-W* expresses in (8), registering only the determinologized usage of the word. However we did not have enough example collocations saying clearly “this is not the true *bakelite* but another kind of plastic”, so our solution does not say expressly “the word can refer to a variety of substances”. Still, the vague wording of the definition and most examples, applicable both to the “true bakelite” and other plastics, together with the absence of the *chemical* label, indicate the determinologized usage of the word.

With regard to the derived adjective *bakelitový*, we used the example with a Trabant mentioned above and stressed both in the definition and in the exemplification the word is used in a determinologized way:

- (9) *ASSC: bakelitový* [...] 1. ► vyrobený z bakelitu • vyrobený z plastu vůbec: *bakelitový vypínač, starý černý bakelitový telefon; bývalý východoněmecký symbol – bakelitový trabant z duroplastu* [...] [1. ► made of bakelite • made of plastics in general: *bakelite switch, old black bakelite telephone; former symbol of East Germany – a bakelite Trabant of duroplast* [...]]

Another dictionary example will be *hogweed*. The *M-W* solution will not be cited, as apparently in American English the word *hogweed* does not refer to the genus *Heracleum* at all; on the other hand, the situations in British English, Czech and Slovak are comparable:

- (10) *OED: hogweed* [...] A large white-flowered weed of the parsley family, native to north temperate regions and formerly used as forage for pigs. – Genus *Heracleum*, family Umbelliferae: several species, in particular the common European *H. sphondylium* and the introduced giant hogweed (*H. mantegazzianum*)
- (11) *SSSJ: bolševník* [...] ► vysoká rozkonárená bylina z čeláde mrkvovitých s dutou stonkou, s velkými tmavozelenými listami a drobnými kvetmi v okolíkoch, rostúca na vlhkých okrajoch lúk a popri potokoch; bot. *b. obrovský* *Heracleum mantegazzianum* zavlečený do Európy z Ázie ako okrasná rastlina, dnes jeden z najnepríjemnejších invázných druhov rastlín, ktorý obsahom fototoxických látok môže poškodiť zdravie človeka [► tall ramose herb of the parsley family with hollow stem, with large dark-green leaves and tiny flowers in umbels, growing in wet meadow margins and at streams; botanically *b. obrovský* = *giant hogweed* *Heracleum mantegazzianum* introduced to Europe from Asia as an ornamental plant, today one of the worst invasive plant species, which can damage human health due to its containing phototoxic substances]
- (12) *ASSC: bolševník* [...] 1. ► vysoká a mohutná rastlina s dutým stonkem, velkými listy a květenstvími drobných bílých květů, původem z Asie, u nás se nekontrolovatelně šíří, bot. bolševník velkolepý *Heracleum mantegazzianum*: *porosty bolševníku; invaze, přemnožení bolševníku; bojovat s agresivním bolševníkem; děti se ošklivě popálily bolševníkem; bolševník se do Čech*



*dostal jako okrasná rostlina* [1. ► tall and large plant with hollow stem, large leaves and inflorescences of tiny white flowers, native to Asia, in our country spread out of control, botanically *bolševník velkolepý* = giant hogweed *Heracleum mantegazzianum*: *vegetation of hogweed; invasion, outbreak of hogweed; fight against the aggressive hogweed; children got seriously burnt by hogweed; hogweed came to Bohemia as an ornamental plant*] 2. ► rostlina s různě dělenými listy a květenstvími drobných bílých květů, bot. rod *Heracleum*: *bolševník je rod z čeledi miříkovitých; všechny bolševníky jsou jedovaté; bot. bolševník obecný Heracleum sphondylium, bolševník perský Heracleum persicum* [2. ► plant with its leaves parted in various ways and with inflorescences of tiny white flowers, botanically genus *Heracleum*: *hogweed is a genus in the parsley family; all hogweeds are poisonous; botanically bolševník obecný = common hogweed, Heracleum sphondylium, bolševník perský = Persian hogweed, Heracleum persicum*]

The British dictionary entry (10) describes whole genus *Heracleum*; however three example sentences out of five (visible only after a click) refer to the *giant hogweed* in collocations like “hogweed which is very dangerous” and “to eradicate hogweed”, and the two remaining examples contain the collocation *giant hogweed* expressly. The Slovak solution (11) describes the genus *Heracleum* first, too, giving the scientific name *bolševník obrovský* as the only example, actually a collocation with its own definition.

Processing the Czech language material for ASSC and finding nearly all occurrences of the word *bolševník* refer to *bolševník velkolepý*, we originally intended to describe only this species and present the entry *bolševník* in the same way as what is now paragraph 1 in (12) – it would be a completely determinologized solution, only working with the laypeople’s point of view and omitting the botanical concept of the term. Nevertheless, not only biologists but also other linguists recommended including the information there are also other species of hogweed than just the notorious giant hogweed; the purely botanical paragraph 2 in (12) is the result, describing the genus *Heracleum* and giving examples of several species. Uncertainty remains whether this solution is not too confusing for dictionary users without biological knowledge, whether they would not read the entry as “describing the same twice”. Let this be an example of the lexicographer’s hard life, trying to be just to the language material, to the scientific concepts of the terms, and to those dictionary users who are laypeople at the same time.

### 3.3 Determinologization C: New Meaning Develops

Type C means developing a new, clearly separate meaning from the original term; the word is transferred to a completely different denotation, and on seeing an example sentence there is usually no doubt whether it belongs to the original terminological meaning or to the determinologized figurative one. The description of *determinologization sensu stricto* by Poštolková (1984) fits well for this case.

The determinologized meaning is based on the metaphor or metonymy. For example, *atmosphere* – in geology ‘a layer of gases surrounding a planet’, determinologized ‘mood of a place or a group of people’. *Allergy* – in medicine ‘hypersensitivity of the immune system to something in the environment’, determinologized ‘extreme disliking for something or someone’. *Neon* – in chemistry ‘a chemical element, among other properties giving bright red light under voltage’, determinologized ‘a shining tube of any color used as an advertising sign’. Journalists and politicians like to animate their utterances with borrowings from the scientific language, so we can read about “the government being in *agony*”, “*anaemic* economic growth” (compare Meyer’s and Mackintosh’s example *anorexic dollar*), “taking new political *azimuth*” or “the national debt reaching *astronomic* heights”. Party leaders describe themselves as *alpha males* (in zoology *alpha* is the leading individual in a wolf pack, group of primates, etc.); a new person is a *comet* of the scene, and so on.



It is usually no problem for lexicographers to distinguish and describe the determinologized figurative meaning. However, the decision must be made as to whether the new meaning is already stabilized – lexicalized – to get its own paragraph within the dictionary entry; this depends on there being a sufficient amount of determinologized examples (clear enough to serve as a dictionary example), their occurring in various sources and over a significant time period. When these conditions are not satisfied, in both the *ASSC* and *SSSJ* one or two determinologized examples would be put at the end of the exemplification part of the terminological entry, labelled as “přen.” / “pren.” (přenesený = figurative) or “expr.” (expressive) and with a short additional explanation. See the example of *celibacy* for *SSSJ*:

- (13) *SSSJ: celibát* [...] ► povinné zrieknutie se manželstva rímskokatolíckych duchovných na základe cirkevného sľubu, bezženstvo: *kňazský c.; rehoľný, kláštorň c.; sľub celibátu*; pren. expr. *žiť v dobrovolnom celibáte bez sexuálnych vzťahov* [► obligatory refraining from marriage for Roman Catholic clergy according to a clerical vow, the state without a wife: *priestly c.; monastic c., c. of friars; the vow of celibacy*; figurative, expressive *to live in self-imposed celibacy without sexual relationships*]

Now the treatment of the adjective *antiseptic* in the four dictionaries will be compared. Note that while the Slovak dictionary (14) probably found no notable determinologized usage and the Czech one (15) only gives one figurative example, in English the determinologized usage is clearly common enough to establish separate paragraphs in a dictionary, the structure of the *M-W* entry (17) being particularly rich:

- (14) *SSSJ: antiseptický* [...] ► súvisiaci s antiseptikom; majúci účinok antiseptík: *antiseptické látky; a. obväz; antiseptické vlastnosti mydla; vypláchnuť ranu antiseptickým roztokom* [► relating to antiseptics; having the effect of antiseptics: *antiseptic substances; a. bandage; antiseptic properties of the soap; to wash a wound with an antiseptic solution*]
- (15) *ASSC: antiseptický* [...] ► pôsobící, účinný proti choroboplodným mikroorganismům na povrchu kůže, sliznic, tkání • ničící tyto mikroorganismy: *antiseptický roztok, sprej, antiseptická ústní voda; slaná mořská voda má antiseptický účinek*; přen. *v domečku vládl antiseptický pořádek* naprostý, až nadměrný [► effective against germs on the surface of the skin, mucous membranes, tissues • destroying these germs: *antiseptic solution, spray, antiseptic mouthwash; salt sea water has an antiseptic effect*; figurative *there was an antiseptic tidiness in the house* total, even excessive]
- (16) *OED: antiseptic* [...] 1 Preventing the growth of disease-causing microorganisms. ‘garlic has powerful antiseptic properties’; ‘his breath smelt of antiseptic mouthwash’  
2 Scrupulously clean or pure, especially so as to be bland or characterless. ‘their squeaky-clean home epitomizes this antiseptic respectability’
- (17) *M-W: antiseptic* [...] 1 a: opposing sepsis, putrefaction, or decay; *especially*: preventing or arresting the growth of microorganisms (as on living tissue) • an *antiseptic* solution  
b: acting or protecting like an antiseptic • an *antiseptic* mouthwash  
2: relating to or characterized by the use of antiseptics • *antiseptic* treatment  
3 a: scrupulously clean: ASEPtic • *antiseptic* surgical instruments  
b: extremely neat or orderly; *especially*: neat to the point of being bare or uninteresting • a spare, *antiseptic* waiting room  
c: free from what is held to be contaminating • an *antiseptic* version of rustic life  
4 a: coldly impersonal • an *antiseptic* greeting  
b: of, relating to, or being warfare conducted with cold precision from a safe distance with few or no casualties on one’s side • *antiseptic* bombings

The last example will be *adrenaline*. The Slovak entry in (18) (the structure of the entry in *ASSC* is identical so I do not cite it) illustrates well all stages of determinologization described so far in this

paper: a term known to laypeople but retaining its terminological status in paragraph 1, the blurred meaning in paragraph 2, the clearly figurative meaning in paragraph 3 along with two idioms:

- (18) SSSJ: *adrenalin* [...] 1. biol. ► hormón drene nadobličiek, ktorý podporuje činnosť srdca, zvyšuje krvný tlak a jeho koncentrácia stúpa pri strese (pri stavoch úzkosti al. záťaže): *hladina adrenalínu; do krvi sa vyplavuje a.* [1. biology ► a hormone from the medulla of the adrenal glands which stimulates heart action, increases blood pressure and its concentration rises under stress (in the state of anxiety or strain): *the level of adrenaline; a. is released into the blood*] 2. hovor. ► stav napätia, vzrušenia; silné emócie: *príchut' adrenalínu; vybit' si a.; film plný adrenalínu; vzduch nabitý adrenalinom* [2. colloquial ► the state of suspense, excitement; strong emotions: *the taste of adrenaline; vent one's a.; a film full of adrenaline; the air charged with adrenaline*] 3. hovor. ► športová aktivita al. činnosť vyvolávajúca napätie a silné emócie účastníkov: *vyznávači, milovníci adrenalínu; užít' si trochu adrenalínu; život bez adrenalínu* [3. colloquial ► a sport or another activity causing suspense and strong emotions of the participants: *lovers, devotees of adrenaline; enjoy some adrenaline; life without adrenaline*] ■ fraz. *dvíhať/zvyšovať adrenalin niekomu* a) vyvolávať emócie, vzrušenie; b) rozčulovať niekoho; *stúpa mu adrenalin* a) je vzrušený, rozrušený; b) je rozčúlený [■ idioms *to raise someone's adrenaline* a) to cause emotions, suspense; b) to annoy someone; *his adrenaline rises* a) he is excited, upset; b) he is annoyed]
- (19) OED: *adrenaline* [...] A hormone secreted by the adrenal glands that increases rates of blood circulation, breathing, and carbohydrate metabolism and prepares muscles for exertion. '*performing live really gets your adrenaline going*'
- (20) M-W: *adrenaline* → EPINEPHRINE. NOTE: *Adrenaline* is used in both technical and nontechnical contexts. It is commonly used in describing the physiological symptoms (such as increased heart rate and respiration) that occur as part of the body's fight-or-flight response to stress, as when someone is in a dangerous, frightening, or highly competitive situation, as well as the feelings of heightened energy, excitement, strength, and alertness associated with those symptoms. In figurative use, it suggests a drug that provides something with a jolt of useful energy and stimulation.

The British version in (19) looks like a purely terminological one, but the example sentence would actually fit for the blurred meaning 2 in the Slovak version in (18) in my opinion; examining more examples visible after a click, collocations like "*charged with adrenaline, I took several deep breaths*" and "*I felt a huge rush of adrenaline*" are of the same kind. The American version in (20) links to a deeply technical description of *epinephrine* but the note about "nontechnical contexts" is remarkable and illustrated by the example sentence "The emerging young voices and their supporters must remain steadfast beyond the current rage, fear, and *adrenaline* of the most recent massacre."

## 4 Conclusion

As for the lexicographically relevant cases of determinologization described in part 3, type A (retaining the scientific background concept of the term) fits well with *determinologization sensu lato* according to Poštolková (1984) or *retention of fundamental domain sense* in the Meyer and Mackintosh model, and type C with *determinologization sensu stricto* or *dilution of the domain sense* where the denotation changed and a new separate meaning or an idiom develops. Type B, blurring of the original meaning, seemingly has not yet been described in theory. Lexicographers are aware that something like this exists, most notably in M-W's examples (8) and (20), while the OED in the same cases holds to a terminological description, although example collocations show a shift in the meaning. It is

often difficult to describe this shift in a dictionary, and it is very difficult to exemplify it clearly. When Meyer and Mackintosh (2000) describe the second stage of determinologization, say type C in this paper, as “when laypersons use words in this category, they do not intend to designate the original domain sense”, we could say for type B “laypersons don’t know whether they intend to designate the original sense because they cannot tell it apart”, not knowing the difference between *depression* and a passing sadness, *acetate* and *cellulose acetate*, and so on, and so it is not easy for lexicographers to look at the related language material and work out what the layperson was actually referring to.

We can see various ways to process the determinologized usage of a term in a dictionary: an intentionally vague definition to cover both the terminological and the determinologized meaning (7), a separate segment of the definition (9), a separate part of the exemplification with an additional explanation (11, 13, 15), a separate meaning-paragraph in the entry (12, 16, 17, 18), an additional note to the entry (20), or such a note instead of the definition itself (8). It is difficult to say which way is the best without a larger and more detailed analysis of determinologized dictionary entries, and there is probably no universal way to treat determinologized words, but many of them need a special approach. In my opinion, any way is valid as long as the determinologization is registered in the dictionary. Explanatory notes such as in (20) are probably most useful for dictionary users who are not always able to decipher the lexicographical way of cutting an entry into paragraphs and sub-paragraphs, using or omitting domain labels and so on, or who can easily fail to notice a figurative example at the end of an exemplification. On the other hand, a general dictionary is no encyclopedia, and the use of additional notes should not be excessive.

As for polysemous entries, it remains an open question how to sort the paragraphs: should the original scientific meaning be first despite the frequency, as in (18) where meanings 2 and 3 are definitely more common outside the scientific literature than meaning 1? Or should the determinologized meaning be first when it is more frequent, as in (12), to help the dictionary user who is probably searching for this one (cf. Lopukhina et al. 2016)?

Example dictionary entries throughout this paper also showed the importance of the exemplification in terminological entries (cf. Taljard 2016), and not only for polysemous entries where the delimitation of meanings must be based on the evidence of the language material. The approach of the *OED*, often giving no example sentence as a fixed part of the entry and only displaying examples after a click, does not seem the best one to me – examination of those additional examples can show a remarkable inconsistency between the examples and the definition of the term, as in the cases of *hogweed* and *adrenaline*. The policy of *ASSC* is to exemplify all one-word headwords including all specialized words (Kochová & Opavská 2016: 156), even though it takes a lot of time to find good examples.

## References

- ASSC: *Akademický slovník současné češtiny*, manuscript.
- Becker, H. (2016). Scientific and technical dictionaries; coverage of scientific and technical terms in general dictionaries. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 393-407.
- Bozděchová, I. (2009). *Současná terminologie (se zaměřením na kolokační termíny z lékařství)*. Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum.
- Buzássyová, K. (1983). Dynamika v odbornej terminológii. In *Jazykovedný časopis* 34, pp. 132-144.
- Buzássyová, K. (2000). Odborná lexika vo všeobecnom výkladovom slovníku. In: K. Buzássyová (ed.) *Človek a jeho jazyk. Na počesť prof. Jána Horeckého*. Bratislava: Veda, pp. 513-523.
- Buzássyová, K. (2001). Z koncepcnej a realizačnej problematiky nového výkladového Slovníka súčasného slovenského jazyka. In *Lexicographica* 99, Bratislava: Veda, pp. 13-23.
- Cabré, M.T. (1999). *Terminology. Theory, Methods and Applications*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

- Dury, P. (2008). The rise of carbon neutral and compensation carbone. A diachronic investigation into the migration of vocabulary from the language of ecology to newspaper language and vice versa. In *Terminology* 14, pp. 230-248.
- Filipec, J., Čermák, F. (1985). *Česká lexikologie*. Praha: Academia.
- Fontenelle, T. (2014). From Lexicography to Terminology: a Cline, not a Dichotomy. In A. Abel, Ch. Vettori et N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 25-45.
- Holubová, V. (2001). K pojetí determinologizace. In M. Žemlička (ed.) *Termina 2000, Sborník příspěvků z II. konference 1996 a III. konference 2000*. Praha a Liberec: Galén a Technická univerzita v Liberci, pp. 157-160.
- Horecký, J. (1956). *Základy slovenskej terminológie*. Bratislava: Vydavateľstvo Slovenskej akadémie vied.
- Horecký, J., Buzássyová, K., Bosák, J. et al. (1989). *Dynamika slovnej zásoby súčasnej slovenčiny*. Bratislava: Veda.
- Janovec, L. (2007). K projevům jazykových vývojových tendencí v současné češtině. In *Naše řeč* 90, pp. 57-66.
- Jedlička, A., Formánková, V., Rejmánková, M. (1970). *Základy české stylistiky*. Praha: Státní pedagogické nakladatelství.
- Kochová, P., Opavská, Z. (eds.) (2016). *Kapitoly z koncepcie Akademického slovníku současné češtiny*. Praha: Ústav pro jazyk český AV ČR, v. v. i.
- Kuchař, J., Roudný, M. (1965). České odborné názvosloví v uplynulém dvacetiletí. In *Naše řeč* 48, pp. 133-144.
- Landau, S. (1974). Of Matters Lexicographical: Scientific and Technical Entries in American Dictionaries. In *American Speech* 49, pp. 241-244.
- Lopukhina, A., Lopukhin, K., Iomdin, B., Nosyrev, G. (2016). The Taming of the Polysemy: Automated Word Sense Frequency Estimation for Lexicographic Purposes. In T. Margalidze, G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 249-256.
- Machač, J. (1964). Odborná terminologie ve výkladovém slovníku. In *Československý terminologický časopis* 3, pp. 65-76.
- Masár, I. (1991). *Príručka slovenskej terminológie*. Bratislava: VEDA.
- Mazière, F. (1981). Le dictionnaire et les termes. In *Cahiers de lexicologie* 39, pp. 79-104.
- Meyer, I., Mackintosh, K. (2000). When Terms Move into Our Everyday Lives: An Overview of De-terminologization. In *Terminology* 6, pp. 111-138.
- M-W: *Merriam-Webster Dictionary*. Accessed at: <https://www.merriam-webster.com/dictionary> [19/03/2018]
- Mžourková, H., Nová, J., Pernicová, H. (2017). Proměny lékařské terminologie v jednojazyčných výkladových slovnících. In *Naše řeč* 100, pp. 215-224.
- Nová, J., Mžourková, H. (2017). Terminology and Labelling Words by Subject in Monolingual Dictionaries – What Do Domain labels Say to Dictionary Users? In *Jazykovedný časopis* 68, pp. 296-304.
- Orgoňová, O., Bohunická, A. (2011). *Lexikológia slovenčiny*. Bratislava: Univerzita Komenského.
- OED: *Oxford English Dictionary*. Accessed at: <https://en.oxforddictionaries.com> [19/03/2018]
- Poštolková, B. (1977). K vlivu odborné terminologie na národní jazyk. In *Slovo a slovesnost* 38, pp. 112-120.
- Poštolková, B. (1980). K specifičnosti významu termínů. In *Slovo a slovesnost* 41, pp. 54-57.
- Poštolková, B. (1984). *Odborná a běžná slovní zásoba současné češtiny*. Praha: Academia.
- Sochová, Z. (1965). Některé novinky v současné slovní zásobě. Odborné termíny jako zdroj rozhojňování neterminologických slovních vrstev. *Naše řeč* 44, pp. 199-205.
- Sochová, Z. (1973). K otázce univerzálního heslového standardu jako podkladu pro dvojjazyčné slovníky. In *Slovo a slovník*, Bratislava: Vydavateľstvo Slovenskej akadémie vied, pp. 193-200.
- SSSJ: *Slovník současného slovenského jazyka*. Accessed at: <http://slovníky.juls.savba.sk> [19/03/2018]
- Taljad, E. (2016). Collocational Information for Terminological Purposes. In T. Margalidze, G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi State University, p. 533-560.
- ten Hacken, P. (2010). The Tension between Definition and Reality in Terminology. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress. 6-10 July 2010*. Leeuwarden/Ljouwert: Fryskke Akademy – Afûk, pp. 915-927.
- Voborská, M. (2001). Termíny v publicistických textech. In M. Žemlička (ed.) *Termina 2000, Sborník příspěvků z II. konference 1996 a III. konference 2000*. Praha a Liberec: Galén a Technická univerzita v Liberci, pp. 170-173.

# **Reports on Lexicographical Projects**





# Thesaurus of Modern Slovene: By the Community for the Community

**Špela Arhar Holdt<sup>1,2</sup>, Jaka Čibej<sup>1,2,3</sup>, Kaja Dobrovoljc<sup>1,3</sup>, Polona Gantar<sup>2</sup>, Vojko Gorjanc<sup>2</sup>, Bojan Klemenc<sup>1</sup>, Iztok Kosem<sup>2,3</sup>, Simon Krek<sup>2,3</sup>, Cyprian Laskowski<sup>2</sup>, Marko Robnik-Šikonja<sup>1</sup>**

<sup>1</sup>Faculty of Computer and Information Science, University of Ljubljana, <sup>2</sup>Faculty of Arts, University of Ljubljana, <sup>3</sup>”Jožef Stefan” Institute

E-mail: [spela.arharholdt@ff.uni-lj.si](mailto:spela.arharholdt@ff.uni-lj.si), [jaka.cibej@ff.uni-lj.si](mailto:jaka.cibej@ff.uni-lj.si), [kaja.dobrovoljc@ijs.si](mailto:kaja.dobrovoljc@ijs.si), [apolonija.gantar@guest.arnes.si](mailto:apolonija.gantar@guest.arnes.si), [vojko.gorjanc@ff.uni-lj.si](mailto:vojko.gorjanc@ff.uni-lj.si), [bojan.klemenc@fri.uni-lj.si](mailto:bojan.klemenc@fri.uni-lj.si), [iztok.kosem@ff.uni-lj.si](mailto:iztok.kosem@ff.uni-lj.si), [simon.krek@ijs.si](mailto:simon.krek@ijs.si), [cyprianadam.laskowski@ff.uni-lj.si](mailto:cyprianadam.laskowski@ff.uni-lj.si), [marko.robnik@fri.uni-lj.si](mailto:marko.robnik@fri.uni-lj.si)

## Abstract

By presenting the Thesaurus of Modern Slovene, the largest open-access collection of Slovene synonyms, this paper describes the concept of a responsive dictionary, a dictionary that allows its data to continuously respond to the changes in language and the feedback from the language community. We begin by briefly summarizing the method of its construction and its technical aspects. A great deal of deliberation and work has been put into interface design, with the aim to make the Thesaurus as user-friendly as possible for all digital media. This is followed by a more detailed description of the types of user input (e.g. synonym suggestions, synonym votes) and feedback (interface improvement suggestions) collected as part of development, as well as the methodology for their implementation. We also touch upon a series of dissemination activities aimed specifically at community building and user involvement. In conclusion, we describe our plans for the future, such as updates to be implemented in version 1.1 of the Thesaurus.

**Keywords:** responsive dictionary, digital lexicography, community, crowdsourcing, thesaurus, Slovene

## 1 Thesaurus of Modern Slovene

The recently published Thesaurus of Modern Slovene (<http://viri.cjvt.si/sopomenke/>) was created at the Centre for Language Resources and Technologies of the University of Ljubljana<sup>1</sup> as part of an effort to establish an infrastructure for Slovene that is comparable to the infrastructures of larger languages. In its current version, the Thesaurus contains 105,473 headwords and 368,117 synonyms, making it the largest automatically generated open-access collection of Slovene synonyms. The database was developed using innovative computational approaches for optimizing data reusability and connectivity. The Thesaurus is based on a range of different language resources (Krek et al. 2017) and enables users to compare synonyms in their collocational context, as well as check their use in the Gigafida reference corpus of modern Slovene.<sup>2</sup> In terms of financial resources, the computer-assisted data preparation was significantly less demanding and more economical than manual processing, as well as significantly less time-consuming. This enables regular updates and upgrades of the resource, making the dictionary a dynamically evolving source of language information. Furthermore, the Thesaurus facilitates a number of ways to involve the user community into the database construction process. The newly developed platform allows users to evaluate entries through voting and add

<sup>1</sup> <https://www.cjvt.si/>

<sup>2</sup> <http://www.gigafida.net/>

suggestions for new synonyms. User activities in the interface are tracked and user suggestions are incorporated into regular updates. The database of the Thesaurus is openly available at the Clarin.si repository under the Creative Commons Attribution-ShareAlike 4.0 International licence.<sup>3</sup>

### 1.1 Creation of the Thesaurus Database

A detailed description of the methodology used in the preparation of the Thesaurus database is presented in (Krek et al. 2017). Here, we summarise the basic idea of the method. The Thesaurus combines language data from two existing reference resources: The Oxford<sup>®</sup>-DZS Comprehensive English-Slovenian Dictionary (Šorli et. at 2006) and the Gigafida reference corpus of written Slovene (Logar and Krek 2012). Both resources comprise authentic language material published after 1991 and offer a solid basis for a description of modern Slovene. The extraction of synonym candidates was based on the manner in which words co-occurred in translation strings of the Oxford<sup>®</sup>-DZS Dictionary (e.g. *abandon* > *zapustiti*; *opustiti*; *odpovedati se*, *odstopiti od*). This information, together with the frequency of word co-occurrences, was the basis for discrimination between ‘core’ and ‘near’ synonyms, with ‘core’ synonyms exhibiting a greater degree of connection to the keyword. The identified links between synonyms were additionally confirmed using the older Dictionary of Standard Slovenian Language (SSKJ). The method formed balanced co-occurrence graphs and used the Personal PageRank algorithm (Page et al 1999) to automatically divide the synonyms into subgroups and rank them according to the degree of semantic relatedness to the keyword. Co-occurrence graphs were used to organize synonyms in the dictionary. The information on the frequency of keywords and synonym candidates in language use was also included in the database.

Expert evaluations (see Figure 1, based on the data from Krek et al. 2017) showed that, although not perfectly accurate, the presented method produced results of sufficient quality. The evaluation involved three linguists who independently rated automatically extracted synonyms from a set of 50 randomly chosen headwords from different part-of-speech categories (with 550 candidates in total). Possible ratings were good, acceptable, and poor. It should be noted that the boundaries of synonymy are difficult to define and heavily depend on the context and the circumstances of language use. Thus, even for expert evaluators, rating the data was not trivial, as seen from data on agreement in Figure 1. However, the percentage of synonyms with at least one positive score was 76.4%, which indicates that it is likely that some dictionary users would be interested in seeing them in the dictionary.



The evaluation has shown that the presence of the majority of synonyms that were rated as poor can be

<sup>3</sup> <https://www.clarin.si/repository/xmlui/handle/11356/1166>

attributed to the methodology used and the content of the initial Oxford<sup>®</sup>-DZS database. In particular, these included literal repetitions within the phrase (*analiza* ‘analysis’ – *podrobna analiza* ‘detailed analysis’, *matematična analiza* ‘mathematical analysis’), multi-word paraphrases similar to the headword (*povratno* ‘reflexively’ – *v refleksivni obliki* ‘in the reflexive form’), masculine-feminine pairs (*nagajivka* ‘mischievous girl’ – *nagajivec* ‘mischievous boy’), sense groups corresponding to the headword but not mutually overlapping in terms of meaning (*krutost* ‘cruelty’ – *nesramnost* ‘rudeness’, *strogost*, ‘strictness’, *neprijaznost* ‘unfriendliness’), and words present in the Oxford<sup>®</sup>-DZS database but rarely used in Slovene (*grotesken* ‘grotesque’ – *hogarthovski* ‘Hogarthian’).

The current database is organized in a self-contained MySQL database, centered on the headword and synonym data, and enriched with corpus-derived collocations and examples (see Section 1.2). It currently contains 105,473 headwords, 368,117 synonyms, 3,353,061 collocations and 2,505,472 examples. It also stores and updates user contributions (synonym suggestions and votes, see Section 3). The Thesaurus interface (see Section 1.3) interacts with the database through a REST API, written in Python and Flask.

## 1.2 Inclusion of Collocations and Corpus Examples

The Thesaurus provides the possibility to explore synonymy in context with the use of corpus data. An important novelty for Slovene language resources in this regard is the option to compare the use of different synonyms with the help of their typical collocates (see Figure 4, which represents the collocations page of the Thesaurus, where e.g. the adjectives *pozitiven* ‘positive’ and *brezskrben* ‘carefree’ can be compared in context through their typical collocates: *pozitivna energija* ‘positive energy’ and *brezskrbne počitnice* ‘carefree vacation’). The information for the comparison was obtained by using the Sketch Difference function in the Sketch Engine tool (Kilgariff et al. 2004).

For each part-of-speech category, a selection of grammatical patterns was made that can be used when comparing synonyms through collocations. The selection was made in line with the preparation of the Collocations Dictionary of Modern Slovene (Kosem et al., in print). For nouns, the patterns are *adjective + noun* (*študijski program* ‘study program’, *študijski načrt* ‘study plan’), *noun + preposition + noun* (*program za prihodnost* ‘program for the future’, *načrt za prihodnost* ‘plan for the future’), *verb + preposition + noun* ( *vključiti v program* ‘include in the program’,  *vključiti v načrt* ‘include in the plan’), and *verb + noun* (*predstaviti program* ‘present a program’, *predstaviti načrt* ‘present a plan’). For verbs, the patterns are *verb + preposition + noun* (*presenetiti pri dejanju* ‘surprise in the act’, *zasačiti pri dejanju* ‘catch in the act’), *verb + noun* (*presenetiti vlomilca* ‘surprise an intruder’, *zasačiti vlomilca* ‘catch an intruder in the act’), *adverb + verb* (*ponovno presenetiti* ‘surprise again’, *ponovno zasačiti* ‘catch in the act again’). For adjectives, the patterns are *adjective + noun* (*pozitivna ocena* ‘positive evaluation’, *spodbudna ocena* ‘encouraging evaluation’), *adjective + preposition + noun* (*pozitiven za gospodarstvo* ‘positive for the economy’, *spodbuden za gospodarstvo* ‘stimulating for the economy’), and *adverb + adjective* (*zelo pozitiven* ‘very positive’, *zelo spodbuden* ‘very encouraging’). For adverbs, the patterns are *adverb + verb* (*odločno zanihati* ‘firmly deny’, *ostro zanihati* ‘strongly deny’), *adverb + adjective* (*odločno zavržen* ‘firmly rejected’, *ostro zavržen* ‘strongly rejected’), and *adverb + preposition + noun* (*odločno proti predlogu* ‘firmly against the suggestion’, *ostro proti predlogu* ‘strongly against the suggestion’).

In addition, examples of use were imported into the dictionary using computational methods for the automatic recognition of good (dictionary) examples (Kilgariff et al. 2008, Kosem 2017). Collocations and examples of use are included in most entries, and all the entries also contain links to the concordances of a particular collocation in the Gigafida corpus, which provide information on the use of the collocation in different contexts. Additional contextual information is provided by domain labels

(e.g. *biologija* ‘biology’), which were added from the Oxford-DZS Comprehensive English-Slovenian Dictionary and help explain the context of use for individual synonyms.

### 1.3 Dictionary Interface Design

In recent years, the Slovene digital lexicography has focused on the systematic collection of empirical data on the habits, wishes and needs of Slovene language users, as well as the possibilities for user involvement in the creation of language resources, primarily through crowdsourcing (Arhar Holdt et al. 2017, Arhar Holdt et al. 2016, Čibej et al. 2016, Gorjanc et al. (ed.) 2017), Čibej et al. 2015). The results of these studies were the basis for the design of the Thesaurus interface. In addition, feedback from future evaluations (possibly collected and analysed through a semi-automatic approach) will be used for further improvement. The interface was developed over a period of approximately one year, along with the concept of the responsive dictionary and the visual identity of the resource. When designing the interface, we took into account the following principles: speed and ease of use (e.g. search bar auto-complete, user-friendly data display, minimal number of clicks), clarity (a light, uncrowded display without unnecessary elements), step-by-step navigation, and motivation for user participation (e.g. Hall of Fame, see Section 2).

### 1.4 Responsive Dictionary as a Concept

With the Thesaurus of Modern Slovene, we are introducing a new type of dictionary called the *responsive dictionary*, so named because it enables the data to continuously respond both to changes in language and the feedback from the language community. Furthermore, it is defined by the following features:

- From the first, the responsive dictionary is developed specifically for the digital medium.
- The initial dictionary database is constructed using advanced computational methods, instantly providing the language community with a large quantity of relevant, albeit still somewhat noisy language information.
- The dictionary interface provides a number of ways for the community to contribute to the expansion of the database and clean up noisy elements.
- The development of a responsive dictionary is never concluded as its data constantly evolves in response to the developments in modern language. The changes are tracked using timestamps in individual entries, while the different versions of the database are stored in a dedicated archive.
- Dictionary development follows a clearly defined methodology for including user contributions and implementing improvements based on information systematically collected from user activities in the dictionary.
- The dictionary and its database are openly accessible under an appropriate license (e.g. CC-BY-SA 4.0).

## 2 Community Building and Inclusion

The Thesaurus of Modern Slovene provides a number of different ways for user involvement. While user involvement is not a new concept in lexicography, our approach differs from collaborative lexicographic practices (Abel and Meyer 2013) that anticipate editorial control of user-added content (e.g. Macmillan Open Dictionary, Leo, Openthesaurus.de) in order to prevent the addition of noisy data (Cristea et al. 2014), or editorial interventions that implement user feedback after the dictionary is already completed (De Schryver and Prinsloo 2000). The responsive dictionary integrates user contributions directly into the process of creating and maintaining the dictionary, while potential



editorial decisions can be taken in consensus/comparison with the decisions of language users. In the dictionary, users can evaluate synonyms and directly add their own suggestions. Technical support for automatically importing existing synonym collections into the Thesaurus is also provided. In addition, users can participate in a public debate on dictionary features, and also have the opportunity to join a group of dedicated user evaluators, i.e. community members that are invited to participate in user evaluations of new versions of the Thesaurus. All these activities are described in more detail in subsections 2.1-2.3.

However, user participation cannot be expected without sufficient user motivation. In the current version of the interface, this challenge is tackled by incorporating motivational elements used in crowdsourcing projects, e.g. the Hall of Fame (a list of users who added the most synonyms) and the Tasks-of-the-Day section (a daily random list of headwords with no user votes; see Figure 2). As for community building, a website, a Facebook profile and a newsletter were created in order to inform the public about the Thesaurus. In addition, a separate project<sup>4</sup> aimed at organizing events for the promotion of the Thesaurus is being carried out in 2018-2019 and will hopefully boost the number of participants in the initial stage of the process as well as provide answers to several open issues: what type of user motivation is the most effective (and what group does it influence the most), how do promotional events affect the number of dictionary users, etc.

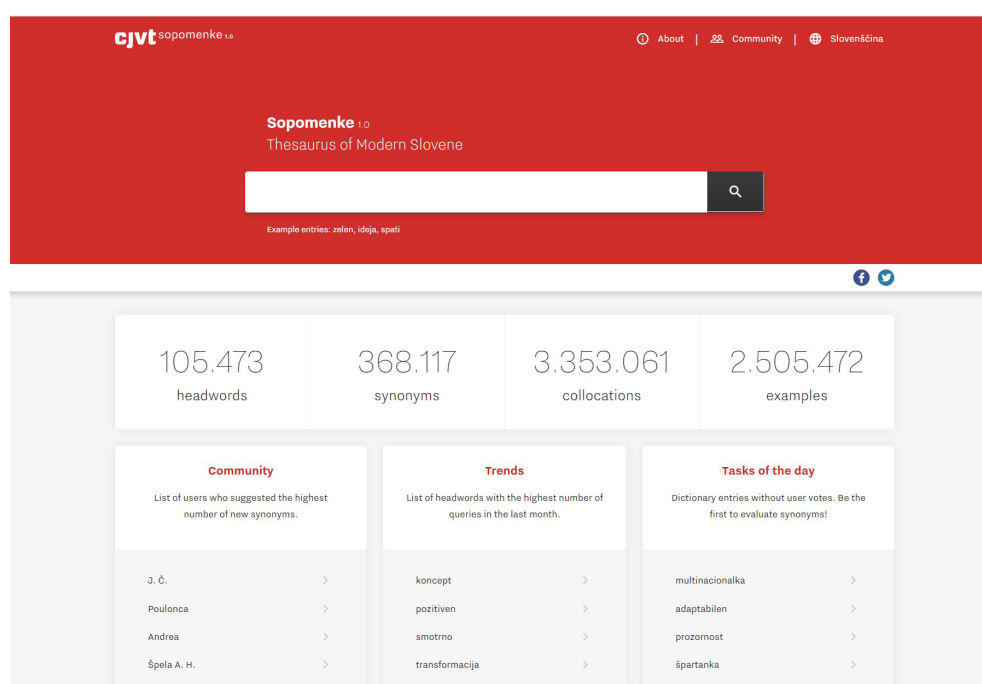


Figure 2: Front page of the Thesaurus of Modern Slovene, with sections for most-active users, trends and tasks of the day.

## 2.1 Synonym Evaluation

As can be seen from Figure 3 showing the synonyms of *pozitiven* ('positive'), the synonym candidates listed under individual thesaurus headwords can be rated as good (appropriate, useful) or poor (noisy, incorrect). In Figure 3, the synonym *pohvalen* ('commendable') has been rated as appropriate, as indicated by the green voting button (displaying a value of 1). Users also have the option to filter the candidates by user rating. With dictionary updates, the ratings will be taken into account in order

<sup>4</sup> <https://www.cjvt.si/promocija-sopomenk/>

to re-arrange or remove data accordingly (e.g. if the users predominantly identify a synonym as inadequate, it can be removed from the entry). User evaluation is not just limited to user-added synonyms but extends to synonyms included in the initial database as well. This allows the community not only to evaluate whether user-added suggestions are valid and constructive, but also to identify potential noisy candidates that are present in the database because of automatic database generation.

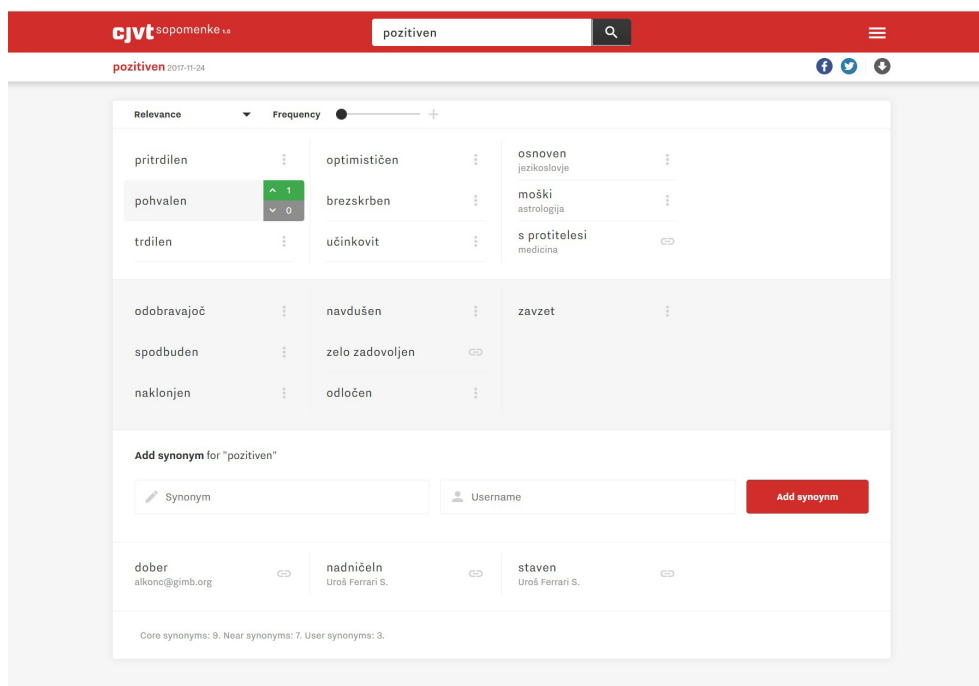


Figure 3: Synonym page for *pozitiven* ('positive'), with separate sections for core, near and user synonyms.

## 2.2 Addition of New Synonyms

Each headword contains a list of core and/or near synonyms (core synonyms are more closely linked to the headword compared to near synonyms; for more detail, see Krek et al. 2017). At the bottom of each entry, a third section is provided for user-added synonyms (Figure 3). This section consists of a form in which users enter their synonym suggestion and user name (no registration is required.). Suggestions are displayed in the entry immediately after submission. As already mentioned, technical support is also provided to automatically import existing synonym collections into the Thesaurus. However, in order to avoid malicious additions and bot entries, the process of adding synonyms is not without restrictions. Once a user has entered a synonym suggestion, the form is substituted by a link that needs to be clicked in order to reopen it. In addition, a number of user activities in the dictionary are logged (more on this in Section 3), which provides enough metadata (e.g. time of submission, IP-address) to remove all contributions from a user that would be identified as malicious. The addition of new synonyms is motivated with the Hall of Fame (a list of users who added the most synonyms), while constructive suggestions are encouraged with the list of best-rated user suggestions.

## 2.3 Discussion of Existing Solutions

Users that want to provide direct feedback on the interface can either send an e-mail to the development team or participate in a discussion that takes place in a dedicated public Facebook group, where users can share links to individual thesaurus entries, ask questions on dictionary data or related language problems, suggest improvements, etc. For instance, since the official publication of the

Thesaurus in March 2018, users have expressed a number of specific needs that should be addressed. First, the users would like to add labels to provide additional contextual information on the suggested synonyms (as shown e.g. by the user suggestion *arcnije (pogovorno)*, ‘*arcnije* (colloquial)’ added to the headword *zdravilo* ‘cure’). Second, in addition to adding synonyms to existing entries, the users would like to add new (or missing) headwords to the dictionary. Third, the users would like to have the option to evaluate user suggestions in one place. In the current version of the interface, user suggestions can only be evaluated within their respective headwords and there is no concise overview of user suggestions (either an overview by individual users or in general). User opinions will be taken into account when preparing thesaurus updates.

### 3 User Data Collection and Dictionary Upgrades

A vast array of data on user activities in the Thesaurus are recorded for further use. User activities are defined as events of different types, as presented in Table 1. For every event, a timestamp and the IP of the user are recorded. In this manner, a sequence of events in a specific session can be reconstructed and potentially malicious entries identified.

Table 1: Data on User Activities.

Event Type	Description	Source Pages
static_page_visit	When a user visits a specific static page of the interface (e.g. the pages with information on the Thesaurus), the log records the visit.	main_page, about_page, community_page, versions_page, impressum_page, 404_page
link_click	The log records when a user clicks on a link that leads to an external page (e.g. the Thesaurus database in the Clarin.si repository).	URL (string)
headword_select	When a user enters a headword, the log records: (I) the entered headword; (II) where in the interface the action took place (from a search window on a specific subpage or by clicking on the example entry links, Tasks of the Day; Trends, or links from Facebook and Twitter).	synonyms_page, collocations_page, main_page, main_page_example, main_page_random, main_page_popular, facebook, twitter
headword_not_found	When a user searches for a headword missing in the database, the log records the provided search string.	headword (string)
synonym_select	When a user selects one of the synonyms on the synonyms page, the log records: (I) the selected synonym (II) the headword under which the synonym is listed; (III) where in the interface the selection took place (e.g. on the synonyms page; from the side menu on the collocations page; from the side menu by using the buttons for previous/next synonym).	synonyms_page, collocations_page, collocations_page_btn_next, collocations_page_btn_prev
collocation_select	When a user clicks a collocate to access the corresponding corpus examples, the log records the set of collocational data the collocate is from.	collocations_page
synonym_go_to_gigafida	When a user clicks the button with the link to the Gigafida corpus from the synonyms page, the log records: (I) the selected synonym, (II) the synonym’s headword; (III) the URL of the link.	synonyms_page

Event Type	Description	Source Pages
example_go_to_gigafida	When a user clicks the button with the link to the Gigafida corpus from the section with corpus examples, the log records: (I) the set of collocational data the collocate is from, and (II) the URL of the link.	collocations_page
synonym_suggest	When a user adds a synonym, the log records: (I) the added synonym and (II) the synonym's headword; (III) the provided username; (IV) whether the entry was successful or not (e.g. if the proposed synonym already existed in the database, which prompts a warning).	synonyms_page
headword_download	When a user downloads entry data, the log records: (I) the current headword; (II) the subpage from which the download button was clicked.	synonyms_page, collocations_page
synonyms_frequency_slider_change	When a user filters synonyms by frequency, the log records: (I) the selected headword, (II) the subpage where the filtering took place, and (III) the type of slider change (left or right).	synonyms_page, collocations_page
synonyms_sort	When a user sorts synonyms according to their features, the log records: (I) the selected headword, (II) the subpage where the sorting took place, and (III) the selected type of sorting (alphabetical, by synonym length, etc.)	synonyms_page, collocations_page
synonym_vote	When a user votes on a synonym, the log records: (I) the synonym the user is voting on and (II) the synonym's headword, (III) whether the vote was added or removed, (IV) the value of the vote (+1 or -1).	synonyms_page, collocations_page
language_change	The log records when a user changes the language of the interface (Slovene to English or vice versa).	language (slv, eng)
share	When a user shares the link to the dictionary content on social media, the log records: (I) the shared URL (II) the source medium (Facebook or Twitter)	medium (Facebook, Twitter), URL

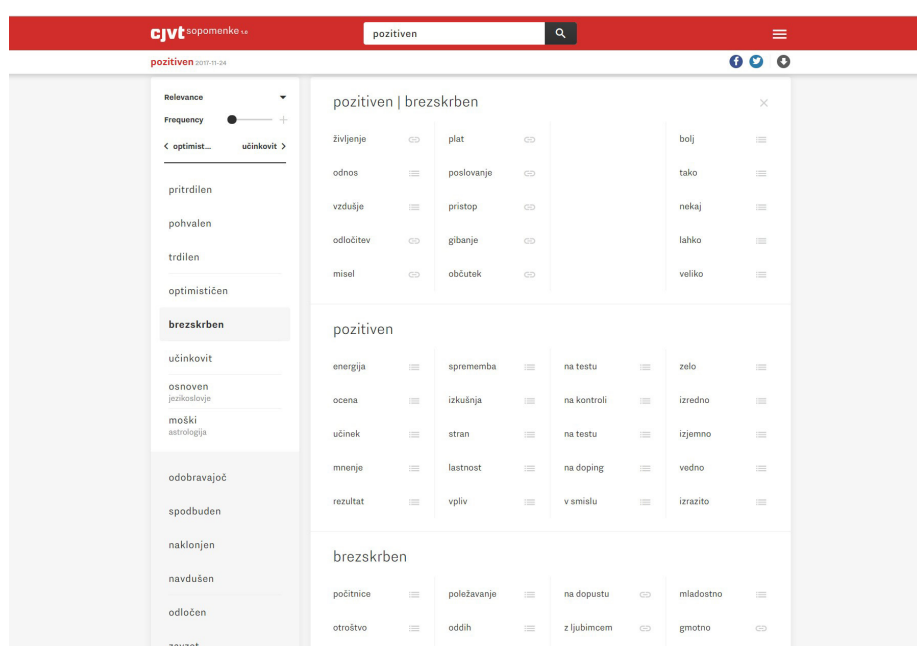


Figure 4: Collocations page with collocates for *pozitiven* ('positive') and *brezskrben* ('carefree').

## 4 Conclusion and Future Work

In the paper, we have presented the Thesaurus of Modern Slovene and the main ideas underlying its construction. The first upgrade of the Thesaurus is planned for the end of 2018 and will provide an opportunity to evaluate and refine the methodology for further steps in dictionary development. Apart from the plans already mentioned in the previous sections of the paper, we point out here the most essential steps of our future work. Firstly, we plan to extend the core database with new synonyms using dense neural vector embeddings (Mikolov et al. 2013), an unexploited source of synonymy detection which projects similar words into a latent vector space where distances preserve the similarity among words. Clustering in this space may provide new candidates for synonyms, especially with variants of embeddings that construct more than one vector per polysemous word (Huang et al. 2012, Peters et al. 2018). Secondly, we intend to use the Pybossa crowdsourcing platform (<https://pybossa.com/>) to prepare a number of specific crowdsourcing tasks to improve the existing database (e.g. removing archaic or idiosyncratic synonyms from literary works). Thirdly, the KOLOS national project (*Collocations as a Basis for Language Description: Semantic and Temporal Perspectives*, ARRS J6-8255) will provide a framework for improving the quality of the collocational data in the Thesaurus. Finally, studies are planned to further address different aspects of user participation and involvement. As mentioned in Section 1.1, the results of the linguistic evaluation regarding the appropriateness of the synonyms included in the core database were not uniform. This can be explained by the fact that synonymy is a contextual linguistic phenomenon, where the true value of synonyms can be evaluated exclusively in their context and within the exact purpose of the user. The collective authorship of the responsive dictionary can thus provide valuable insight regarding the extent to which user knowledge corresponds to expert linguistic assessments. On the other hand, the perception of the language community regarding the novelties brought by the responsive dictionary concept will also be addressed: a study is underway combining think-aloud protocols and semi-structured interviews with representatives of selected user groups in order to obtain feedback on dictionary functionality (e.g. the usefulness and the clarity of the interface), as well as opinions on the democratization of the dictionary creation process, different issues regarding user involvement (e.g. the question of authorship, ethical issues, and quality control), the dynamic nature of a responsive dictionary, etc. The results will help pinpoint the potential weaknesses of the concept that will have to be given priority in the future.

## References

- Arhar Holdt, Š., Kosem, I. & Gantar, P. (2016). Dictionary user typology: the Slovenian case. In T. Margalitadze & G. Meladze (eds) *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 179-187.
- Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (2017). Value of language-related questions and comments in digital media for lexicographical user research. *International journal of lexicography*, 30 (3), pp. 285-308.
- Cristea, D. Forăscu, C., Răschip, M. & Zock, M. (2014). How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008*, Marrakech, Morocco.
- Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing, pp. 70-83.
- Čibej, J., Gorjanc, V. & Popič, D. (2016). XVII EURALEX International Congress, 6-10 September, 2016, Tbilisi. Analysing translators' language problems (and solutions) through user-generated content. In T. Margalitadze & G. Meladze (eds) *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 158-167.



- De Schryver, G. M. & Prinsloo, D. J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds): *Proceedings of the 9th EURALEX International Congress*. Institut für Maschinelle Sprachverarbeitung: Stuttgart, Germany, pp. 197-209.
- Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2017). *Dictionary of modern Slovene: problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts.
- Huang, E. H., Socher, R., Manning, C. D. & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp. 873-882.
- Kilgariff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*, EURALEX 2004 Lorient, France July 6–10, 2004. Lorient: Universite de Bretagne-sud, pp. 105–116.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris. (eds) *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425–432.
- Kosem, I. (2017). Dictionary examples. In V. Gorjanc, P. Gantar, I. Kosem and S. Krek (eds.) *Dictionary of Modern Slovene: problems and solutions*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, pp. 174-194.
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (in print). Collocations Database and Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana University Press, Faculty of Arts: Ljubljana.
- Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017*. Leiden, Netherlands, pp. 93-109.
- Logar, N. & Krek, S. (2012): New Slovene corpora within the "Communication in Slovene" project. In *Prace Filologiczne*, 63, pp. 197-207.
- Meyer, C. M. & Abel, A. (2017): User Participation in the Era of the Internet. In P. A. Fuertes Olivera (ed.): *The Routledge Handbook of Lexicography*. London: Routledge, pp. p. 735-753.
- Mikolov, T., Yih, W.T. & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746-751.
- Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of NAACL*.
- SSKJ, Slovar slovenskega knjižnega jezika [Electronic version] (2014). A. Bajec et al. (eds.). Ljubljana: Založba ZRC, Znanstvenoraziskovalni center SAZU. Available at: <http://www.fran.si>.
- Šorli, M., Grabnar, K., Krek, S. & Košir, T. (2006). Oxford-DZS comprehensive English-Slovenian dictionary. In *Proceedings XII EURALEX international congress*. Edizioni dell'Orso: Università di Torino: Academia della Crusca, pp. 631-637.

## Acknowledgements

The paper was prepared as part of two national projects, *Nova slovnica sodobne standardne slovenščine: viri in metode* (New grammar of contemporary standard Slovene: sources and methods, ARRS J6-8256) and *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (Collocations as a Basis for Language Description: Semantic and Temporal Perspectives, ARRS J6-8255). The development of the Thesaurus was funded by the ARRS P6-0215 research program (*Slovene language – basic, contrastive, and applied studies*) and the infrastructural programme of the Centre for Language Resources and Technologies at the University of Ljubljana. The interface was developed by Studio Kruh in collaboration with Leon Noe Jovan. The crowdsourcing development benefited from the COST Action CA16105: *European Network for Combining Language Learning with Crowdsourcing Techniques*.

# Dictionary of Verbal Contexts for the Romanian Language

**Ana-Maria Barbu**

*Institute of Linguistics, Romanian Academy, Bucharest*

*E-mail: anamaria.barbu@g.unibuc.ro*

## Abstract

This paper presents a dictionary of verbal contexts for Romanian, which comprises 600 verbs and over 2,000 meanings with one or more valency patterns. It is manually built but based on corpus information, and is developed both for teaching Romanian to foreigners, by its printed version, and for computational linguistics, by its XML format and consistent principles and conventions of the design. The dictionary is rich in information, including lexical, grammatical and semantic features of the complements, morphosyntactic variants occupying an argument position, dependencies between complements induced by control, raising and predication phenomena and verbal alternations, as variants of valency patterns with the same meaning. The paper offers details about all this information, the building procedure and some problems that needed to be solved during our work. This enterprise is far from being finished, because further work has to be done to improve the actual encoding and add new types of information, such as semantic roles or diathesis uses, for growing the number of entries and for getting different kinds of generalizations.

**Keywords:** verb pattern, verbal context, valency, argument, complement, Romanian verbs

## 1 Introduction

This study presents a work focused on the concept of verb valency, seen as a configuration of dependents (i.e. arguments or complements) lexically required by a verb (or a predicate, in general). The concept of valency, as is well known, is rooted in Tesnière's theory of dependency grammar, however, as Herbst (1999: 2) points out, the beginnings of a valency theory as such are not found in works within dependency grammar, but within foreign language teaching; and they are marked by the first valency dictionaries, such as Helbig and Schenkel (1968) or Engel and Schumacher (1976). Since then, both valency theory and valency dictionaries have shown their scientific potential and utility. Nowadays, important efforts are being made to build such dictionaries in more and more languages, especially with the support of computational and corpus linguistics, which, in turn, is a main beneficiary of such a valuable resource.

In this vein, the *Dictionary of Verbal Contexts for Romanian Language* (hereafter DVCRL) we present here represents a similar endeavor. This dictionary is manually built but mainly based on internet data, and comprises 600 verbs and over 2,000 meanings with one or more verb patterns (or complementation patterns), both in machine-readable format and in printed version (see Barbu 2017). This work started about ten years ago and has, as models of that time, FrameNet (Johnson & Fillmore 2000), VerbNet (Kiper et al. 2000), CPA (Hanks & Pustejovsky 2005, Hanks 2006) and VALLEX 2.0 (Žabokrtský & Lopatkova 2007), as described and analyzed in Barbu (2008). The dictionary building had two phases. The first was developed by five linguists, within a national project spanning three years, which collected about 3,000 verb entries.<sup>1</sup> In this phase, the description of valencies was limited mainly to obligatory complements (minimal valencies) and to the verb meanings in the Romanian Explanatory Dictionary (DEX 1998). The lack of further funding and the need to solve the

<sup>1</sup> The first version of the dictionary was created within the CNCSIS project nr. 1156/2005, during the years 2005-2007.

shortcomings observed in the previous version urged us to take over the improvement of the dictionary on our own, being aware of how important is to obtain such a linguistic resource, and that such work has to be done as well as possible. In these conditions, the enterprise advanced very slowly and could only cover much fewer entries, due to the fact that it was supported by only one linguist, who for a while was working outside his job duties.

The improvements made in the second phase mainly address the following aspects. On the one hand, we have paid special attention to facultative or optional complements<sup>2</sup> and adjunct-like elements (or modifiers), also encoded in FrameNet, Vallex and CPA. For instance, any directed-movement verb has a Path, including Origin and Goal, as its lexically-governed complements, but they are facultative because can be omitted depending on the communication context, like Path and Goal in *John has gone from the school*. Furthermore, a verb like *a căuta* ‘search’ is very frequently accompanied by a Locative modifier, unlike other types of modifiers (e.g. Time or Manner), which justifies the inclusion of a Locative element in the complementation pattern of this verb. Due to these facts, we preferred the term *verbal context*, instead of *verb valency*. While *verb valency* (predicate-argument structure or subcategorisation frame) generally refers to *minimal* number of arguments required by a verb for accomplishing its meaning, *verbal context* includes, as understood here, adjunct-like elements which are frequently used in events centered on a certain verb, in the same line as Fillmore’s *semantic frame*.

On the other hand, we have consolidated the concept of *argument position*, allowing morpho-syntactic variants (Barbu & Ionescu 2007:45), in the same vein as that mentioned later in (Przepiórkowski et al. 2014: 2786), as we show below. Another improvement refers to the specification of alternations, considered in a more restricted sense than Levin’s (1993), that is, as sets of verb patterns with, roughly, the *same* meaning. The meanings themselves of each verb had to be re-thought, as we point out below in the next section. We have not left aside the information indicating dependencies between different complements of the same pattern either. Such dependencies are induced by control and raising phenomena or small clauses. It is worth mentioning that all the verb contexts are valid for declarative sentences and the active voice, and that many regular transformations or variations, which do not have a lexical nature, are not caught in the encoding of this dictionary.

In what follows we present some aspects of the building procedure used in this project, with a short explanation of why we had to dispense with DEX meanings and re-think them. In Section 3, we detail the information structure of an entry and verb pattern, while we also explain some of our encoding options. This section deals with the topics of alternations and XML representation, as well. Some challenging aspects of our enterprise are treated in Section 4, and we conclude the paper with general remarks and plans for further work.

## 2 The Building Procedure

Due to the fact that DVCRL is manually built, it is not founded on valency samples extracted from a large corpus, as other approaches are, but on meaning-oriented bases, by tightly correlating meanings with contexts in use. Thus for each verb meaning taken from monolingual dictionaries, we build contexts by intuition and check them on the internet (seen as a corpus, cf. Kilgariff & Grefenstette 2003), which can bring up new, unexpected contexts.<sup>3</sup> The search online cannot be done by using part-of-speech tags, and therefore for obtaining rarer structures we used some lexical expressions we know are used, eventually adapting them with the wildcards allowed by Google search engine.

<sup>2</sup> Actually, in this paper we settle a new distinction between facultative and optional complements.

<sup>3</sup> Note that the Romanian language has had a large part-of-speech annotated corpus only since the end of 2017, and therefore we could not adopt a corpus-oriented methodology.

With this we have noticed, quite surprisingly, that a corpus, however big it could be, does not cover a good number of language facts. Despite these difficulties, in our opinion, this work method, which combines corpus analysis with native speaker intuition, has some advantages over an exclusively corpus-oriented approach, because it ensures benefits like the following: access to less frequent uses of a verb (for less used meanings); human generalization of the information unavailable in a corpus (e.g. semantic roles or selectional restrictions); and the correspondence between the valency frame and the meaning, ready to use, for instance, in computational semantic disambiguation.

In the first phase, for each verb, we adopted its meanings from the Romanian Explanatory Dictionary (DEX 1998) and assigned the minimal verb contexts for each of them, as described in Barbu and Ionescu (2007) and Barbu (2008). But this turned out to be a bad start, because we could not obtain a convenient correspondence between the patterns of a verb and its DEX meanings. For instance, for the transitive verb *a aplauda* ‘applaud’ DEX offers only the meaning “to express satisfaction, approve or admiration for something by clapping the hands”. By a careful inspection of the verb contexts in use, one can find two embarrassing facts: 1. the subject of this verb can refer to entities without hands (e.g. institutions or organizations) so that no clapping of hands can happen – actually it shows up a new meaning: “to praise”; 2. when clapping of hands is involved, and only in this case, the direct object can be omitted. In order to capture these peculiarities in our encoding, we needed to set two different semantic restrictions: +human for “human creature (with hands, feet, eyes, etc.)” and +person for “physical or juridical person”, and two different verb patterns equivalent (in Romanian) to those in example (1) (see Figure 1 below for the structure of such an entry):

(1) applaud

I.

1. NP [nom +human], 2. (fac.) NP [acc]

to clap the hands (for expressing the approval of someone or something): *The audience applauded (actors) at the end of the show in standing ovation.*//

II.

1. NP [nom +person], 2. NP [acc]

to praise: *Coalition applauds New York City's universal physical education initiative.*//

Note that in the first verb pattern (I.) the noun phrase in accusative (NP[acc]) can be omitted (it is marked by (fac.)), which is mirrored in the pattern example by the omissible word *actors* placed inside parentheses. Instead, in the second verb pattern (II.) the direct object is obligatory, otherwise odd utterances such as *\*Coalition applauds* are obtained.

As one can see, different semantic restrictions (e.g. +human versus +person) can trigger and justify different verb patterns. Actually, finding what particularizes a certain pattern was the most important challenge of our work.

As a general method of dictionary building, we keep DEX as a reference dictionary but we have to reconsider the verb meanings in DEX. Not every meaning in DEX also deserves to be mentioned in our dictionary, because there are some meanings with the same valency structure. At the same time, we discovered meanings not registered in DEX. The problem of this meaning mismatch has been previously touched on by Herbs et al. (2004: xxxviii), who claim: “Establishing senses according to their valency patterns in some cases results in a rather different identification of senses than in conventional dictionaries”.

The inventory of the verb entries is established according to decreasing occurrence frequencies of the verbs extracted from a newspapers corpus. This led us to include, in DVCRL, verbs used in mass-media rather than in the core vocabulary of Romanian, so that verbs such as *a mânca* ‘eat’ or *a dormi* ‘sleep’ could be missing.



For editing the dictionary, we used the Professional Lexicography Software TshwaneLex (tshwaned-je.com). It has many functions, but we only mention the XML editor, the possibility of RTF or HTML export and the so-called WYSIWYG (“what you see is what you get”) view. These help to easily get the dictionary in machine-readable and printed formats. Even if the software is provided with DTDs (Document Type Definition schemata) appropriate for bi- and mono-lingual dictionaries, we had to get rid of them and to build a DTD specific to the information structure of verbal context entries (see §3.5). Before writing the DTD, one has to get the final form of the information structure and XML annotation schema, because it is very risky to change the DTD after beginning the work of dictionary expansion. In general, adding new elements is possible, if child relations are not changed. Therefore, we firstly worked out, in text format, a good number of verbs (with one or several meanings/patterns) in order to capture the relevant and general information structure.

### 3 The Information Structure

#### 3.1 The Structure of an Entry

An entry is headed by a verb lemma and has the information structure shown in Figure 1, illustrating the verb *a concura* ‘take part in a tournament’ (I), ‘compete’ (II) or ‘act together’ (III). For each lemma, one or more *verbal contexts* are given, numbered with I, II, etc. Such a verbal context includes:

a) Morphological information about the verb. For instance, the context III is used only at the 3<sup>rd</sup> person (singular or plural) and it is marked by V[3], accordingly. This type of information can also refer to specific mood or tense, as well as to pronominal or negation clitics of the verb, and it is enclosed inside square brackets preceded by V (see also V[se] in Figure 1 context II, which indicates that the verb has the reflexive clitic *se* in that verb pattern).

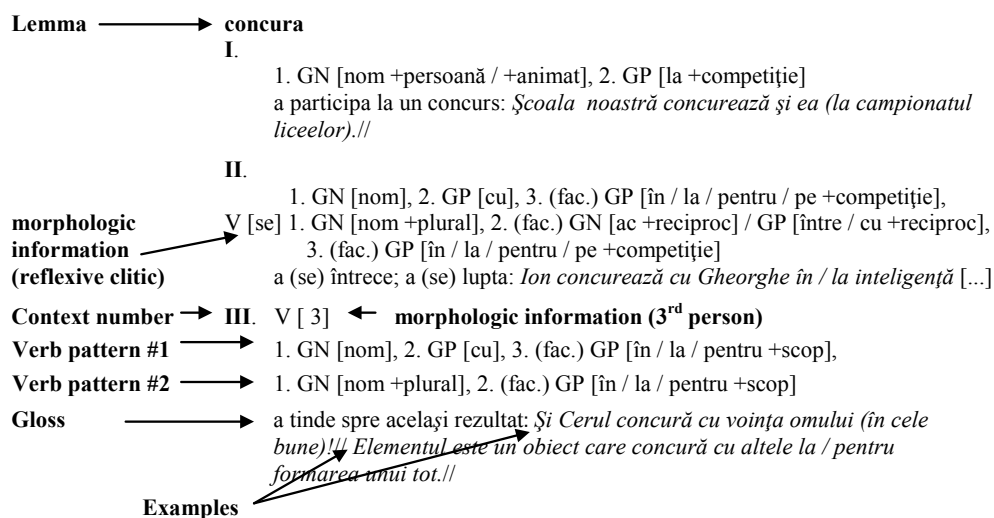


Figure 1: Entry structure.

b) One or more *verb patterns*. As Figure 1 shows, verb alternations (of two or more patterns) are described. The structure of a verb pattern is explained in sub-section 3.2. In the first phase of the project we considered that the morphological information applies to a certain context/meaning and we have assigned this information to the context level (see context III), that is higher than the pattern level. During the work, it turned out that such an information type was also needed at the pattern level in cases of alternations. For instance, context II displays the reciprocal alternation, which has a variant



implying, in Romance languages, the use of a reciprocal/reflexive clitic on the verb. This fact compelled us to make a late change to the dictionary DTD, fortunately without any perturbation of the existing annotation scheme.

c) One or more glosses by means of synonyms and sometimes paraphrases. In general, we strive to choose the synonyms that can replace the lemma verb in the given examples. If examples found in corpus fit the same verb context but require non-synonymic glosses, then several senses are provided (marked with a., b., etc.).

For instance, in the example (2), the Romanian verb *a cuprinde* ‘embrace’ has three meanings: a. to hold close with the arms, usually as an expression of affection: *John embraced Mary around her waist*, b. to surround or enclose: *The mist embraced the hills* and c. to include or contain as part of something broader: *The contract embraces the elements of work*.

## (2) *cuprinde*

1. GN [nom], 2. GN [ac]

a. a îmbrăţişa: *Ion a cuprins-o pe Maria de mijloc.*//

b. a închide în sine: *Ceaţa a cuprins dealurile.*//

c. a conţine; a fi alcătuit din: *Contractul cuprinde elementele muncii prestate.*//

d) One or more examples. We give examples for each pattern alternation and each description variant marked with ‘/’. We propose our own examples whenever they express a common use, or are extracted from internet (and adapted by shortening or person name deletion) when they show special uses or meanings.

e) One or more idioms centered on the lemma (if it is the case). Idioms can be considered verbal contexts with lexicalized complementation, and this is the reason for including them in our description. The verb *a ajunge* ‘reach’ has the idiom in example (3), where what follows the ‘=’ mark is its gloss. An idiom can have a fixed part and a mobile one, differentiated in (3) by two character types: italics and normal, respectively. The mobile part can take different references (see *cuiva* ‘to somebody’ in (3)) or can be inflected.

(3) *a-i*            *ajunge* *cuiva*            *cuţitul*    *la os*    = a fi într-o situaţie disperată  
to-clitic.dat reach somebody.dat knife.the to bone = to be in a desperate situation

## 3.2 The Structure of a Verb Pattern

A verb pattern describes the complements of the verb lemma corresponding to a certain meaning and it has the information structure shown in Figure 2.<sup>4</sup>

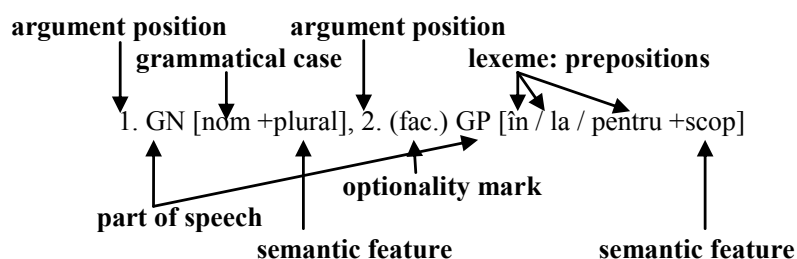


Figure 2: Verb pattern.

<sup>4</sup> GN=noun phrase (NP), GP=preposition phrase (PP), nom=nominative case, ac=accusative case, +plural=plural, fac.= facultative, în/la/pentru=in/to/for, +scop=goal.

As one can see in Figure 2, a verb pattern contains one or more *argument positions*. We talk about *positions* instead of arguments or complements, because such a position can be occupied by different types of phrases. For instance, the predicative element of the verb *be* can be a noun phrase: *He is a doctor*, an adjective phrase: *He is very happy* and so on, and all of these types of phrases occupy the same position (the predicative one) as variants. For Romanian, a free word order language, there is no connection between the positions in the verb pattern and the complements order in utterances. However, by convention, the first position is reserved for the subject.

As is well known, a position is obligatory or not, but it is worth pointing out some further aspects of this here. By obligatory complement one usually understands a complement which cannot be excluded from the immediate context of the verb. However, many linguists, such as Allerton (1975), Fillmore (1986) or Herbst (1999), point out that there are unexpressed complements that have to be obligatorily recovered from the larger linguistic or extra-linguistic context. In other words, such a null (or recoverable) complement, even if it is not expressed, is not deleted from the argument structure of the verb. In DVCRL, we consider such a complement obligatory as well, and consequently we do not mark it as optional; however, in examples we put it in parenthesis as omissible in immediate context of the verb. At the same time, the optional complements are those which are legitimated by the meaning of the verb but can be relevant in some contexts and irrelevant in others. Such complements are marked in DVCRL with (fac.) ‘facultative’ (see Figure 2). More than that, we have found a special case of optional complements, when any of several complements of the same verb pattern can be omitted but at least one has to be expressed. For instance, in Romanian there are light verbs whose meanings are strictly determined by their complements. Such a verb is *a da* ‘give’ which acquires a motion meaning if it is accompanied by motion complements (i.e. Goal, Origin or Path). In the example (4)a the verb has all the complements expressed.

(4)

- (a) El   dă           bagajele   jos   din tren           pe geam.  
He   gives       the luggage down from the train through the window.  
‘He throw his luggage out of the train window.’
- (b) El   dă   bagajele   jos.  
He   gives the luggage down.
- (c) \*El   dă   bagajele.  
He   gives the luggage.

In (4)b there is only one complement (the adverb *jos* ‘down’), expressing the Goal, which cannot be further deleted without making the utterance ungrammatical for the intended meaning, (4)c. For these cases of optional complements, we set the mark (opt.) ‘optional’, obligatorily appearing in a series of two or more complements in a verb pattern and having the meaning that any complement in the series can be missing but not all of them. The difference between (fac.) and (opt.) is therefore given by the fact that in the former case the complement can be missing independently of the other complements, while in the latter case the complement can be missing only if at least one of the other complements marked by (opt.) is expressed.

The necessity of working with argument positions was actually dictated by the optional complements, which raise the question as to whether or not they can co-occur when expressed. If they do co-occur in standard configuration (that is, not involving coordination, restriction or appositive specification, for instance), then, logically, they occupy different argument positions. This turns out to be the case for obligatory complements too, when establishing the morpho-syntactic variants (i.e. phrases of different parts-of-speech) assigned to a certain argument position. In other words, the variants exclude each other, so that the complement of *quantity* (marked +quantity) of the verb *weigh*, for example, can be expressed by a noun phrase (e.g. *The bag weighs two pounds*) or an adverb phrase (e.g. *The*

*bag weighs heavily*) but not by both at the same time (e.g. \**The bag weighs two pounds heavily*). Such a verb should have the verb pattern in example 5, where ‘/’ separates variants:

(5) 1. NP[nom +object], 2. NP[nom +quantity] /AdvP[ +quantity]

The concept of argument position with morpho-syntactic variants is also encountered in Przepiórkowski et al. (2014: 2786), but the test used there to detect the occupants of the same position is not the co-occurrence principle but the coordination. In our opinion, coordination is too sensitive to all kind of linguistic and extra-linguistic factors, so that a coordination failure could lead to incorrect results. For instance, in Romanian an example such as *Sacoşa cântăreşte două livre şi greu* ‘The bag weighs two pounds and heavily’ is very odd. By applying the coordination principle the two phrases should occupy different positions, and that is wrong.

A position can be fulfilled by one phrase (or more) (see Figure 2). A phrase is characterized by its part-of-speech, grammatical or lexical information and semantic information. Grammatical information refers to grammatical cases for noun and adjective phrases (see *nom* for nominative in Figure 2), or refers to mood for verb phrases (i.e. participle, gerund or supine). Lexical information is offered for governed preposition and adverb phrases (see the prepositions series in Figure 2), and it specifies conjunctions or relative adverbs in subordinating clauses. Semantic information (preceded by +) mainly represents semantic (or selectional) restrictions, but it can reach the generality of a semantic role such as Goal (see *+scop* in Figure 2) or Source, Cause etc., in order to characterize, for instance, a series of possible prepositions with appropriate meaning without specifying them explicitly. Note that this dictionary does not use Thematic Roles nor micro-roles (cf. Hartmann et al. 2013), nor even grammatical functions. We focus on “visible”, uncontroversial marks such as grammatical cases (note that Romanian is a fully inflectional language), which express, in general, certain Thematic Roles and grammatical functions (e.g. accusative is Patient/Theme and direct object).

### 3.3 Alternations

As pointed out in Kettnerová et al. (2012: 434), despite the growing attention paid to alternations in theoretical linguistics (especially since Levin’s seminal work (1993)), there are few valency lexica approaching this topic. DVCRL includes alternations explicitly and consistently. However, in contrast with Kettnerová et al., we have completely ignored what the authors call “grammaticalized alternations”, more precisely the regular transformations of diatheses and other Romanian linguistic phenomena, such as possessive dative or object duplication. Even among the morpho-syntactic variations on the argument positions, which Levin counts as alternations too (see, for example, 1993:43 Preposition drop alternation), there are some *regular* phenomena such as replacing an NP with a free-relative clause: *My friend believes me* can be replaced by *Who is my friend believes me* or a locative, time or manner complement, by its corresponding relative clause: *I go there*, by *I go where I like*. These kinds of replacements apply for any verb and any argument position, so it would be redundant to mention them in the dictionary.

Instead, we have encoded what Kettnerová et al. call “lexical alternations”, to which we have added the reciprocity transformation, which is word-oriented, despite its regular transformation pattern. In Figure 1, both verbal contexts II and III display reciprocity alternation, but of different types. While the first requires a reflexive clitic and permits an omissible reciprocal complement (e.g. *unul pe altul* ‘each other’) such as in (6)a, the second accepts neither clitic nor reciprocal complement, (6)b.

- (6) (a) Ion concura cu George în afaceri. ↔ Ion şi George se concureau (unul pe altul) în afaceri.  
 John competed with George in business. ↔ John and George competed (with each other) in business.

(b) Viziunea regizorului concura cu cea a actorilor. ↔ Viziunea regizorului și a actorilor (\*se) concureau (\*una pe alta).

The director's vision concurred with that of the actors. ↔ The vision of the director and that of the actors concurred.

Unlike Kettnerová et al., we use neither general alternation rules, nor situational participants (e.g. Agent, Recipient, etc.). The distinction between the variants of the same alternation is done through semantic restrictions, because often an alternation does not imply a simple reorganization of the situational participants but a random complement variation, like in the example (7) for the verb *a achita* 'pay', where (a) and (b) are valency variants of the same meaning "to give money for a commercial product":

(7) (a) 1. NP [nom +person], 2. NP [acc +goods]

Ion a achitat băutura.

John paid the drink

'John paid for the drink.'

(b) 1. NP [nom +person], 2. NP [acc +money], 3. PP [pentru 'for' +goods]

Ion a achitat un euro pentru băutura.

John paid one euro for the drink.

Notice that in Romanian the commercial product can be expressed either by an NP in accusative (7) a when the paid price is not expressed, or by a *pentru* 'for'-PP (7)b, when the price is specified as direct object.

It is very likely that when DVCRL covers almost all the Romanian verbs one may obtain valuable generalizations about alternations by capturing the recurrent patterns in some rules.

### 3.4 Dependencies Between the Complements

In order to capture the dependencies between the complements of a verb pattern, expressing control or raising phenomena or case agreement, we have used a set of conventions.

If a verbal complement has, as its subject, the NP in nominative (that is, the same subject as the lemma verb), its part-of-speech is indicated simply by V, like in example (8) for the subject-control verb *a promite* 'promise':

(8) 1. NP [nom +person], 2. V [să 'to'], 3. NP [dative + person]

Ion promite Mariei să o iubească întotdeauna.

Ion.nom promises Maria.dat to her.clitic.acc love forever.

'Ion promises Maria to love her forever.'

The verb *promite* has the same subject as its verbal complement in subjunctive (marked with V[să 'to']): *să iubească* 'to love', namely the NP in nominative: *Ion*, while the NP in dative: *Maria* shows to whom is addressed the promise.

If the verbal complement has as its subject an NP different from the one in nominative, then we use VP (without any semantic restriction), like in example (9) for the object-control verb *a recomanda* 'recommend':

(9) 1. NP [nom +person], 2. NP [dat +person], 3. VP [să 'to']

Ion îi recomandă Mariei să plece.

Ion recommends to Maria to leave.

This time, the subject of the verb in subjunctive *să plece* 'to leave' is the person to whom the recommendation is addressed, namely the reference of the NP in dative.

When a complement is a subordinating clause that can have a subject that is different from the other NP-complements, we use the mark VP as well, but we also indicate a semantic restriction in the set +fact, +question, +cause, +goal etc. for that clause. The example (10) shows such a case.

- (10) 1. NP [nom +person], 2. NP [acc +money], 3. (fac.) VP [ +goal]  
 Ei cheltuie bani ca el să aibă succes.  
 They spend money for him to be successful.

The situation of case agreement concerns cases in which a complement is a predication of another complement, like in example (11) for the verb *a considera* ‘consider, regard as’, where the argument position 3 refers to the argument position 2 and must have the same grammatical case, here the accusative. These are structures of small clauses. The predication can be expressed by an adjective phrase (AdjP) or by a noun phrase, as variants of the same position.

- (11) 1. NP [nom +person], 2. NP [acc], 3. AdjP [acc] / NP [acc]  
 Eu îl consider pe el bun/prieten.  
 I consider him good/my friend.

### 3.5 XML Encoding

Any XML (eXtensible Markup Language) encoding needs a DTD (Document Type Definition) that is a schema ensuring a logical and consistent structure of all the dictionary entries. The TshwaneLex DTD editor allows lexicographers to tailor the appropriate schema for every kind of dictionary, without the need for an IT expert. In Figure 3 we give a sketch of the DTD used in DVCRL and an XML example. The main lines in designing a DTD concern the information types captured in the encoding (identified by the element names), their embedding structure (see the indentation expressing a child relation) and the multiplication number of the corresponding elements (see the superscript symbols). An element encompasses other elements (with internal structure) and/or attributes (with terminal values).

In DVCRL schema, an entry, named Lemma, has ArgUnit and Expression as child elements and the attribute LemmaSign taking the verb lemma as its value. The ArgUnit element, which has to be at least one or more (see superscript +), encodes the verbal context and has a numbering attribute ArgUnitNr. The children of ArgUnit are the morphological information, which can appear once or can be missing (LemmaFeature with superscript (0,1)), one or more verb patterns (ArgStructure<sup>+</sup>) and one or more numbered senses (Sense<sup>+</sup>). Further, each element has the structure displayed in Figures 1 and 2. Note that the meanings of the superscript symbols are the following: \* – zero or more, 1 – strictly one, and inside brackets there are lists of possible attribute values.

The TshwaneLex editor offers the possibility of assigning specific layout characteristics to DTD elements and attributes, thus relieving lexicographers of the need to worry about these. This facility can trigger, sometimes, the need for element restructuring and small schema modifications. For instance, for treating several expressions of a lemma as items of the same type we should include them into an Expression List element. In other words, the final form of a DTD can be motivated not only by the information structure, but by layout features as well. In the second column of Figure 3, one can see how DTD is reflected in XML encoding, by the example of the first verbal context of the verb *concura* (see verbal context I in Figure 1).<sup>5</sup>

5 The XML variant of this dictionary can be obtained by request from the author.



DTD schema	XML example for <i>concura</i> I.
Lemm:LemmaSign = <i>text</i> ArgUnit <sup>+</sup> : ArgUnitNr = <i>integer</i> LemmaFeature <sup>*</sup> : ReflForm <sup>(0,1)</sup> = [ <i>se, își, o, îl</i> ] NegForm <sup>(0,1)</sup> = [ <i>nu</i> ] FlexForm <sup>(0,1)</sup> = [ <i>3sg</i> ] ArgStructure <sup>+</sup> : LemmaFeature <sup>*</sup> : ReflForm <sup>(0,1)</sup> = [ <i>se, își, o, îl</i> ] NegForm <sup>(0,1)</sup> = [ <i>nu</i> ] FlexForm <sup>(0,1)</sup> = [ <i>3sg</i> ] ArgPosition <sup>+</sup> : PositionNr = <i>integer</i> PositionState <sup>(0,1)</sup> : PositionState = [ <i>fac., opt.</i> ] Argument <sup>+</sup> : POS <sup>1</sup> = [ <i>GN, GAdj, GAdv, GP, GV, V</i> ] Afeature <sup>1</sup> : Gram <sup>*</sup> = <i>text</i> Sem <sup>*</sup> = <i>text</i> Sense <sup>+</sup> : SenseNumber = <i>integer</i> Definition <sup>+</sup> : Definition <sup>+</sup> = <i>text</i> Example <sup>+</sup> : Example <sup>+</sup> = <i>text</i> Expression <sup>*</sup> : LemmaSign = <i>text</i> Definition = <i>text</i>	<Lemma LemmaSign="concura" <ArgUnit ArgUnitNr="1"> <ArgStructure > <ArgPosition PositionNr="1"> <Argument POS="GN"> <AFeature Gram="nom" Sem="+persoană / +animat"/> </Argument> </ArgPosition> <ArgPosition PositionNr="2"> <Argument POS="GP"> <AFeature Gram="la" Sem="+competiție"/> </Argument> </ArgPosition> </ArgStructure> <Sense SenseNumber="1"> <Definition Definition="a participa la un concurs"/> <Example Example="Școala noastră concurează și ea (la campionatul liceelor)."/> </Sense> </ArgUnit> [...] </Lemma>

Figure 3: XM Encoding.

## 4 Challenging Aspects

The most challenging task in this approach was to establish the appropriate characteristics of a verb pattern, especially regarding the semantic features (or selectional restrictions) of each complement. This has been done by resorting to corpus examples but also to human intuition, because it is not always possible to reach, in a huge corpus like the internet, all the structures that are in use, especially those less frequent ones. Actually, a dictionary like DVCRL is comprised not of verb patterns found in corpus, but what it is *possible* to be found, described in a concise manner. We tailor an intuitive verb pattern, by starting from a meaning of a verb, and then verify it on corpus examples. After gathering the examples, the question that arises is what the complements have in common. Thus, certain semantic preferences can show up and they have to be defined in the most general way as possible. At the beginning we chose intuitive semantic tags such as +human, +object, +authority, +event, etc. We then extended this list as needed, although using the tags already defined as much as possible, in order not to increase the list excessively. At the end, we removed the semantic features assigned to less than three verbs and included them explicitly in the verb definition or as a note. For instance, one of the meanings of *a se ambala* is “get running uncontrollably fast” and requires a subject with the semantic feature +horse. Because this feature was singular, we have replaced it with +animate and specified in the definition that it is about a horse. Finally, we obtained a list of 90 semantic tags (for about 2,000 meanings).

Many verb patterns differ only by the semantic features of their complements. For instance, in Romanian the verb *amputa* ‘amputate’ has the medical use (of cutting off a limb) but also a general or

metaphorical meaning, namely “to remove a part of something with bad consequences” like in the example “[...] but people apparently have no problem with billions being squandered on an *amputated budget* in Europe” (<http://www.europarl.europa.eu/>). The latter meaning applies to budget, texts, personality, future, and so on, to the effect that no specific semantic preferences can be delimited. These two meanings are assigned the verb patterns in example 12, respectively:

- (12) (a) NP [nom +animate], 2. NP [ac +limb]  
 (b) NP [nom], 2. NP [ac]

As one can notice, the verb pattern in (12)b includes, somehow, that in (12)a. However, which is better: to conflate or to keep them separately? Such questions arose throughout our work.

## 5 Conclusion

This study describes a verb valency lexicon for the Romanian language – here valency is understood in a broad sense, corpus and user-oriented. This dictionary is the first endeavor at this level of comprehensiveness and consistency for this language, and it is intended to be a valuable resource for both computational linguistics and teaching Romanian to foreign students. Besides the entry description, we tried to bring out some aspects of our experience, such as the necessity to detach the meanings of verb patterns from those in an usual monolingual dictionary. We consider this fact to be important because, as we know, many tasks of word sense disambiguation use such monolingual dictionaries as references, which do not reflect the formal aspects found in a corpus, as a valency dictionary does, and thus the computational performance is diminished. We also touch problems less discussed in the literature, such as a new type of optional complements, the argument position concept based on co-occurrence test and the dependencies between complements of the same verb pattern.

The dictionary presented here is just a starting point. There is a lot of room for improving the informativeness and the consistency of its entries. Information about semantic roles, diatheses and collocation preferences waits to be added, as does increasing the number of entries. We also hope to get valuable feedback from the dictionary’s users as soon as possible, in order to use this as supplementary guidelines for further work.

## References

- Allerton, D. J. (1975). Deletion and proform reduction. In *Journal of Linguistics*, 11, pp. 213-238.
- Barbu, A.M. (2008). First Steps in Building a Verb Valency Lexicon for Romanian. In P. Sojka et al. (eds.), *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence (LNAI) 5246, Springer, pp. 29-36.
- Barbu, A.M. (2017). Dicționar de contexte verbale. Editura Universității din București.
- Barbu, A.M. & Ionescu, E. (2007). Designing a Valence Dictionary for Romanian. In *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP2007)*, 27-29 September 2007, Borovets, pp.41-45.
- DEX (1998). Dicționarul Explicativ al Limbii Române. Institutul de Lingvistică „Iorgu Iordan –Al. Rosetti” al Academiei Române, Editura Enciclopedic Gold.
- Engel, U. & Schumacher H. (1976). Kleines Valenzlexikon deutscher Verben, Tübingen: Narr.
- Fillmore, C. (1986). Pragmatically Controlled Zero Anaphora. In *Proceedings of the 12th Annual Meeting of the Berkeley Linguistics Society*, BLS12, Berkley, 15-17 February 1986, pp. 95-107.
- Hanks, P., Pustejovsky J. (2005). A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, X-2, pp. 63-82.
- Hanks, P. (2006). The Organization of the Lexicon: Semantic Types and Lexical Sets, <http://www.cs.cas.cz/sem-web/download/06-11-hanks.doc> [03/20/2018].

- Hartmann, Iren & Haspelmath, Martin & Taylor, Bradley (eds.) 2013. Valency Patterns Leipzig. Leipzig: Max Planck Institute for Evolutionary Anthropology, <http://valpal.info> [2018-03-08].
- Helbig, G. & Schenkel W. (1968). Wörterbuch zur Valenz und Distribution deutscher Verben, Leipzig: Enzyklopädie.
- Herbst Th. (1999). English valency structures - a first sketch. In *Erfurt Electronic Studies in English*, 6, <http://web-doc.gwdg.de/edoc/ia/eese/rahmen22.html> [03/07/2018].
- Herbst, Th. & Heath, D. & Roe, Ian F. & Götz, D. 2004. A Valency Dictionary of English. A Corpus-based Analysis of the Complementation Patterns of English Verbs, Nouns, and Adjectives. Berlin – New York: Mouton de Gruyter.
- Johnson, C. & Fillmore, C. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, Seattle WA, pp. 56-62.
- Kettnerová, V., Lopatková, M. & Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In *Proceedings of the 15th Euralex International Congress 2012*, 7-11 August 2012, University of Oslo, Norway, pp. 434-443.
- Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. In *Computational Linguistics*, 29(3), pp. 333-347.
- Kipper, K. & Dang, H. T. & Palmer, M. (2000). Class-based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, 30 July-3 August, pp. 691-696.
- Levin, B. (1993). English Verb Classes and Alternations. A Preliminary Investigation. Chicago and London: The University of Chicago Press.
- Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F. & Świdziński M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, 26-31 May 2014, pp. 2785–2792.
- Žabokrtský, Z. & Lopatkova, M. (2007). Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. In *The Prague Bulletin of Mathematical Linguistics*, 87, pp. 41-60.

## A Sample French-Serbian Dictionary Entry based on the *ParCoLab* Parallel Corpus

**Saša Marjanović<sup>1</sup>, Dejan Stosic<sup>2</sup>, Aleksandra Miletic<sup>2</sup>**

<sup>1</sup>Faculty of Philology, University of Belgrade, <sup>2</sup>CLLE, Université de Toulouse, CNRS, Toulouse

E-mail: [sasa.marjanovic@fil.bg.ac.rs](mailto:sasa.marjanovic@fil.bg.ac.rs), [dejan.stosic@univ-tlse2.fr](mailto:dejan.stosic@univ-tlse2.fr), [aleksandra.miletic@univ-tlse2.fr](mailto:aleksandra.miletic@univ-tlse2.fr)

### Abstract

It has already been shown in the state-of-the-art in lexicography that the bilingual dictionary making process can be improved by relying on parallel corpora. The aim of this paper is to present such an application of the *ParCoLab* parallel corpus, a searchable trilingual 11 million token electronic database of aligned texts in French, Serbian and English, developed at the University of Toulouse (France) in cooperation with the University of Belgrade (Serbia). In this paper, we first point out the shortcomings of the leading general French-Serbian dictionaries, which were made using traditional lexicographic methods. We pay special attention to the treatment of the equivalents offered. Taking the case of the French adjective *sale* 'dirty' as an example, we show that the *ParCoLab* parallel corpus makes it possible to: 1) have quick and easy access to meanings missing from the existing dictionaries and to corresponding equivalents; 2) find new equivalents that are not included in any of the existing dictionaries, and which are in some cases the most common translation solutions; 3) order equivalents by their relative corpus frequency; and 4) disambiguate different usages through adequate contextual examples. The solutions we offer are shaped into a sample dictionary entry.

**Keywords:** parallel corpus, lexicography, dictionary, bilingual, French, Serbian

### 1 Introduction

General and phraseological Serbian-French dictionaries have already been examined by Marjanović (2013a, b) taking the example of animal metaphors and similes processing as an example. In both cases, the author underscored the incoherent selection of content, the unsystematic processing and inadequate sense representation, poor or lack of illustrative material and the lack of guidelines on the use of entries and their equivalents. However, only one paper has been published so far on the existing French-Serbian dictionaries (Stanojević-Knežević 2005), which is in fact a bibliography of French-Serbian lexicography. The paper suggests that French-Serbian dictionaries are outdated, incomplete and full of mistakes. However, there is no mention of whether the effectiveness of existing French-Serbian dictionaries has been tested on real texts, as Bujas (1975) did in Yugoslav lexicography for the English-Serbo-Croatian dictionary that he was editing. He gave his students the task of examining how well the equivalents from the latest edition could be applied to translating selected newspaper and magazine articles. The analyzers read the assigned texts thoroughly, checked every word in the dictionary and carefully annotated their findings. After taking all the students' proposals into consideration, Bujas introduced 2,200 new items into the new edition of the dictionary and concluded that the verification of the effectiveness of a bilingual dictionary on contemporary texts can significantly improve the quality of the dictionary (cf. Bujas 1975: 204). Luckily, such manual verification of effectiveness is not necessary anymore: owing to the advent of information technologies, this can be accomplished using electronic parallel corpora (cf. Hartmann 1994: 291-292).

## 2 Parallel Corpora and Bilingual Lexicography

Two types of electronic corpora fall under the term ‘parallel corpus’ – translation corpora and comparable corpora (Atkins & Rundell 2008: 476). However, this term is mostly used to signify translation corpora in the metalexicographic literature, and it will also be used in this sense throughout this paper. A translation parallel corpus is an electronic database that contains texts written in one language and their translations into one or more languages. The corresponding texts are paired at the sentence level. Such a corpus, among other things, makes it possible to find translation equivalents and contexts in which they are used swiftly and easily by comparing aligned pairs of texts (cf. Atkins & Rundell 2008: 478). This is why one of the applications of parallel corpora is in the bilingual dictionary making process. It has been unambiguously proved on specific examples that parallel corpora, regardless of the language pair, can lead to a set of good equivalents, usable in lexicography, which are not listed in existing dictionaries (see, among others, Hartmann 1994; Roberts 1996; Roberts & Montgomery 1996; Dickens & Salkie 1996; Teubert 2002; Citron & Widmann 2006; Salkie 2008; Goossens 2012; Perdek 2012; Perko & Mezeg 2012; Zavaglia & Galafacci 2014). Furthermore, it has been shown that parallel corpora are useful when paired bilingual texts, which both represent the translation of the same source text in a third (so-called *pivot*) language, are compared with each other (e.g. the French-German corpus of Plato’s *Republic*, in Teubert 2002). Considering that the use of a parallel corpus can help the lexicographer gain insight into which equivalents are used in translation, and in which contexts, these resources can largely compensate for the lexicographer’s intuition and increase the objectivity of lexicographic work.

However, while the usefulness of parallel corpora is not called into question in the metalexicographic community, for a long time there has been no mention of dictionaries based on parallel corpora in applied lexicography. Atkins and Rundell’s (2008: 477) survey on the EURALEX discussion list as to whether there is a single dictionary publisher who uses a parallel corpus brought no new findings. The only dictionary systematically based on a parallel corpus was the *Bilingual Canadian Dictionary* (Roberts 1996; Roberts & Montgomery 1996), unfortunately unfinished. Furthermore, the parallel corpus was only “used at the end of the translation stage to ensure that no good equivalents have been missed” (Roberts & Montgomery 1996: 460).

Some of the reasons for this situation include the poor availability of parallel corpora for most language pairs, the high cost of creating a new parallel corpus, the inadequate size of existing parallel corpora, and the unreliability of the translations they include (cf. Salkie 2008). Whereas little can be done about the first three problems, when it comes to the last reason, it has been shown that even bad translations can serve a purpose in the dictionary making process (cf. Marjanović 2017: 487–492): the frequency of incorrect translations in certain language units indicates problematic points, to which lexicographers should pay more attention. Atkins and Rundell (2008: 478) also stress that parallel corpora offer too much evidence, which has then to be carefully considered by the lexicographer. A detailed examination of all corpus findings slows the lexicographer down, which is not profitable in commercial lexicography. This obstacle, however, is considerably alleviated by new technologies that contribute to the automation of lexicographic work. For all these reasons parallel corpora are gaining steam in modern e-lexicography (see Héja 2010; Lindemann 2013; Lindemann et al. 2014; Škrabal & Vavřín 2017).

The goal of this paper is to show how existing lexicographic processing, and especially the processing of equivalents in French-Serbian dictionaries, can be improved by relying on a new language resource, the *ParCoLab* parallel corpus. First, we provide an overview of existing traditional French-Serbian dictionaries on a specific example and point to their shortcomings (Section 3); next, in Section 4, we describe the *ParCoLab* corpus, and finally, in Section 5, we present the results of the analysis of the



corpus-based equivalents and their lexicographic usability. Section 6 then summarizes the findings and offers a sample dictionary entry based on the results of the analysis.

### 3 The Big Four: French-Serbian Dictionaries and Their Shortcomings

In order to check the quality and to address the shortcomings of the existing French-Serbian (Serbo-Croatian, Croatian) dictionaries, in this section, we will illustrate and discuss in detail the processing of equivalents on the example of a common French entry *sale* ‘dirty’ in the following four leading general bilingual dictionaries: Marković (1980), Jovanović (1991), Putanec (1995) and Točanac et al. (2017). We chose this entry because we believe it to be a representative sample: it is polysemous, its correspondent in the primary sense (*prljav* ‘dirty’) belongs to a wide synonym set at the systemic level, and, depending on the context, it can be translated with a set of equivalents. By analyzing this entry we can gain insight into how lexicographers order different senses, how they discriminate them and how they list multiple equivalents. The excerpt from the four aforementioned dictionaries looks as follows:

**sale** [sal] *adj.* prljav, nečist; gadan, odvratan; pokvaren || *c’est une affaire sale* to je prljav posao; *c’est un caractère sale* odvratan karakter; *linge sale* prljavo rublje; *avoir les mains sales* imati prljave ruke; *être sale comme un porc, comme un peigne* biti vrlo prljav (Marković 1980)

**sale** [sal] *a.* (posle imenice) prljav, nečist, prašnjav, blatnjav, neuredan, odvratan; neprijatan, nepristojan; (pre imenice) *fam.* pokvaren; gadan (Jovanović 1991)

**sale** [sal] *a.* zamazan, nečist, prljav, gadan; osiromašen (o rudi); mutan (o boji); nesiguran (o obali); ružan, nepristojan, pokvaren, gadan (Putanec 1995)

**sale** [sal] *adj.* 1. prljav, nečist, neuredan • *mains* ~s prljave ruke • *histoire* ~ prljava priča • *argent* ~ prljav novac (od šverca, ...) 2. *fam.* grozan, užasan, odvratan, gadan • ~ *temps* grozno vreme, *fig.* loši dani • ~ *coup* nizak, težak udarac • ~ *type* odvratan tip (Točanac et al. 2017)

All four dictionaries are made using traditional lexicographic methods, mostly via bilingual adjustment of French monolingual dictionaries. In other words, lexicographers read the definitions and examples in French dictionaries and noted equivalents they could recall. Therefore, we base our analysis on the senses of the French word *sale* represented in the *Le Petit Robert* dictionary, which was the one that all lexicographers used, according to their respective bibliographies. In order to conduct our analysis, we established eight different senses: (1) ‘covered or marked with an unclean substance’, (2) ‘(of a colour) not bright or pure’, (3) ‘immoral, or dishonest’, (4) ‘concerned with sex in a lewd or obscene way’, (5) ‘(of money) obtained through illegal or disreputable means’, (6) ‘used to emphasize how bad something is’, (7) ‘used to emphasize one’s disgust for something’, and (8) ‘used to emphasize one’s disgust for someone’. Table 1 shows which of these senses are present in the four dictionaries:

Table 1: Distribution of senses in the analyzed French-Serbian dictionaries

Dictionary / Sense	1	2	3/4	5	6	7	8
Marković 1980	+	-	+	-	+	-	-
Jovanović 1991	+	-	+	-	+	-	+
Putanec 1995	+	+	+	-	-	-	-
Točanac et al. 2017	+	-	+	+	+	+	+

Two facts can be noticed immediately: only two senses are present in all four dictionaries, and the largest dictionary (Putanec 1995) processes only the first three senses. However, two other senses were introduced into Putanec’s dictionary, with the equivalents *osiromašen* and *nesiguran*, but these

two senses are too technical, so they should not have been processed in a general dictionary, especially not when common senses under items 5-8 from general language were omitted. The most comprehensive dictionary is Točanac et al. (2017), but the senses are not well discriminated there, which is something we will address later on.

By reviewing the excerpts, we can also note that senses in the first three dictionaries are listed in a linear manner and that they are separated by a semi-colon, while in the fourth one, the senses are explicitly separated by numbers. Marković introduces senses 3 and 6 with “examples”<sup>1</sup> (*c’est une affaire sale* and *c’est un caractère sale*), while Točanac et al. (2017) do the same for senses 4, 5, 7 and 8: *histoire sale*, *argent sale*, *sale coup* and *sale type*. In the first dictionary, the examples are always given in a separate block, while in the second, they are listed within the senses they refer to. The latter implies: 1) that the sense introduced by non-contextual equivalents must be represented in the examples, or 2) if more senses are grouped in one because they share the same equivalent, the equivalent can be used in all examples. This means that the equivalents “grozan, užasan, odvratan, gadan” from the second sense are potentially interchangeable in the “examples” (e.g. *grozno* for *užasno*, *odvratno*, *gadno vreme*), which justifies grouping several equivalents. However, such replacements are not possible in all “examples” in the first sense: (*prljave / nečiste / neuredne ruke*, but *\*nečista / \*neuredna priča*, *\*nečist / \*neuredan novac*). This inconsistent approach and overlapping are justified to a certain extent by the target group, that is, Serbian users who intuitively know that the equivalent *prljav* in examples *prljave ruke*, *prljava priča*, *prljav novac* – and, consequently, the adjective *sale* in source collocations – does not have the same meaning. However, native French speakers use these dictionaries as well, and this would not be obvious to them.

When it comes to explicit sense indicators, we note that only Putanec (1995) – with three secondary senses, but not the fourth – and Točanac et al. (2017) – with the “example” *argent sale* – state the sense discriminator. However, it is unclear why the sense discriminator was provided only for that collocation, but not for *histoire sale*. In both cases, the adjective *prljav* is not used in its primary sense, which was introduced through non-contextual equivalents. There is no sense discriminator either in the case when the equivalent is polysemous (cf. *pokvaren* in Marković 1980 and Jovanović 1991). Consequently the user cannot know which sense is the right one.

Furthermore, within a single sense, there are equivalents that introduce a completely different meaning. For instance, in Jovanović (1991), the equivalent *odvratan* ‘repulsive’ and *gadan* ‘disgusting’ in Putanec (1995) are unjustifiably among the set of equivalents that denote the primary sense of the entry *sale*. These mistakes occur in other entries of those dictionaries too. We can also note that close, but different secondary meanings were grouped into the same sense (e.g. in Putanec 1995: *ružan*, *nepristojan*, *pokvaren*, *gadan*, where the first two denote sense 4, while the others denote sense 3 and 6-8 respectively). Although lumping in general is a justified procedure in bilingual lexicography (Adamska-Sałaciak 2006: 76-79), it can only be applied when two or more senses share one equivalent or a set of the same equivalents, which is not the case in our example. Due to the vagueness of meaning and overlapping of senses, we could not separate equivalents for our third and fourth senses with certainty in the dictionaries analyzed, so they were presented together in Table 1.

Interestingly, the order of equivalents in these four dictionaries is not the same, not even with the primary sense: three dictionaries give the equivalent *prljav* in first place, which is also the formal correspondent of the French entry, while in Putanec (1995), it is listed in third place. What is more, it is surprising that the interlingual hyponym *zamazan* is in first place. With regard to formal correspondence, it should be pointed out that this does not allow users to perceive explicitly that the equivalent

1 The term *example* is placed in quotation marks when it is not used in the metalexicographic sense, but in the sense the authors of the four dictionaries used in their forewords.

*prljav* has a polysemantic structure and that it mostly corresponds to the French one, as we will see in sections 5 and 6. Because of that, these dictionaries would not be able to serve as an aid to language acquisition (cf. Tarp 2008: 195-198).

Finally, equivalence itself is understood quite broadly. Often, interlingual hyponyms, co-hyponyms and even interlingual hypernyms are listed, even when it is unjustified. Examples of this can be found in the excerpts from three dictionaries (Jovanović 1991: *prašnjav* ‘dusty’, *blatnjav* ‘muddy’, *neuredan* ‘untidy’; Točanac et al. 2017: *neuredan* ‘untidy’; Putanec 1995: *zamazan* ‘daubed’).

This detailed analysis of the four dictionaries confirms that there are plenty of shortcomings in traditional French-Serbian lexicography with regard to sense representation and processing of equivalents: not all of the common senses were processed; those that were processed are often intertwined and incoherently listed, and they are not systemically discriminated with indicators or examples; equivalence is understood broadly and polysemous equivalents are not repeated even when they are used in different senses. Since we consider the processing of the entry *sale* to be a representative sample, the findings should be understood as a general image of equivalent processing in the analyzed dictionaries, as well as in French-Serbian lexicography in general. This stance is supported by the results of the analysis of similes (Marjanović 2017: 493-551) and prepositions (Stosic et al., forth.) in French-Serbian dictionaries. In Section 5, we will examine if and how the *ParCoLab* parallel corpus, briefly presented in Section 4, can contribute to improving French-Serbian lexicography.

#### 4 A new resource: the PARCOLAB Parallel Corpus

*ParCoLab* is a searchable trilingual database of aligned texts in French, Serbian and English, which has been developed by the research unit CLLE-ERSS (CNRS and the University of Toulouse-Jean Jaurès, France) in cooperation with the Romance Department of the Faculty of Philology, University of Belgrade (Serbia). The corpus is freely available at the following address: <http://parcolab.univ-tlse2.fr/>. *ParCoLab* contains original texts in one of the three aforementioned languages, and their translations in one or both remaining languages. In other words, it is based on three distinct subcorpora, each having a different pivot language. To date, the entire corpus contains 11.1 million tokens. In the past year, the number of tokens has increased by four million (7.1 million mid-2017, see Miletic et al. 2017: 158).

Aside from quantitatively enriching the corpus, it has also been diversified by introducing many different text types. So far, literary texts and their translations make up the largest share of the corpus, because they are the most readily available. However, there are also a certain number of legal and political texts, film subtitles, web content and biology texts. The detailed distribution of tokens according to genre is listed in Table 2:

Table 2: Distribution of tokens per text type and language

Text type	English	French	Serbian	Total tokens
Literary texts	1,919,428	4,257,773	3,495,363	9,672,564
Legal texts	233,556	291,996	79,679	605,231
Web content	229,006	186,256	63,018	478,280
Film subtitles	48,383	125,919	104,935	279,237
Biology	0	40,759	35,113	75,872
Politics	0	9,529	8,576	18,105
TOTAL	2,430,373	4,912,232	3,786,684	11,129,289

The corpus data are stored in an XML format based on the TEI P5 Guidelines. The alignment of the original texts with their translations was performed using an algorithm integrated in *ParCoLab*; no external resources were used for this task. The algorithm proceeded in descending order, creating one-to-one alignments, first at chapter level, then at paragraph level, and finally at sentence level. Errors were signaled by the tool and corrected manually, which guarantees the reliability of the corpus alignments. As of yet, only a small part of the corpus includes morphosyntactic and syntactic annotations, but our current short-term efforts are focused on this task (for further information, see Miletic et al. 2017: 160-162; Miletic et al. 2016; Miletic & Urieli 2017). Queries are carried out via the Elasticsearch search engine, which is well adapted to querying data in NoSQL databases, and a query form offers quiet very good search possibilities.

Regardless of the fact that *ParCoLab* is unbalanced and small compared to large-scale web corpora, it already contains a large number of words pertaining to the core vocabulary, which enables its usefulness for lexicographic purposes in the process of making new general French-Serbian dictionaries to be tested. Our conviction is supported by the authors referred to in Section 2, who proved the lexicographic usefulness of parallel corpora that are even smaller than *ParCoLab*.

## 5 The ParCoLab Corpus Evidence vs Dictionary Evidence

We examined the same French word as in Section 3 in order to compare corpus equivalents with the ones found in the dictionaries. First, we limited the search to original French texts with their Serbian translations. We found 80 occurrences, out of which only two were not from literary texts. They encompassed six out of eight senses, and we found the following equivalents (listed alphabetically): *bezobrazan*, *gadan*, *nečist*, *odvratan*, *podao*, *pokvaren*, *prljav*, *zamrljan*. However, such a short list was unexpected. In order to increase the number of occurrences, and potential equivalents, we expanded the search to French translations of Serbian and English original texts. By doing this, we were able first verify whether, as stated by Citron and Widmann (2006: 255), better translation solutions are found in the source language (L1) when searching through the translated language (L2), that is, in our case, when Serbian equivalents of the French word *sale* (L2) are searched in original Serbian texts (L1). Second, we were able to verify whether, in our case, we can reach interesting translation solutions through a third *pivot language* (cf. Teubert 2002).

The content we searched through included the French-Serbian, the Serbian-French and the English-French-Serbian subcorpora, and contained approximately four million tokens, from which we extracted 277 occurrences. In Table 3, we list a detailed distribution of occurrences according to sense, text type and original language. In five cases, we could not ascribe a single sense to the analyzed lexeme (due to the unclear translation), so we listed those cases in a special column, represented by “?”:

Table 3: Distribution of occurrences per sense, text type and language

Original Language	Text type	Senses								
		1	2	3	4	5	6	7	8	?
English	Literary texts	27	1	0	0	0	4	0	1	3
	Film subtitles	2	0	0	1	0	0	0	0	0
French	Literary texts	49	3	1	0	0	4	5	15	1
	Film subtitles	1	0	1	0	0	0	0	0	0
Serbian	Literary texts	119	5	1	0	0	5	2	7	1
	Film subtitles	0	0	0	0	0	0	0	18	0
TOTAL		198	9	3	1	0	13	7	41	5



The numbers from Table 3 show that seven out of eight isolated senses of the lexeme *sale* are attested in the corpus and that, in its current version, the database can represent a polysemantic structure of the lexeme *sale*. Also, the number of different equivalents has been increased. We evaluated the relevance of their inclusion in a French-Serbian dictionary by annotating them according to the level of *lexicographic potential* (LP) (Perdek 2012: 382-386)<sup>2</sup>. Following Perdek (2012), we established four levels of LP – *high*, *medium*, *low* and *zero* LP. *Zero* LP includes incorrect equivalents, while *high* LP includes ideal equivalents which can be used in most common contexts of a certain sense. *Medium* (cf. *average*, in Perdek 2012) encompasses good equivalents which may enter the dictionary, but need to be indicated in a proper way, because they are limited to specific contexts or require translation transformations. *Low* refers to good equivalents which correspond to a single context.

Out of the total number of equivalents, 230 (83%) are characterized by a high and medium level of LP, which is a quite encouraging result for using *ParCoLab* for lexicographic purposes. Through English, we reached good equivalents such as *mastan*, *neopran*, *običan*, *zamašćen*. Since there are far more equivalents from Serbian original text, they will be presented later, especially under the last sense (cf. below). In the following section, we will thoroughly analyze all the translation equivalents in order to pinpoint the specific features of each sense. This will allow us to more easily compare corpus equivalents with existing lexicographic equivalents.

## 5.1 Corpus Findings Per Sense

The primary sense of the lexeme *sale* is, as one would expect, the most common in *ParCoLab*: out of 198 tokens, the equivalent *prljav* is listed in 161 cases, which means that its LP is undoubtedly high. This corroborates our observation that the solution from Putanec (1995) is poor (cf. Section 3). Furthermore, if we disregard 10 cases when there was no translation, or when it was wrong, there are a dozen more equivalents for the primary sense in *ParCoLab* (see Table 4). The lexicographically marked equivalents are in italics in the table. However, in the corpus, there are no occurrences of the lexicographic equivalents *blatnjav* and *neuredan*.

Table 4: Distribution of the first sense equivalents per LP

LP	Equivalents
High LP	<i>prljav</i> (161), <i>nečist</i> (4x), <i>neopran</i> (3x), <i>mastan</i> (2)
Medium LP	<i>zaprljan</i> (3x), <i>uprljan</i> , <i>umašćen</i> , <i>izmašćen</i> , <i>zamazan</i> , <i>zamrljan</i>
Low LP	<i>garav</i> , <i>prašnjav</i> , <i>pun prljavštine</i> , <i>zagađen</i> (2x)

Such a rich set of translation equivalents can help the lexicographer who is working on a French-Serbian dictionary to select the corresponding equivalents much more swiftly and easily, depending on the target group, its needs and the situations in which the dictionary will be used.

Only Putanec (1995) lists the second meaning (‘of a color, not bright or pure’) and processes it with the equivalent *mutan*, which is a co-hyponym of the French entry. For this meaning, we found the equivalent *prljav* seven times in *ParCoLab* (e.g. *prljavobela*), while the equivalents *siv* and *gadan* appeared once each (e.g. *žutosiva*). The first equivalent (*prljav*) has a high LP, the second (*siv*) a medium one, because it is used only with certain color adjectives (e.g. white, yellow, green, blue). Both equivalents are part of compounds according to orthography norms, and their use needs to be carefully illustrated in the dictionary. The third equivalent (*gadan*) is an example of a bad equivalent with zero LP.

<sup>2</sup> See also *OK*, *fuzzy* and *false*, in Lindemann *et al.* (2014), as well as the notion of Basic vs Rich Translation Equivalence, in Dickens and Salkie (1996).



The third meaning ('immoral' or 'dishonest') is illustrated by three citations; in two of them, the equivalent is *prljav* (1)-(2), and in one, it is *kvaran* (3):

- |  |   |
|--|---|
| (1) Pour elle, le peuple est quelque chose de <i>sale</i> et de rusé, et Prerovo, un repaire de loups. | Narod — to je njoj nešto <i>prljavo</i> i lukavo. Prerovo — vučja jazbina!            |
| (2) C'est lâche, c'est <i>sale</i> , et petit, et commun de calomnier une femme !                      | To je kukavički, to je <i>prljavo</i> , i sitničarski, i nisko, klevetati jednu ženu! |
| (3) C'était <i>sale</i> , mais bien joué.  | Bilo je <i>kvarno</i> , ali se dobro izvukao.   |

In our opinion, the equivalents are adequate only in the first and third citations. In the second, it would have been better if the equivalent *pokvaren* had been used, as suggested by Marković (1980), Jovanović (1991) and Putanec (1995) in their dictionaries. It would also correspond to the third translation situation: in fact, since he cheated the girl by lying to her, the young man agrees that the means was immoral, but that the goal has been achieved. Neither of the corpus equivalents is listed in the dictionaries analyzed, which is disappointing since the first equivalent has the same polysemantic structure as the entry (cf. Section 3). Putanec (1995) offers the equivalent *gadan* with *pokvaren*, which is unjustified, since these two lexemes are not interchangeable.

For the fourth sense ('concerned with sex in a lewd or obscene way'), we found one citation in *ParCoLab*:

- |  |  |
|--|--|
| (4) Ce mec m'envoyait des trucs très sexuels et sales et il y avait une fille dans son avatar. | Jedan momak mi je slao hiperseksualne, <i>gadne</i> stvari, a na avataru mu je bila devojka. |
|--|--|

The translation equivalent proposed (*gadan*) has a medium LP. In other words, out of context, it would not refer to the adequate meaning, so it should be contextualized in the dictionary. What is more, the best option would be to structurally transform *gadna stvar* into the morphologically related noun *gadost*. When it comes to dictionaries, only Točanac et al. (2017) processes this sense and lists the equivalent *prljav* in the collocation *histoire sale*. This equivalent has a high LP, but, aside from it, with the same collocation, the adjective *mastan* is more natural. The equivalents *nepristojan* (Jovanović 1991; Putanec 1995), as well as *ružan* (Putanec 1995) correspond to the same collocation. Jovanović's equivalent *neprijatan* is only an interpretation of the sense, so it can be considered that its LP is zero.

The citation for the fifth sense ('of money, obtained through illegal or disreputable means') was not found in *ParCoLab*. We assume that it could be found in newspaper articles, which have yet to be included in the *ParCoLab*. Among the four dictionaries, only Točanac et al. (2017) lists the collocation *argent sale* and its corresponding equivalent *prljav novac*.

For the sixth meaning ('used to emphasize how bad something is'), we have found 13 occurrences in the corpus. In one case, the translation is not adequate, while in six cases, the equivalent *gadan* was used (e.g. 5, 6), and in one case – *prljav* (7):

- |  |  |
|--|--|
| (5) Oh, il a vécu un <i>sale</i> moment.   | O, taj je preživeo <i>gadan</i> trenutak.  |
| (6) Alors Athénaïs vomit les plus sales injures, les invectives les plus obscènes sur les magistrats et les grenadiers [...].                                | Tada Atenaida stade da izbacuje najgadnije psovke i najbestidnije pogrde na činovnike i grenadire [...].                                       |
| (7) Pour ne pas ébruiter une si <i>sale</i> affaire, car je suis dans l'impossibilité de justifier la conduite de mon père, je vous écris au dernier moment. | Da se ne bi raščula jedna tako <i>prljava</i> stvar, jer ja sam u nemogućnosti da opravdam postupak svog oca, pišem vam u poslednjem trenutku. |

A larger number of occurrences of the first equivalent can be ascribed to the text type from which the citations stem. In ordinary Serbian language, especially in example 6, the synonymous lexemes *odvratan*, *grozan*, *ružan*, *užasan* would also be used in that sense: Marković (1980) lists two equivalents

(*gadan* and *odvratan*), Točanac et al. (2017) offers four (*grozan*, *užasan*, *odvratan*, *gadan*). We cannot deduce further guidelines on the order of the equivalents based on corpus information, because the number of occurrences is small.<sup>3</sup>

There are seven occurrences of the seventh sense ('used to emphasize one's disgust for something'). In three occurrences, the equivalent is *prljav* (in one case, the translation is incorrect), and *lopovski*, *podao* and *gadan* occur once (e.g. 8, 9, 10 and 11, respectively):

- |  |  |
|--|--|
| (8) Je ne veux même pas prononcer son <i>sale</i> nom de Shiptar.  | Neću ni da izgovorim njegovo <i>prljavo</i> šiptarsko ime.   |
| (9) Vous l'avez érigé dans votre <i>sale</i> patelin.  | Podigli ste ga u vašem <i>lopovskom</i> mestu.   |
| (10) [...] et méprisant la guillotine de 89 comme une <i>sale</i> vengeance.   | [...] a puno preziranja prema gijotini iz 89 kao prema jednoj <i>podloj</i> osveti.                      |
| (11) [...] regretter de [...] n'avoir à lui offrir qu'une <i>sale</i> soutane de prêtre dont elle aura peur et dégoût! | [...] žaliti što [...] moći joj ponuditi samo <i>gadnu</i> sveštenučku mantiju koje se ona boji i gnuša! |

Alongside these, we excerpted one quite interesting translation solution: in one citation, the adjective *sale* has been left out from the translation, but an expressive lexeme was used for the noun with which it stands in original:

- |  |   |
|--|---|
| (12) j'ai quitté la <i>sale</i> baraque à Deneulin, je descends demain au Voreux avec douze Belges | Ostavio sam onu Denelenovu <i>straćaru</i> , silazim sutra u Vore sa dvanaestoricom Belgijanaca |
|--|---|

None of the corpus equivalents are lexicographically processed, since the seventh sense was treated only in Točanac et al. (2017) within the example where the equivalents *nizak* and *težak* were provided for *sale*, which is a justifiable lexicographic solution. Unlike the previous ones, it is more difficult to present the equivalents for this sense in the dictionary, because the choice of equivalents depends on the context: it is, therefore, necessary to present it with examples.

The final sense ('used to emphasize one's disgust for someone') was only processed in Točanac et al. (2017) as the example of *sale* type with the equivalent *odvratan* tip, and there are 41 citations in *ParCoLab* for it. If we put aside the two citations with a non-existent translation and the one with an incorrect translation, there are numerous equivalents left. More than half of the citations belong to the literary text type; in them, we have identified the following equivalents: *gadan* (8x), *prljav* (4x), *odvratan* (2x), *smrdljiv*, *pogan*, *bezobrazan*, *običan*. By analyzing each individual case, we come to the conclusion that the equivalents that occur more than once have either medium or low LP, while others have either a high or medium LP. We will list the examples of good translation solutions.

- |  |   |
|--|---|
| (13) Te voilà collé au mur, <i>sale</i> crapule !  | Sad si ti sabijen uza zid, <i>gadna</i> huljo!  |
| (14) <i>Sales</i> youpins, [...] vous avez crucifié mon Dieu et vous voulez ma peau ;  | <i>Prljavi</i> gadovi, [...] razapeli ste moga boga i sad hoćete moju kožu;             |
| (15) J'ai vu Mouquet, tu vas encore au Volcan, où il y a ces <i>sales</i> femmes de chanteuses.                                | Videla sam Mukea, ideš opet u »Vulkan«, gde su one <i>odvratne</i> pevačice.            |
| (16) Jusqu'à présent, c'était du gâteau, <i>sales</i> Youpins, mais c'est fini.  | Dosad vam je bilo lepo, <i>smrdljivi</i> Čivuti, ali sad je tome kraj.                  |
| (17) « Tu chantes, <i>sale</i> petite souris ! » Il lui serre le cou, le secoue et cherche à lui briser la tête contre le mur. | Pevaš, mišiću <i>pogani</i> ! — steže mu vrat, trese ga i hoće glavu o zid da mu slupa. |

3 With regard to this sense, we have to point out that in three corpus citations, the idioms *être/se trouver dans une sale passe* and *être dans un sale pétrin* occur with good equivalents. In one case, *sale* is used in the translation of the Serbian idiom: *Nisu mi čista posla* → *Ça m'a l'air d'une sale affaire*. Since phraseology is not the topic of this paper, we will not expand on this any further.

- (18) Partie devant eux, la Mouquette s'exclamait dans l'escalier noir, en les traitant de *sales* mioches et en menaçant de les gifler, s'ils la pinçaient. Muketa, koja je pošla ispred njih, vikala je niz mračne stepenice, nazivala ih *bezobraznom* dečurlijom i pretila da će ih išamarati ako je budu štipali.

We also found equivalents where structural transformations occurred: *jarac*, *poganija* (both with low LP) and *gad*. The lexeme *gad* is interesting because it is precisely the equivalent that could have been used for the example *sale type* in Točanac et al. (2017):

- (19) il [...] cria qu'il ferait se repentir un jour le *sale monde* qui manquait de reconnaissance [...]]. vikao [je] da će se jednoga dana pokajati ti *gadovi* koji ne znaju za zahvalnost [...].

In film subtitles, whose source language is Serbian, we found six different equivalents, which all are good, but have a medium LP. Those are *najobičniji* (2x) and *prljav*, and then equivalents where structural transformations of the type Adj + N → N + Adj occurred: *pička* (4x), *đubre [jedno]* (3x) and *stoka*. Aside from that, in two cases from films, and two cases from literary texts, there are no adjectives in Serbian, we only find the noun with the pejorative suffix (cf. example 25). We will list a few examples of good translation solutions:

- |   |  |
|---|--|
| (20) Lâche-moi, <i>sale</i> skinhead !        | Pusti me, <i>pičko</i> ćelava!                   |
| (21) Tu ferais quoi ? <i>Sale</i> menteur !   | Šta bi mi pokazao, <i>đubre jedno</i> lažljivo!  |
| (22) T'es mort, <i>sale</i> Tchétvik !        | Krvavu ti nedjelju jebem, <i>stoko</i> četnička! |
| (23) Violeta est une <i>sale</i> pute.        | Violeta je <i>najobičnija</i> kurva.             |
| (24) De <i>sales</i> capitalistes.            | <i>Prljavi</i> kapitalisti.                      |
| (25) Qu'est-ce qu'il y a ? <i>Sale</i> pute ! | Šta je, šta je, <i>kurvetino</i> ?               |

Considering that this common sense has not been processed in French-Serbian dictionaries, all the listed solutions from *ParCoLab* represent valuable content for identifying new equivalents. With this sense, we also see a drastic difference between text types, which justifies the introduction of film subtitles in the *ParCoLab* database. Furthermore, all the equivalents from films are located in original Serbian texts, so we can confirm the aforementioned claims by Citron and Widmann (2006) that equivalent processing can be improved by searching equivalents in source language texts.

Seeing as *sale* is part of the core vocabulary and that it is relatively frequent, we can assume that the result will be approximately the same with other frequent and common words with a similar profile (cf. Marjanović et al., forth.). The same can also be expected with a majority of highly frequent grammatical items (cf. Stosic et al., forth.). Let us recall once again that it has been demonstrated in the metalexigraphic literature (see references in Section 2) that corpora of a narrower scope than *ParCoLab* can make a significant contribution to the dictionary making process. Therefore, the *ParCoLab* parallel corpus can already help lexicographers to verify the effectiveness of listed lexicographic equivalents and/or extraction of translation equivalents in general in their work on medium-size dictionaries, such as the dictionaries we analyzed in section 3.

## 6 The Sample French-Serbian PARCOLAB-based Dictionary Entry

Based on the results of the analysis of existing dictionaries (cf. Section 3) and bearing in mind the equivalents with a high and medium LP that we extracted from the *ParCoLab* parallel corpus, we offer a sample corpus-based dictionary entry, that satisfies both the reception and L2→L1 translation needs of native speakers of Serbian and the production, and L1→L2 translation needs of French users. The sample is given in Figure 1:

**SALE** /sal/ *adjectif*

1. (pas propre) **prljav**, **nečist**, [personne] **musav**; **maines sales** prljave ruke; **vaisselle sale** prljavi **ou** neoprani sudovi; **il a les cheveux sales** prljava **ou** masna mu je kosa
2. [couleur] **prljav**; **blanc sale** prljavob[ij]elo; **jaune sale** sivožut, žutosiv
3. (immoral) **prljav**, **pokvaren**, **kvaran**
3. (obscène) **prljav**, **nepristojan**, **bezobrazan**; **histoires sales** masne priče; **dire des choses sales** govoriti gadosti **ou** perverzije
4. [argent] **prljav**
5. (désagréable) [temps, habitude, affaire etc.] **odvratan**, **gadan**, **grozan**, **užasan**; **quel sale temps !** kako je vr[ij]eme odvratno **ou** ružno!
6. **PÉJ. proklet**, **smrdljiv**, **VULG. jeben**; **cette sale voiture** ta prokleta kola; **sale capitaliste** prljavi kapitalist[a], ⇔ **đubre kapitalističko**; **sale menteur !** ⇔ **đubre lažljivo!**, ⇔ **VULG. pičko lažljiva!**, *(souvent traduit par un nom péjoratif de sens augmentatif)* lažovčino!

Figure 1: A sample French-Serbian corpus-based dictionary entry

The sample carefully presents the polysemy of the entry, discerns its senses based on contextual information found in *ParCoLab*, and offers sets of corpus-based equivalents, which are ordered according to the number of occurrences. We provide the information on the use of equivalents and illustrate them contextually. In several cases, especially within the last sense, we also added equivalents found through association by reading the citations from *ParCoLab* and their translations (e.g. *proklet* ‘damned’, *jeben* ‘fucking’). However, this should not be seen as a deviation of the corpus approach to lexicography because, as stressed in the literature (cf. Roberts & Montgomery 1996; Lindemann 2013: 252), corpus content is the raw material that lexicographers need to adjust to the type of dictionary they are working on.

## 7 Concluding remarks

The existing French-Serbian lexicography, based exclusively on traditional methods, suffers from a number of shortcomings, as demonstrated through the analysis of the entry *sale* in the four leading French-Serbian dictionaries. The most prominent issue concerns the unsystematic equivalent processing, as well as the lack of authentic illustrative material and sense discriminators. In this paper, we have shown that by relying on the 11.1M French-Serbian-English parallel corpus *ParCoLab*, these shortcomings can be remedied or at least lessened to a large extent. Based on the subcorpus of aligned French and Serbian original and translated texts of approximately four million words, we extracted 277 occurrences of the French adjective *sale* and their Serbian translations. We classified these occurrences based on their sense. This allowed us to establish that *ParCoLab* currently contains seven out of the eight senses. Based on the extracted material, we listed the equivalents for every sense. The results of this paper indicate that using *ParCoLab* can lead to a set of equivalents, a large number of which are not included in the existing dictionaries. In some cases, those equivalents were the most common translation solutions. We have also shown that the overwhelming majority of equivalents (around 83%) have a high or medium LP. Based on the results of the analysis, we have offered a sample entry of the future French-Serbian *ParCoLab* based dictionary. Thus, we have shown that the content of the parallel corpus *ParCoLab* in its current scope can contribute to improving the existing French-Serbian dictionaries. That is the main purpose of this paper.

However, we must bear in mind that the number of extracted equivalents would have been higher if the corpus had been larger and more diversified in terms of genre, which is something the *ParCoLab*



research team is working on. At the same time, we need to mention that the methodology applied in this paper is based on manual equivalent extraction from *ParCoLab*. While quite reliable, such a method is very time-consuming; therefore, we are aware that the extraction process should be automated in the work on specific commercial dictionaries. For this reason, it is necessary to develop tools for automatic equivalent extraction, such as *The Translation Equivalents Database* (Treq), developed at the Institute of the Czech National Corpus (cf. Škrabal & Vavřín 2017), or the *Bilingual Word Sketches*, developed within *The Sketch Engine* tool (cf. Baisa et al. 2014), which quantify pairs of extracted equivalents in terms of their relative and absolute frequency respectively. Considering that texts in *ParCoLab* will be lemmatized, and morphosyntactically and syntactically annotated in the near future (cf. also Miletic 2018), the next phase is to align texts at word-level, which will enable us to develop the application for automatic *ParCoLab* equivalent extraction. Such a tool would contribute to making the described methodology completely applicable in the work on good commercial French-Serbian corpus-based dictionaries.

## References

- Adamska-Sałaciak, A (2006). *Meaning and the Bilingual Dictionary: The Case of English and Polish*. Frankfurt am Main: Peter Lang.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Baisa, V., Jakubiček, M., Kilgariff, A., Kovář, V. & Rychlý, P (2014). Bilingual word sketches: the translate button. In A. Abel et al. (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 505-513.
- Bujas, Ž. (1975). Testing the Performance of a Bilingual Dictionary on Topical Current Texts. In *Studia Romanica et Anglicana Zagradiensia*, 39, pp. 193-204.
- Citron S., Widmann T. (2006). A Bilingual Corpus for Lexicographers. In E. Corino et al. (eds.) *Proceedings of the 12th EURALEX International Congress*. Torino: Edizioni dell'Orso, pp. 251-255.
- Dickens, A., Salkie, R. (1996). Comparing Bilingual Dictionaries with a Parallel Corpus. In M. Gellerstam et al. (eds.) *Proceedings of the 7th EURALEX International Congress*, Göteborg: Göteborg University, pp. 551-559.
- Goossens, D. (2012). Translation equivalents in translation corpora and bilingual dictionaries: the case of approximators in English and French. In: R. Vatvedt Fjeld, J.M. Torjusén (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, pp. 514-522.
- Hartmann, R.R.K. (1994). The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography. In W. Martin et al. (eds.) *Proceedings of the 5th EURALEX International Congress*. Amsterdam: Vrije Universiteit, pp. 291-297.
- Héja, E. (2010). The Role of Parallel Corpora in Bilingual Lexicography. In: N. Calzolari et al. (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valetta: European Language Resources Association (ELRA), pp. 2798-2805.
- Jovanović, S.A. (1991). *Srpskohrvatsko-francuski rečnik*. Beograd: Prosveta.
- Le Petit Robert de la langue française*. Accessed at: <https://www.lerobert.com>. [05/03/2018]
- Lindemann, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. In *Procedia – Social and Behavioral Sciences*, 95, pp. 249-257.
- Lindemann, D., Manterola, I., Nazar, R., San Vicente, I. & Saralegi, X. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density language pair. In A. Abel et al. (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 563-576.
- Marjanović, S. (2013a). Le traitement des métaphores animalières dans les dictionnaires serbe-français. In *Godišnjak Filozofskog fakulteta*, 38(3), pp. 117-128.
- Marjanović, S. (2013b). Obrada poredbenih frazema u srpsko-francuskim rečnicima. In S. Gudurić, M. Stefanović (eds) *Jezici i kulture u vremenu i prostoru*, 2 (2). Novi Sad: Filozofski fakultet, pp. 255-268.
- Marjanović, S. (2017). *Poredbene frazeme s komponentom comme/kao u francuskom i srpskom jeziku*. PhD thesis. Faculty of Philology, University of Belgrade, Belgrade, Serbia.



- Marjanović, S., Stošić, D. & Miletic, A. (forth.). Paralelni korpus *ParCoLab* u službi srpsko-francuske leksikografije. In J. Novaković, M. Srebro (eds.) *Srpsko-francuske književne i kulturne veze u evropskom kontekstu*. Novi Sad: Matica srpska.
- Marković, R. (1980). *Francusko-srpskohrvatski rečnik*. Beograd: BIGZ.
- Miletic A., Stosic D. & Marjanović, S. (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. In K. Ekštejn, V. Matoušek (eds) *Text, Speech, and Dialogue. TSD 2017*. Lecture Notes in Computer Science, vol. 10415. Cham: Springer, pp. 156-164.
- Miletic A., Urieli, A. (2017). Non-projectivity in Serbian: Analysis of Formal and Linguistic Properties. In S. Montemagni, J. Nivre (eds.) *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, Linköping University Electronic Press, pp. 135–144.
- Miletic, A. (2018). *Un treebank pour le serbe : constitution et exploitations*. PhD thesis. University of Toulouse - Jean Jaurès, Toulouse, France.
- Miletic, A., Fabre, C. & Stosic, D. (2016). Mise au point d'une méthode d'annotation morphosyntaxique fine du serbe. *Conférence conjointe JEP-TALN-RECITAL 2016*, Paris, pp. 506-513. Accessed at: <https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/V02-TALN.pdf>. [05/11/2017]
- Perdek, M. (2012). Lexicographic potential of corpus-derived equivalents. The case of English phrasal verbs and their Polish equivalents. In: R. Vatvedt Fjeld and J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, pp. 376-388.
- Perko G., Mezeg, A. (2012). Uporaba francosko-slovenskega vzporednega korpusa pri slovarski analizi nekaterih mejnih področij idiomatike. In M. Šorli (ed.) *Dvojezična korpusna leksikografija : slovenščina v kontrastu: novi izzivi, novi obeti*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp 12-34.
- Putanec, V. (1995). *Francusko-hrvatski rječnik*. Zagreb: Školska knjiga.
- Roberts, R., Montgomery, C. (1996). The Use of Corpora in Bilingual Lexicography. In M. Gellerstam et al. (eds.) *Proceedings of the 7th EURALEX International Congress*, Göteborg: Göteborg University, pp. 457-464.
- Roberts, R. (1996). Parallel-Text Analysis and Bilingual Lexicography. Accessed at: <http://www.dico.uottawa.ca/articles-fr.htm>. [01/12/2017]
- Salkie, R. (2008). How can lexicographers use a translation corpus?. In X. Richard et al. (eds.) *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*. Hangzhou: Zhejiang University. Accessed at: <http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Salkie.pdf>. [01/12/2017]
- Stosic, D., Marjanović, S., Miletic, A. (forth.). Corpus parallèle *ParCoLab* et lexicographie bilingue français-serbe : recherches et applications. In M. Srebro, J. Novaković (eds.) *Serbica*.
- Škrabal, M., Vavrin M. (2017). The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In. I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Brno: Lexical Computing CZ, pp. 124-137.
- Stanojević-Knežević, M. (2005). Rečnici francuskog jezika u Srbiji s Bibliografijom od 1904. do 2004. In M. Pavlović, J. Novaković (eds.) *Srpsko-francuski odnosi 1904-2004*. Beograd: Društvo za kulturnu saradnju Srbija-Francuska, Arhiv Srbije, pp. 219-233.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and the Non-Knowledge*. Tübingen: Max Niemeyer Verlag.
- Teubert W. (2002). The role of parallel corpora in translation and multilingual lexicography. In B. Altenberg, S. Granger (eds.) *Lexis in contrast: corpus-based approaches*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 189-214.
- Točanac, D., Dinić, T. & Vidić, J. (2017). *Francusko-srpski rečnik*. Beograd: Zavod za udžbenike.
- Zavaglia, A., Galafacci, G. (2014). Corpus, Parallélisme et Lexicographie Bilingue. In A. Abel et al. (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 587-597.



# Historical Lexicography, Etymology



# Lexicography in the Eighteenth-century Gran Chaco: the Old Zamuco Dictionary by Ignace Chomé

**Luca Ciucci**

*Language and Culture Research Centre, James Cook University*

*E-mail: luca.ciucci@jcu.edu.au*

## Abstract

The *Vocabulario de la lengua zamuca* is the only extant dictionary of Old Zamuco, an extinct Zamucoan language spoken in the 18th century in the abandoned mission of *San Ignacio de Samucos*, located in the northern Chaco lowlands of South America. This document was until now inaccessible to scholars, but has now been thoroughly studied by the present author, and found to contain very rich data, which establish its fundamental importance for linguistic studies on Zamucoan and Chaco languages. The critical edition of the dictionary is currently under publication. The original author of the dictionary, the Jesuit Father Ignace Chomé (1696–1768), reveals his brilliant linguistic intuition and remarkable skills in the collection and representation of linguistic data. One of the theoretical challenges he had to face was the threefold system of nominal suffixation, which is an absolute linguistic rarity. The present paper will show how the author dealt with the main word classes of Old Zamuco, that is verbs, nouns and adjectives. Chomé structured the entries for these lexemes in such a way as to provide plenty of information on inflectional and derivational morphology. The data from the dictionary permit new and interesting insights on the grammar of Old Zamuco.

**Keywords:** historical lexicography, lexicography of extinct languages, morphology, South American languages, Zamucoan

## 1 The Old Zamuco Language<sup>1</sup>

Old Zamuco is the earliest documented language of the Zamucoan family. It was spoken in the Jesuit mission of *San Ignacio de Samucos*, which was part of the *Jesuit Missions of Chiquitos* in southeastern Bolivia. The mission was founded in a remote and today unknown location of the northern Chaco in 1724. While the lingua franca in the other missions of the region was Chiquitano, the only language spoken in San Ignacio was Old Zamuco. This was part of a strategy aiming at the evangelization of the still uncontacted Zamucoan tribes of the region (Combès 2009: 82). When the mission was abandoned in 1745, its inhabitants were transferred to the neighboring missions, where they were gradually assimilated by the Chiquitano speaking population, and lost their language. Nowadays the Zamucoan family consists of two languages, Ayoreo and Chamacoco. Ayoreo is spoken by about 4,500 people in the Chaco lowlands of southeastern Bolivia and northern Paraguay, while the speakers of Chamacoco, about 2,000, traditionally inhabit the department of Alto Paraguay in Paraguay. Old Zamuco is lexically very close to Ayoreo (Kelm 1964), but comparative studies (Ciucci & Bertinetto 2015, 2017) show that the language shares a number of archaic features with Chamacoco, which is lexically very different from Ayoreo. Old Zamuco is therefore of fundamental importance for the reconstruction of Proto-Zamucoan.

1 **Abbreviations and conventions.** The data for Old Zamuco, Ayoreo and Chamacoco are reported in the respective orthography. The Spanish translations in the dictionary have been faithfully transcribed according to Chomé's orthography. In the text I have used the following abbreviations: 1, 2, 3 = first, second, third person; BF = base form; F = feminine; FF = full form; FS = feminine singular; FP = feminine plural; GF = generic form; IF = indeterminate form; IRLS = irrealis; M = masculine; MP = masculine plural; MS = masculine singular; NEG = negation; P = plural; REFL = third person reflexive; RLS = realis; S = singular. Chomé also makes use of abbreviations in his dictionary. Here they are reported with the respective translations: 3.<sup>a</sup> = third person; Abs. = generic form (Spanish *absoluto*); Fem. = feminine; N. = negation; pl. = plural; pos. = 'possession'; v. = see (Latin *vide*).



Old Zamuco was first studied by the founders of San Ignacio, Agustín de Castañares (1687-1744) and Domenico Bandiera (1693-1765). However, most of the available documentation is due to the work of Ignace Chomé (1696-1768), an admired polyglot with remarkable skills of linguistic analysis (Hervás y Panduro 1784: 47). He was the leading figure for linguistic studies in the *Jesuit Missions of Chiquitos*, being the author of a number of works on Chiquitano and Old Zamuco (Astorgano Abajo 2007: 744-746). Chomé was sent to San Ignacio de Samucos at the end of 1737, and in a letter dated 17th of May 1738 he communicated that he had learned Old Zamuco in five months (Possoz 1864: 94-97). Chomé wrote a grammar of Old Zamuco, *Arte de la lengua zamuca*, and a dictionary, *Vocabulario de la lengua zamuca*, which was a bidirectional dictionary between Old Zamuco and Spanish. While the grammar was published posthumously in 1958 (Chomé 1958 [before 1745]), the dictionary remained until recently inaccessible to scholars, which led the present author to prepare a critical edition of it (Ciucci, forthcoming). Chomé's grammar and dictionary are the main sources on Old Zamuco, with the dictionary containing approximately eight times the amount of data of the grammar. Other minor sources are wordlists and short texts collected by Hervás y Panduro (1784, 1786, 1787a,b) and by d'Orbigny (Lussagnet 1961, 1962). A few more data can be found in Clark (1937: 127-128), who published some materials which Hervás y Panduro used for his linguistic works.

## 2 The Old Zamuco Dictionary

The *Vocabulario de la lengua zamuca* is the only extant dictionary of Old Zamuco. When the Jesuits were expelled in 1767, the dictionary was in the inventory of objects of the mission of *Santiago de Chiquitos* (Brabo 1872: 522). It is not known when and why this and other volumes were transferred to La Paz, but there is evidence that in the first half of the 20th century the dictionary was kept at the Geographical Society of La Paz, where Vargas Ugarte saw the manuscript in 1931. Later, the dictionary was considered lost (Lussagnet 1961), and was then located at the Central Library of the *Universidad Mayor de San Andrés* (UMSA) in La Paz (Combès 2009), where it is currently kept. It is part of a collection of six volumes of grammatical and lexicographical studies on Chiquitano and Old Zamuco, attributed to Ignace Chomé. In 2014 Pier Marco Bertinetto and the present author obtained official permission from UMSA to prepare a critical edition of the whole collection. The upcoming edition of the Old Zamuco dictionary (Ciucci, forthcoming) is the first outcome of this project, which began at *Scuola Normale Superiore di Pisa* and is now continuing at James Cook University. In 2016 the collaboration of the UMSA Central Library with the two universities involved in the project led to the inclusion of this collection in the register of the UNESCO Memory of the World program by the UNESCO Regional Committee for Latin America and the Caribbean.

The volume containing the work is a leather-bound manuscript in folio form with 243 numbered pages. The text of the dictionary begins on page 2 and covers 242 pages. According to Vargas Ugarte (1931: 151) the manuscript consisted of two parts: (i) a Spanish-Old Zamuco section, which was 62 folios long (124 pages), and which is now lost; (ii) the Old Zamuco-Spanish section (the remaining 242 pages). The second part corresponds to what is left of the dictionary, and still has all the folios/pages counted by Vargas Ugarte, with the exception of the initial pages, immediately following the lost first section. As a result, the manuscript contains no title, which can only be obtained from Vargas Ugarte's (1931) description. Interestingly, the lost first section is dated 1739, the second 1738. Considering the original length of the two sections, the first had shorter entries, possibly with the main intention to cross-reference with the corresponding Old Zamuco-Spanish entry of the word, where Chomé offers detailed semantic and grammatical information.

The Old Zamuco-Spanish section of the dictionary was originally longer, because there were already some gaps in 1931, as one can see in Vargas Ugarte. The entry for the word *natu* is followed by the

one for *toriga*, which is actually incomplete, because the citation form was in the previous (lost) page, although it can easily be reconstructed thanks to the data contained in other parts of the dictionary. The first complete entry after *natu* is *torona*.<sup>2</sup> It is impossible to quantify how many pages are missing, but this gap can clearly be seen in the manuscript. Between pages 235 and 236 there are two or more pages missing, because the entry following *udda* is *utac*. This was not noted by Vargas Ugarte, but considering the number of folios he mentions these were already missing in 1931. Since there is no gap in the numbering of the pages, the obvious conclusion is that page numbers were added recently.

The manuscript has two columns. Chomé wrote the entries in the left column and used the right one to add further information or new entries later. For this reason, it is easy to identify later additions in the dictionary. Chomé completed the first version of the Old Zamuco-Spanish part in 1738. This is a remarkable achievement, considering that this first version includes most entries, and that Chomé had been able to do such a huge task within a year, and in the very year in which he had also learned the language. Moreover, the way in which the dictionary is structured shows that Chomé had acquired a deep understanding of Old Zamuco. It is possible that he also used some materials produced by his predecessors in the mission: for instance, in his grammar he reports data on verb moods and tenses from a small notebook by Augustín de Castañares (Chomé 1958: 147), but this does not lessen the value of Chomé's achievements.<sup>3</sup> We do not know when Chomé wrote his grammar, but it was very likely written after the dictionary, which is referred to many times in the former. In the dictionary one can also see some parts of text that appear to be copied from the grammar, but they are later additions in the right column. After finishing the first version of the dictionary, Chomé not only continued to add entries, but also checked the data several times, as proved by the many corrections in the manuscript. The dictionary is not signed by Chomé, but comparison with autograph letters by Chomé kept in Rome<sup>4</sup> shows that the calligraphy of the manuscript is that of Chomé, authenticating this as the original version of the work.<sup>5</sup>

The orthography of Old Zamuco is based on Spanish. This is the alphabetical order adopted in the dictionary: a, b, c {=/k/}, ç {=/s/}, ch, d, e, g {=/g/, but <gü>, and often <gu>, = /w/}, h, i, m, n, o, p, q, r, s, t, u {occasionally: u, ü}, v, y {=/j/}, z {=/s/}. However, Chomé does not strictly follow this order. The Old Zamuco-Spanish part is divided into many sections according to the first two or three letters of the entries: AA, AB, AC, ACH, AD, AE, AG, AH, AIB, AIC, and so on, until ZU at the end of the dictionary. The dictionary contains a very rich set of data. The remaining part of the dictionary comprises 2,110 entries (including some incomplete entries). The text of the critical edition (Ciucci, forthcoming) has about 96,763 words. For comparison, the edition of the grammar (Chomé 1958) has about 22,220 words. Extracting the Old Zamuco words from the dictionary results in a corpus of about 43,667 words. By contrast, if one adds up only the parts in Old Zamuco from all other available sources (excluding d'Orbigny, who collected his data more than 60 years after the Jesuit period), the resulting corpus has about 5,422 Old Zamuco words. This means that the dictionary offers about eight times the amount of data contained in the grammar and other minor sources from the Jesuit period.

The main word classes in Old Zamuco are verbs, nouns and adjectives, which form the majority of lexical entries. The Old Zamuco dictionary is rich in morphological information on these word classes, but presupposes that the reader has some knowledge of the inflectional morphology of Old

2 The description of the dictionary by Vargas Ugarte (1931: 151) contains two mistakes: (i) the total number of folios was 179 and not 279; (ii) the first complete entry after *natu* is not *sorona*, but *torona* (Vargas Ugarte confused <t> with <s>), which follows *toriga*.

3 This notebook by Augustín de Castañares is lost and is only known because it is mentioned in the Old Zamuco grammar.

4 *Archivum Romanum Societatis Iesu* (Roman Archives of the Society of Jesus), Fondo Gesuitico 751: 283; Fondo Gesuitico 752: 247, 251, 255, 259, 260, 261, 262.

5 By contrast, my analysis of the manuscript of the Old Zamuco grammar has revealed that it is a copy done by some missionaries. The original manuscript is lost.

Zamuco, as described in Chomé's grammar. In the following, I will deal with the inflection of these word classes, in order to show how Chomé organizes his dictionary.

### 3 Nominal Suffixation in Chomé's Dictionary

Old Zamuco is a fusional language. Nouns and adjectives have the same suffixation, so that they will be referred here as nominals. Every nominal suffix expresses gender (masculine vs feminine), number (singular vs plural) and a third grammatical system marking the "nominal form" (Bertinetto 2014a; Ciucci 2016). In all Zamucoan languages nominals distinguish a base form, a full form and an indeterminate form. The base form marks a noun phrase which carries out nominal predication, as in (1).<sup>6</sup> It is called the base form, because its singular coincides with the root or stem (depending on the nominal), and it is the base for any morphological operation.

- (1) *Nani-onnoe* [uom-io].  
indigenous\_man-MP.FF good-MP.BF  
'The indigenous people are good.' (Chomé 1958: 128)

By contrast, the full form and indeterminate form indicate that the noun phrase works as a core or peripheral argument (2).

- (2) [*Desi-oddoe*] *dac*.  
boy-MP.FF 3.come  
'The boys come.' (Chomé 1958: 128)

The difference between the full and indeterminate form involves semantics, because the indeterminate form has non-specific referent.

- (3) *A-gu* [*cucha-tic*].  
1S.RLS-eat thing-MS.IF  
'I eat something.' (unspecified) (Chomé 1958: 132)

Within the noun phrase, the possessor appears in the full form (4).<sup>7</sup> In Old Zamuco, adjectives follow the head of the noun phrase (5), and only the last element of the sequence (e.g. *uzodaddoe* in (5)) appears in the form and number required by the syntactic context. All preceding nominals (excluding the possessor) are in singular base form (e.g. *horâ* in (5)), independently of the referent's number. The head of the NP and the adjectival modifier only agree in gender.

- (4) *A-hu* [chuguper-itie] *unno-tie*.  
1S.RLS-pluck bird-MS.FF wing-MS.FF  
'I pluck a bird's wing.' (Chomé 1958: 128)
- (5) *Uom-onoe* *iyogueciñum* [*horâ* *uzoda-ddoe*]  
good-MP.FF 3.RLS.flee companion.M/FS.BF bad-mp.ff  
'The good (people) flee from bad companions.' (Ciucci, forthcoming)

This grammatical system is an absolute rarity, only found in the Zamucoan family. In his grammar Chomé described this unknown system combining the cases of Classical Latin and those of Old French (nominative vs the so-called *cas régime*) with the notion of nominal tense, which he knew

<sup>6</sup> The morphological expression of predicativity on nouns it is a rare feature (Bertinetto, Ciucci & Farina, forthcoming).

<sup>7</sup> According to Chomé, the possessor can also be in base form if the head of the NP is in full form (Chomé 1958: 128). However, this structure coincides with that of compounds, whose first element is in (singular) base form: cf. the compound *yote maitae* (3.FS.FF) 'branch of river' < *yote* (3.MS.BF) 'water' + *maitae* (3.FS.FF) 'finger'. The degree to which Chomé was aware of the ambiguity between possession and compounding is a problem that needs further investigation.

from Old Guaraní. As a result Chomé's description is obscure, because the labels he uses often do not correspond to the actual function of a given element. Table 1 features the nominal tripartition of Old Zamuco after reinterpretation of Chomé's data (Ciucci 2016: 682-690). Note that there are two types of full form. The first is the one normally used, while the second, indicated between parentheses as "full form II", consists of allomorphs in competition with the standard full form.<sup>8</sup> It was rarely used in Old Zamuco, and in Ayoreo it is only found in a few relics. In Zamucoan some affixes can undergo nasal harmony.<sup>9</sup> Below and in the following tables I do not provide allomorphs whose distribution is due to nasalization.

Table 1: The nominal tripartition of Old Zamuco.

	Singular		Plural	
	Masculine	Feminine	Masculine	Feminine
<b>Base Form</b>	-Ø	-Ø, -e	-o, -yo	-i, -yi
<b>Full Form</b>	-tie	-tae	-oddoe	-yie, -iyie
<b>(Full Form II)</b>	(-ré, -dde)	(-ac)	(-ao, -iao)	(-ai)
<b>Indeterminate Form</b>	-nic, -ric, -tic	-nac, -rac, -tac	-nigo, -rigo, -tigo	-rigui

Below are a feminine and a masculine noun from the dictionary (6-7). For ease of reading, I have supplemented words abbreviated by the author and I have omitted philological details which are not relevant here, thus reporting only the text of the upcoming critical edition. Each entry is followed by its English translation, with some glosses for the words I want to point out.

- (6) **Garomié**, *garomietae*, pl. *garomie*, *garomieyie*, cola de los animales.  
 [**Garomié** (FS.BF), *garomietae* (FS.FF), pl. *garomie* (FP.BF), *garomieyie* (FP.FF), animal tail.]
- (7) **Hubedda**, *hubeddatie*, pl. *hubeddao*, *hubeddaddoe*, grande moscardon.  
 [**Hubedda** (MS.BF), *hubeddatie* (MS.FF), pl. *hubeddao* (MP.BF), *hubeddaddoe* (MP.FF), big blowfly].

For all nominals Chomé systematically reports, in the same order: (i) the singular base form, which is the citation form; (ii) the singular full form; (iii) the plural base form; and (iv) the plural full form. The paradigm is then followed by the Spanish translation of the word. If the nominal has both masculine and feminine forms, the citation form is usually the masculine singular base form, as in the adjective *dugoc* (8). After the translation, the feminine singular base form (*dugogué*) is reported. Very often Chomé adds examples showing the use of the word: in this case the noun phrase *dahec dugoc* 'narrow road' and *pit dugoc* 'narrow stick'; note that when the noun phrase is provided out of context, even the second element of a N + A sequence is always in singular base form. Finally, derivations are also indicated, such as *ducotiga* 'narrowness', an abstract noun formed with the derivational suffix *-tiga*.

- (8) **Dugoc**, *dugotie*, pl. *dugoch*, *dugogodd*, estrecho, angosto, puntiagudo, como olla que no tiene asiento o botija, Fem. *dugogué*; *dahec dugoc*, camino angosto, estrecho; *pit dugoc*, palo estrecho; *dugotiga*, angostura.  
 [**Dugoc** (MS.BF), *dugotie* (MS.FF), pl. *dugoch* (MP.BF), *dugogodd* (MP.FF), tight, narrow, sharp-pointed, like pot which has no bottom or earthenware jug, Fem. *dugogué* (FS.BF); *dahec* (MS.BF) *dugoc* (MS.BF), tight, narrow road; *pit* (MS.BF) *dugoc* (MS.BF), narrow stick; *dugotiga* (MS.BF), narrowness.]

Some nominals have an epicene singular base form, such as *icha* (9). In this case, both masculine and feminine forms are listed together in the same canonical order established by Chomé.

<sup>8</sup> In the available data, one can see that the standard full form and the full form II have the same uses.

<sup>9</sup> For instance, the masculine plural full form suffix *-oddoe* (-/odoe/) can nasalize into *-onoe* (ex. (1)) or *-onoe* (ex. (5)).



- (9) *Icha*, *ichatie*, *ichatae*, pl. *ichao*, *ichai*, *ichaddoe*, *ichayie*, nuevo, nueva cosa; *borau icha*, ropa nueva; *guigueda icha*, casa nueva.

[*Icha* (M/F.BF), *ichatie* (MS.FF), *ichatae* (FS.FF), pl. *ichao* (MP.BF), *ichai* (MP.BF), *ichaddoe* (MP.FF), *ichayie* (FP.FF), new, new thing; *borau* (FS.BF) *icha* (M/F.BF), new clothes; *guigueda* (MS.BF) *icha* (M/F.BF), new house].

Chomé's dictionary is an invaluable source of information on the morphology of the base and full form, as well as gender derivation. By contrast, the indeterminate form is never provided in the dictionary, so that, apart from the general rules provided in the Old Zamuco grammar (Chomé 1958: 128), relatively little is known about the morphology of the indeterminate form. However, the dictionary is rich in examples and in some of them the indeterminate form is used. The same considerations apply to the so-called "full form II" (see Table 1). The little attention paid to the indeterminate form and full form II in the dictionary is due to the fact that in the language these forms occurred much less frequently than the standard full form, as one can see in the corpus of clauses and sentences extracted from the dictionary and the grammar.

The full form is also the most commonly used form in Ayoreo and Chamacoco. Although noun phrases in base form are restricted to nominal predication, thus making the plural base form also infrequent, the singular base form is the form to which suffixes are added in order to obtain the rest of the paradigm. This is particularly evident in *dugoc* /dugok/ (MS.BF) (8): the second velar consonant in *dogogoddoe* /dugogodoe/ (MP.FF) or *dugogué* /dugoge/ (FS.BF) can only be explained by the fact that the respective suffixes (-*oddoe* and -*e*) were added to the singular base form *dugoc* /dugok/, with subsequent voicing of /k/ in intervocalic position. In the base form plural *dugochō* /dugotçō/, /tç/ is also due to palatalization of /k/, which merges with the suffix -*yo* (-/jo/). The singular base form was also used to build the rest of the paradigm in the other Zamucoan languages, but this is less evident. While in Old Zamuco the singular base form is always shorter than the singular full form, in Ayoreo the full form often underwent phonetic erosion, so that it can coincide with the base form (10) or be even shorter. In Chamacoco the base form is no longer productive, and tends to be replaced by the full form, which in the masculine also loses the final segment(s) (11).

- (10) Old Zamuco: *chaboto* (FS.BF), *chabototae* (FS.FF) 'bat'  
 Ayoreo: *chaboto* (FS.BF), *chaboto* (FS.FF) 'bat'  
 Chamacoco: *sabito*<sup>?</sup> / *sabite*<sup>?</sup> (FS.BF), *sabitita* (FS.FF) 'bat'
- (11) Old Zamuco: *cucha* (MS.BF), *cuchatie* (MS.FF) 'thing'  
 Ayoreo: *cucha* (MS.BF), *cuchai* (MS.FF) 'thing'  
 Chamacoco: no base form, *kuchit* (MS.FF) 'thing'

This has consequences for lexicography, because in the main dictionary of Ayoreo (Higham *et al.* 2000) the singular full form is the citation form. Ulrich and Ulrich (2000) do not follow consistent criteria for nominals, although they prefer the full form as the citation form. Finally, the present author has used the singular full form as the citation form for Chamacoco (Ciucci 2013).

#### 4 Nouns Inflected for Possessor

In Zamucoan all nouns can be possessed, with the exception of proper nouns. There is however a difference between nouns which express the possessor through a prefix ("inflected for possessor") and those which do not mark the possessor ("uninflected for possessor"). This opposition between nouns inflected and uninflected for possessor is found in most Chaco languages (Fabre 2007). Table 2 features the possessive inflection of Old Zamuco. Table 2 is not exhaustive, because one can find interesting exceptions (concerning not only possessive morphology) in the dictionary, which will be discussed in future papers.



Table 2: Old Zamuco possessive inflection (adapted from Ciucci &amp; Bertinetto 2017).

1s	2s	3	REFL	1p	2p	GF
y-V-root ch-V-root z-V-root	Ø-a/V-root	PREFIXAL NOUNS: d-V-root g-V-root	d-a/V-root	ay-V-root az-V-root	ay-V-root az-V-root	p-V-root d-V-root g-V-root Ø-Ø-root
		THEMATIC NOUNS: Ø-V-root				
		RADICAL NOUNS: Ø-Ø-root				

There is no distinction in the third person between singular and plural, but between a reflexive third person (REFL), which is coreferent with the subject, and a “plain” third person, non-coreferent with the subject. The latter person, henceforth referred to as third person, is the least predictable of the paradigm and its morphological shape is related to the inflectional classes, as reported in the column of the third person. Many possessive nouns also have a form which expresses an unspecified or no possessor, called “generic form” (GF) (Bertinetto 2014a; Ciucci 2016). Such an unspecified possessor marker for nouns with possessive inflection is found in most Chaco languages (Campbell & Grondona 2012: 646). In Zamucoan, the alienability vs inalienability of a noun correlates with the presence of a generic form or lack thereof.

When a noun inflected for a possessor has no generic form, its citation form is the third person of the singular base form, as in *canariga* (12). The following forms are also in the third person. After the translation, the abbreviation *pos.* for ‘possessive’ (Spanish: *posesivo*) introduces the possessive inflection. Chomé systematically reports: (i) the first person singular; (ii) the second person singular; (iii) the third person (sometimes omitted if it coincides with the second singular); (iv) the first person plural and the second person plural, which are homophonous. All of these persons are always in singular base form. The only person not provided is the reflexive third person. This is because it is very regular and can be obtained by adding *d-* to the second person singular (Chomé 1958: 142). One can however observe some forms of reflexive third person in the examples of the dictionary: e.g. *ch-etig-are* (3.RLS-stretch\_out) *d-amaneca-tie* (REFL-arm-MS.FF) ‘s/he stretches out her/his arm’.

- (12) **Canariga**, *canarigatie*, pl. *canarigao*, *canarigannoe*, amistad mala, luxuria, pos. *yiganariga*, *aganariga*, *canariga*, pl. *ayiganariga*; *canarimiecoda*, muy luxurioso; v. *coda*.  
[**Canariga** (3.MS.BF), *canarigatie* (3.MS.FF), pl. *canarigao* (3.MP.BF), *canarigannoe* (3.MP.FF), bad friendship, lechery, pos. *yiganariga* (1S.MS.BF), *aganariga* (2S.MS.BF), *canariga* (3.MS.BF), pl. *ayiganariga* (1P/2P.MS.BF); *canarimiecoda*, very lecherous; v. *coda*.]

If a noun has the generic form, this can be indicated after the possessive inflection, preceded by the abbreviation Abs. (Spanish *absoluto* ‘absolute’), as in *didai* (13). The generic form is in singular base form, as usual, and the citation form is the third person.

- (13) **Didai**, *didaitie*, pl. *didaio*, *didaioddoe*, pata de los animales, pié del hombre, çapato, pos. *yiriddai*, *ariddai*, *diddai*, pl. *ayiriddai*, Abs. *piriddai*.  
[**Didai** (3.MS.BF), *didaitie*, pl. *didaio*, *didaioddoe*, animal leg, human foot, shoe, pos. *yiriddai* (1S.MS.BF), *ariddai* (2S.MS.BF), *diddai* (3.MS.BF), pl. *ayiriddai* (1P/2P.MS.BF), Abs. *piriddai* (GF.MS.BF).]<sup>10</sup>

However, Chomé mostly chooses the generic form as the citation form, as in *carup* (14), while the third person is provided only within the possessive inflection. This is consistent with the fact that in

<sup>10</sup> *Didai* and *diddai* are the same form. Chomé is not consistent with the use of double consonants in the orthography: they were possibly used in order to indicate the length of the preceding vowel (Kelm 1964: 462).

the grammar he analyzes the possessive inflection as if it were built from the generic form (Chomé 1958: 140-142). Actually, it is more appropriate to consider the third person, rather than the generic form, as the base of the possessive paradigm (Ciucci 2016). This could be the reason why Chomé sometimes uses the third person as the citation form, despite the presence of a generic one, as in *didai* (13). This fluctuation between third person and generic form has also to do with the fact that their formation is lexically idiosyncratic, so that this raised the problem of which one to choose as the entry word. From a modern perspective, the generic form should not be put in relation with the whole paradigm, as Chomé did, but just with the third person, which is the base of the possessive paradigm (Ciucci 2016). Indeed, it is possible to identify both nouns where the third person derives from the generic form, and nouns where the opposite occurs (Ciucci & Bertinetto 2017: 323).

- (14) **Carup**, *carubitie*, pl. *carubio*, *caruboddoe*, *cuerda*, *soga*, pos. *yigarup*, *agarup*, *igarup*, pl. *ayigarup*. [**Carup** (GF.MS.BF), *carubitie*, pl. *carubio*, *caruboddoe*, *rope*, *cord*, pos. *yigarup* (1S.MS.BF), *agarup* (2S.MS.BF), *igarup* (3.MS.BF), pl. *ayigarup* (1P/2P.MS.BF).]

## 5 Verbs and Verbal Paradigms in the Old Zamuco Dictionary

All Zamucoan languages are tenseless (Bertinetto 2014b).<sup>11</sup> Old Zamuco verbs display a distinction between realis and irrealis mood, which has partly disappeared in Ayoreo and Chamacoco (Ciucci & Bertinetto 2015). Table 3 briefly illustrates the verb inflection of Old Zamuco. The third person is the most idiosyncratic, and its shape determines the inflectional class of the verb. It is the only person which lacks singular vs plural distinction.

Table 3: The verb inflection of Old Zamuco (adapted from Ciucci & Bertinetto 2015).

	REALIS	IRREALIS
1s	<i>a-V-root</i>	<i>y-/ch-/z-V-root</i>
2s	<i>d-a/V-root</i>	<i>Ø-a/V-root</i>
3	<i>ch-/t-/z-Ø-(V)-root</i>	<i>d-/n-/Ø-(V)-root</i>
1p	<i>a-V-root-suffix</i>	<i>y-/ch-/z-V-root-suffix</i>
2p	<i>d-a/V-root-suffix</i>	<i>Ø-a/V-root-suffix</i>

Below is a typical dictionary entry for a verb. In the following analysis, the English translation will be divided into four parts (16-19).

- (15) **Airaha**, *daraha*, *chiraha*, pl. *airahago*, *darahao*, *saber*, *aprehender*, *aprender*, N. *ca chiraha*, 3.<sup>a</sup> *ca diraha*; *pirahac*, *sabido*, etc., pos. *chirahac*, *arahac*, *irahac*, pl. *ayirahac*; *pirahazore*, *el que sabe*, *conoce*; *pirahariga*, *conocimiento*, etc.; *Tupâde iraharigatie ome cuchaddoe ca iruericuz*, *es infinita la sabiduria de Dios*; *airaha ezabedayie*, *sé leer*; *airaha poriyie*, *saber trepar en arbol*; *airaha teutie*, *bolver en si el que desvariaba*; *e airaha*, *ainarañumé*, *resabiado ser*; *airaha quitic*, *comprender*, *alcanzar con el entendimiento*; *ca diraha dirire peatic*, *es basto*, *rudo*, *tupido de entendimiento*; *ca araha dirire peatic*, *sois una bestia*, *lo tienen por suma injuria*; *irahezore*, *vel*, *iraheque*, *el que no sabe*; *chirahezore*, *yo soi el que no sabe*, *arahezore*, *tu eres el que no sabe*, etc., Fem. *iraheto*; *deachatie iraheque*, *vel*, *irahezore*, *cuchuzoda dateputigatie*, etc., *si alguno no supiere la gravedad del pecado*, etc.

The citation form is always the first person singular realis, followed by the other persons of the irrealis, and by the Spanish translation in the infinitive (16). In addition, Chomé shows the negation of

<sup>11</sup> More precisely, Zamucoan languages belong to a group of languages which Bertinetto (2014b) calls “radical tenseless” languages. Radical tenselessness is a rare feature.

first singular and third person. Since the negation falls within the scope of the irrealis, what Chomé actually does is to provide the irrealis of the first singular and third person, preceded by the negative particle *ca*. This scheme recurs systematically for all verbs reported in the dictionary (excluding of course defective and uninflectable verbs).

- (16) [*Airaha* (1S.RLS), *daraha* (2S.RLS), *chiraha* (3.RLS), pl. *airahago* (1P.RLS), *darahao* (2P.RLS), to know, to learn, to understand, N. *ca* (NEG) *chiraha* (1S.IRLS), 3.<sup>a</sup> *ca* (NEG) *diraha* (3.IRLS)]

Following the system used in Latin dictionaries, Chomé establishes the first person realis (which he considers the present indicative) as the citation form.<sup>12</sup> Since the first realis has the prefix *a-*, the vast majority of entries for verbs (i.e., all but irregular and defective ones) are found in the letter *A*, which covers 152 of the 242 pages of the dictionary, for a total of 900 entries out of 2,110. In the grammar and in the examples of the dictionary, one can see some irregular verbs with first person realis beginning with *o-*, possibly owing to deletion of prefix *a-*, such as ‘to give’: *ozi* (1S.RLS), *izi* (3.RLS). Unfortunately, there is limited information on these verbs, which are particularly interesting for diachronic reasons (Ciucci & Bertinetto 2015), because the letter *O* was in the lost part of the manuscript. In Chamacoco the citation form of the verb is the third person realis (Ulrich & Ulrich 2000; Ciucci 2013), while in the Ayoreo dictionary by Higham *et al.* (2000) the verb’s theme is the citation form, except that very often also the third person is provided when it is irregular.<sup>13</sup>

Entries for verbs provide a large part of the verb paradigm. The only missing persons are the second singular irrealis, and the first and second plural irrealis. However, these were easy to reconstruct for the reader who had some knowledge of Old Zamuco grammar: indeed the second singular irrealis only differs from its realis counterpart in that it lacks the prefix *d-*. The first and second persons plural irrealis are derived from the first and second singular irrealis by adding the same pluralizing suffixes as the respective realis persons.

In his grammar, Chomé gives a complicated picture of verb inflection, identifying a number of moods and tenses (Chomé 1958: 143-148) which actually do not exist, because they are just the combination of verb forms with independent adverbial particles (e.g., expressing past or future reference). This is not surprising, because Chomé’s view of grammar is mostly based on the Latin model. In the dictionary, however, Chomé had the brilliant intuition that the verb system could be reduced to the opposition between what he considered the present indicative (the realis mood) and the forms used when the verb is negated (the irrealis mood).

From the verb root one can derive a noun with passive meaning: *pirahac* (17), which one can translate as ‘known’, ‘who/what is known’, ‘known person/thing’. It is usually provided in the generic form, followed by the possessive inflection, according to the same scheme seen in §4. Chomé also provides the nomen agentis (*pirahazore*) derived from the verb root by means of the suffix *-zore* or *-gore*, to which he often adds the abstract noun (*pirahariga*). There are reasons to surmise that the abstract noun is not always directly derived from the verb root, but it can also be derived from the passive deverbal noun (Ciucci 2016: 493-508). The deverbal noun with passive meaning and the nomen agentis are provided immediately after the verb, because they are considered participles by Chomé (Chomé 1958: 145). Note that in this example only the generic form of the nomen agentis and the abstract noun are reported, possibly because their possessive inflection is identical with that of *pirahac*.

- (17) [*pirahac* (GF.MS.BF), known, etc., pos. *chirahac* (1S.MS.BF), *arahac* (2S.MS.BF), *irahac* (3.MS.BF), pl. *ayirahac* (1P/2P.MS.BF); *pirahazore* (GF.MS.BF), the one who knows; *pirahariga* (GF.MS.BF), knowledge, etc.]

<sup>12</sup> The same criterion is frequently used by early lexicographers of Native American languages (Smith-Stark 2007: 60-61).

<sup>13</sup> Ayoreo has no distinction between third person realis and irrealis.

After deverbal nouns one can find examples of expressions or clauses, sometimes even sentences, in order to see the uses of the verb (18). The dictionary is also very rich in examples, particularly in entries for verbs, but also in entries for nouns and adjectives.<sup>14</sup> These are very interesting data for linguistic analysis. Note that Chomé very often gives examples which are translated with the Spanish infinitive: in this case the respective Old Zamuco verb is in the first person singular realis, that is in the citation form (below, I have only glossed the relevant verbs).

- (18) [*Tupâde iraharigatie ome cuchaddoe ca iruericuz*, the wisdom of God is infinite; *airaha* (1s.RLS) *ezabedayie*, I know how to read; *airaha* (1s.RLS) *poriyie*, to know how to climb a tree; *airaha* (1s.RLS) *teutie*, to come round, the one who was delirious; *e airaha* (1s.RLS), *ainarañumé*, to be knowing; *airaha* (1s.RLS) *quitic*, to understand, to get to understand; *ca diraha* (3.IRLS) *dirire peatic*, they have bad understanding; *ca araha* (2s.IRLS) *dirire peatic*, you are a brute, they consider it the supreme insult;]

The deverbal nouns from the verb ‘to know’ have two antonyms (19). From the available data, one can see that this is not very frequent in Old Zamuco,<sup>15</sup> and this is why Chomé reports them, along with some examples, in the final part of the entry. Note that here the translations by Chomé are not literal. Since *iraheque* (literally ‘what is unknown’) and *irahezore* can be used interchangeably, Chomé assigns to both the meaning of *irahezore*. Here both nouns are in third person, which suggests that there was no generic form. *Iraheto* is the regular feminine form of *irahezore*. Finally, the last example is part of a sentence: *cuchuzoda dateputigatie* ‘the gravity of the sin’ is the subject and *deachatie iraheque* or *deachatie irahezore* carries out nominal predication (owing to the final element in base form). The literal meaning is ‘the gravity of the sin is the unknown thing (*iraheque*) of someone’ or (implying some sort of personification) ‘the gravity of the sin is the unknowing one (*irahezore*) of someone’.

- (19) [*irahezore* (3.MS.BF), or, *iraheque* (3.MS.BF), the one who does not know; *chirahezore* (1s.MS.BF), I am the one who does not know (lit. my unknowing one), *arahezore* (2s.MS.BF), you are the one who does not know (lit. your unknowing one), etc., Fem. *iraheto* (3.FS.BF); *deachatie iraheque*, or, *irahezore*, *cuchuzoda dateputigatie*, etc., if someone did not know the gravity of the sin, etc.]

In (20) there is a defective verb from the dictionary: one can see that its paradigm is limited to third person realis and irrealis, with the third person realis as the citation form.

- (20) **Zora**, assomar el sol al horizonte quando nace, assomar los sembrados; *e chacaddoe zora*, ya nace lo que sembré; *guiede zaî zora*, recién apunta el sol; N. *ca norâ*.  
[**Zora** (3.RLS), to come up the sun over the horizon at sunrise, to come up what is sown; *e chacaddoe zora*, what I sowed already comes up; *guiede zaî zora*, recién apunta el sol; N. *ca* (NEG) *norâ* (3.IRLS).]

The verb *zora* has an interesting feature: it shows the prefix *z-* (/s/-) for the third person realis. In Chomé’s grammar one can find some hints of a group of verbs with this prefix (Ciucci 2016: 227-228), but this was an open question, because the only example provided in the grammar was difficult to interpret (Ciucci & Bertinetto 2015: 17). However, the new data from the dictionary permit the identification of a small group of verbs characterized by the third person realis prefix *z-*, by the third person irrealis prefix *n-*, and by the first irrealis prefix *z-*. Note that *n-* here is not due to nasal harmony, but is precisely selected by this group of verbs. In (21) I report the dictionary entry for one of them. In the translation I have only glossed the verb paradigms, and have segmented the prefixes which distinguish this group of verbs from the others.

<sup>14</sup> The fact that there are almost no examples in (6-9) and (12-14) is due to the fact that I have purposely chosen short entries in order to focus on the main features of the dictionary.

<sup>15</sup> This is also indirectly confirmed by comparison with Ayoreo. In Chamacoco, deverbal nouns have mostly disappeared.



- (21) *Aoz*, *daoz*, *zoz*, pl. *ahoco*, *dao*zo, echar fuera animales, etc., N. *ca zoz*, 3.<sup>a</sup> *ca noz*; *pozic*, echado fuera, pos. *zocic*, *aozic*, pl. *azozic*; *pozigore*, el que echa fuera; *aoz ore*, forçalos, echalos fuera de su fortaleza; *aoz gacayie ahâ inaguatae*, acorralla las vacas; *aoz cavayuoddoe*, arrea los cavallos; *yote ozic*, broça que trae el rio; v. *zozic*; *yonuratie zoz oyidoddoe*, *pachaddoe*, la avenida trae pescado, broza, etc.

[*Aoz* (1S.RLS), *daoz* (2S.RLS), *z-oz* (3.RLS), pl. *ahoco* (1P.RLS), *dao*zo, to cast out animals, etc., N. *ca z-oz* (1P.RLS), 3.<sup>a</sup> *ca n-oz* (3.IRLS); *pozic*, cast out, pos. *zocic*, *aozic*, pl. *azozic*; *pozigore*, the one who casts out; *aoz ore*, force them, deprive them of their strength!; *aoz gacayie ahâ inaguatae*, pen the cows!; *aoz cavayuoddoe*, spur on the horses!; *yote ozic*, brushwood which the river brings; see *zozic*; *yonuratie zoz oyidoddoe*, *pachaddoe*, the flood brings fish, brushwood, etc.]

## 6 Conclusions

The Old Zamuco dictionary by Chomé is not only the main source on this language, but also a document of exceptional interest, owing to the quality and quantity of data. The way in which Chomé deals with the unusual nominal suffixation of Old Zamuco is original and reveals a deep understanding of the language. The criteria used by Chomé are also very effective with respect to the forms of verbal and possessive inflection which he decided to provide in each entry. The dictionary is very rich in grammatical information. As far as suffixation is concerned, one can identify at least 856 masculine paradigms and 487 feminine paradigms (including those with epicene singular base form). Considering also masculine nominals for which suffixation is not provided, the corresponding feminine is indicated 533 times. In the dictionary are also 930 nouns for which the possessive inflection is reported, along with the paradigm of about 850 verbs and verbal periphrases (excluding uninflected verbs).

As shown in this paper, Chomé's data could hardly be interpreted without previous descriptive and comparative research on Zamucoan, carried out by Pier Marco Bertinetto and the present author. The analysis of the dictionary is also an example of how diachronic studies, such as Ciucci & Bertinetto (2015, 2017), can feed into the interpretation of synchronic data, such as those collected by Chomé.

The dictionary permits to make interesting additional findings with respect to what was previously known (Ciucci 2016): the group of verbs with third person prefix *z-* is just one example. Another concerns nominal suffixation: in the grammar, Chomé never documented the suffix *-tac* for the feminine singular indeterminate form (see Table 1). However, its cognate is found in Ayoreo and Chamacoco, so that one could question whether it was also present in Old Zamuco (Ciucci 2016: 747-748). The dictionary solved the dilemma with a few occurrences of *-tac*. Possessive classifiers are a further example: they are present in all Zamucoan languages, but they could not be studied in Old Zamuco, owing to the paucity of information in the grammar. By contrast, the dictionary contains interesting data on classifiers, which can thus be compared with Ayoreo and Chamacoco (Ciucci & Bertinetto, forthcoming). The dictionary also allows us a better understanding of Old Zamuco syntax. With the previously available data it was not possible to know whether Old Zamuco had so-called para-hypotactical structures, as in Ayoreo or Chamacoco (Bertinetto & Ciucci 2012). By the many examples in the dictionary, one can see that Old Zamuco employed a large number of paratactical structures, but there is no example of para-hypotaxis. The importance of the dictionary is not limited to Zamucoan, indeed, Ciucci (2014) has identified a number of borrowings between Zamucoan and other Chaco families, particularly Guaycuruan and Mataguayan, so that the data in the dictionary are also relevant for studies on language contact. Kelm (1964: 815) interprets some examples from Chomé's grammar in ethnographical terms and identifies similarities with the Ayoreo society. Ultimately, the dictionary is also an interesting source of anthropological information about the Old Zamuco speaking people, and a first analysis confirms that Ayoreo and Old Zamuco have similar cultural background (see Ciucci, forthcoming 2019).



## References

- Astorgano Abajo, A. (ed.) (2007). *Lorenzo Hervás y Panduro, Biblioteca jesuítico-española (1759-1799)*. Madrid: LIBRIS.
- Bertinetto, P.M. (2014a) [2009]. Ayoreo. In M. Crevels, P. Muysken (eds.) *Lenguas de Bolivia*, Tomo 3: Oriente. La Paz: Plural Editores, pp. 369-413. [English version (2009). Ayoreo (Zamucó). A grammatical sketch. In *Quaderni del Laboratorio di Linguistica* 8 n.s.]
- Bertinetto, P.M. (2014b). Tenselessness in South American indigenous languages with focus on Ayoreo (Zamucó). In *LIAMES* 14, pp. 149-171.
- Bertinetto, P.M. & Ciucci, L. (2012). Parataxis, hypotaxis and para-hypotaxis in the Zamucóan languages. In *Linguistic Discovery* 10(1), pp. 89-111.
- Bertinetto, P.M., Ciucci, L. & Farina, M. (forthcoming). Morphologically expressed non-verbal predication.
- Brabo, F.J. (1872). *Inventarios de los bienes hallados a la expulsión de los jesuitas y ocupación de sus temporalidades por decreto de Carlos III*. Madrid: Imp. y Esterotipia de W. Rivadeneyra.
- Campbell, L. & Grondona, V. (2012). Languages of the Chaco and Southern Cone. In L. Campbell, V. Grondona (eds.) *The Indigenous languages of South America. A Comprehensive Guide*. Berlin: De Gruyter Mouton, pp. 625-668.
- Chomé, I. (1958) [before 1745]. Arte de la lengua zamuca (Présentation de Suzanne Lussagnet). In *Journal de la Société des Américanistes de Paris* 47, pp. 121-178.
- Ciucci, L. (2013). Chamacoco lexicographical supplement. In *Quaderni del Laboratorio di Linguistica* 12 n.s.
- Ciucci, L. (2014). Tracce di contatto tra la famiglia zamucó (ayoreo, chamacoco) e altre lingue del Chaco: prime prospezioni. In *Quaderni del Laboratorio di Linguistica* 13 n.s.
- Ciucci, L. (2016) [2013]. *Inflectional morphology in the Zamucóan languages*. Asunción: CEADUC.
- Ciucci, L. (ed.) (forthcoming). *Ignace Chomé. Vocabulario de la lengua zamuca - Edición crítica y comentario lingüístico*. Madrid/Frankfurt: Iberoamericana Verfuert Verlag.
- Ciucci, L. (forthcoming, 2019). A culture of secrecy: the hidden narratives of the Ayoreo. In A. Storch, A. Hollington, N. Nassenstein, A.Y. Aikhenvald (eds.) *Creativity in language: secret codes and special styles*. A special issue of the *International Journal of Language and Culture*.
- Ciucci, L. & Bertinetto, P.M. (2015). A diachronic view of the Zamucóan verb inflection. In *Folia Linguistica Historica*, 36(1), pp. 19-87.
- Ciucci, L. & Bertinetto, P.M. (2017). Possessive inflection in Proto-Zamucóan: a reconstruction. In *Diachronica* 34(3), pp. 283-330.
- Ciucci, L. & Bertinetto, P.M. (forthcoming). Possessive classifiers in Zamucóan.
- Clark, Ch.U. (1937). Jesuit letters to Hervás on American languages and customs. In *Journal de la Société des Américanistes* 29, pp. 97-145.
- Combès, I. (2009). *Zamucos*. Cochabamba: Instituto de Misionerología.
- Fabre, A. (2007). Morfosintaxis de los clasificadores posesivos en las lenguas del Gran Chaco (Argentina, Bolivia y Paraguay). In *UniverSOS* 4, pp. 67-85.
- Hervás y Panduro, L. (1784). *Catalogo delle lingue conosciute e notizia della loro affinità, e diversità*. Cesena: Gregorio Biasini.
- Hervás y Panduro, L. (1786). *Aritmetica delle nazioni e divisione del tempo fra l'Orientali*. Cesena: Gregorio Biasini.
- Hervás y Panduro, L. (1787a). *Vocabolario poliglotta con prolegomeni sopra più di CL lingue*. Cesena: Gregorio Biasini.
- Hervás y Panduro, L. (1787b). *Saggio pratico delle lingue*. Cesena: Gregorio Biasini.
- Higham, A., Morarie, M. & Paul, G. (2000). *Ayoré-English dictionary*. Sanford, FL.: New Tribes Mission. 3 vols.
- Kelm, H. (1964). Das Zamucó: eine lebende Sprache. In *Anthropos* 59, pp. 457-516, 770-842.
- Lussagnet, S. (1961). Vocabulaires Samuku, Morotoko, Poturero et Guarañoka précédés d'une étude historique et géographique sur les anciens Samuku du Chaco bolivien et leur voisins. In *Journal de la Société des Américanistes de Paris* 50, pp. 185-243.
- Lussagnet, S. (1962). Vocabulaires Samuku, Morotoko, Poturero et Guarañoka (suite et fin). In *Journal de la Société des Américanistes de Paris* 51, pp. 35-64.
- Possoz, A. (1864). *Vie du R. P. Ignace Chomé de la Compagnie de Jésus*. Douai: Dechristé.
- Smith-Stark, Th.C. (2007). Lexicography in New Spain (1492-1611). In O. Zwartjes, R. Arzápalo Marín, Th.C. Smith-Stark (eds.) *Missionary Linguistics IV / Lingüística misionera IV*. Lexicography. Amsterdam: John Benjamins, pp. 3-82.

- Ulrich, M. & Ulrich, R. (2000). *Diccionario Ishiro (Chamacoco) - Español / Español - Ishiro (Chamacoco)*. Asunción: Misión a Nuevas Tribus Paraguay.
- Vargas Ugarte, R. (1931). Contribución a la bibliografía de las lenguas americanas. In *Boletín del Instituto de Investigaciones Históricas* 13, pp. 148-155.

## Acknowledgements

First and foremost, I would like to thank the *Universidad Mayor de San Andrés* (UMSA) in La Paz for its support to the project of critical edition of the volumes of grammatical and lexicographical materials attributed to Ignace Chomé, in particular to the rector, Prof. Waldo Albarracín Sánchez, to the vice-rector, Prof. Alberto Quevedo Iriarte, and to Marilyn Sánchez Rada, director of the UMSA Central Library and member of Memory of the World National Committee of Bolivia. I also acknowledge the indispensable help of Simona Di Noia of the Italian Embassy in La Paz. I would like to express my gratitude to the following scholars for their help and suggestions: Alexandra Y. Aikhenvald, Pier Marco Bertinetto, Alice Cavinato, Isabelle Combès, Mauro Costantino, R. M. W. Dixon, Brigitta Flick, Bernhard Hurch, Irene Lorenzini and Gabriella Erica Pia. I am entirely responsible for any mistake this work may contain.



# Historical Corpus and Historical Dictionary: Merging Two Ongoing Projects of Old French by Integrating their Editing Systems

**Sabine Tittel**

*Heidelberg Academy of Sciences and Humanities*

*E-mail: sabine.tittel@urz.uni-heidelberg.de*

## Abstract

To combine corpus data with dictionary data has two advantages: (i) It embeds the vocabulary of the corpus texts within the overall system of the language, and it semantically disambiguates the texts. (ii) The corpus data enrich the dictionary and shed new light on the comprehension of the vocabulary. The retrospective integration of corpus data into a dictionary is a task that has to focus on two aspects, (i) on the integration of the word forms, and (ii) on the semantic integration of the words. This second aspect continues to be an important issue, particularly for historical languages. Automated solutions do not exist. In this paper, we present the retrospective integration – both with a graphical and a semantic focus – of the corpus of Old French legal texts, *Documents linguistiques galloromans* (with approx. 800,000 attestations of Old French lexemes), into the *Dictionnaire étymologique de l'ancien français* (with 83,000 dictionary entries). We have implemented a semi-automated process resulting in a time-saving editorial workflow to accomplish the data integration. Further, we have created a twofold publication concept for the dictionary entries that makes for a straightforward way of enriching the dictionary with the valuable material of the domain of Old French law.

**Keywords:** historical lexicography, corpus linguistics, Old French, dictionary writing system, scholarly digital text edition, history of law

## 1 Introduction

The retrospective integration of two large and long-standing projects of Medieval French, i.e., the corpus of Old French legal documents, *Documents linguistiques galloromans* – DocLing, and the comprehensive dictionary of the Old French language, *Dictionnaire étymologique de l'ancien français* – DEAF, is a challenging task. Such a retrospective approach cannot profit from the advantages that a newly created corpus lexicographic venture has. The latter typically defines its linguistic corpus in the initial phase of the project as a well-integrated part of the system architecture. An example in the field of Old French is the *Dictionnaire Électronique de Chrétien de Troyes* – DÉCT (Kunstmann 2007–2014).

Instead, combining the two long-established projects DocLing and DEAF has to deal with distinct data formats that are specific for the corpus and the dictionary, respectively, and with the adaptation of a tailor-made electronic dictionary writing system.

In this paper, we present our data integration that focuses on two main aspects. Firstly, we have implemented the integration with respect to the word forms attested in the corpus texts. This means that we merge the approximately 800,000 word occurrences of DocLing within the parts of the approximately 83,000 DEAF entries that present the graphical realizations of the lexemes. Secondly, we have implemented the integration with respect to the meaning of the words. This means that we perform a semantic mapping of the word occurrences within the DEAF entries. This second aspect is not an obvious task and continues to be a challenging issue, particularly for historical languages.

From the point of view of historical linguistics, the data integration has two main benefits: (i) It enriches the dictionary with data from the discourse tradition of medieval law that had previously been widely unnoticed by lexicography. (ii) It embeds the vocabulary of the corpus texts within the overall system of the medieval French language. At the same time, it creates a means for the semantic disambiguation of the vocabulary and, thus, for the understanding of its meaning. We believe our approach is promising for the integration of other corpora and dictionaries.

This paper is structured as follows: In Section 2, we introduce the two lexical resources DEAF and DocLing. Section 3 explains our integration approach, and Section 4 presents the semi-automated workflow of the data integration: First, we discuss the mapping of the data models and the rules for the data import and export. Then, we show the steps of the semi-automated workflow and the graphical user interfaces we developed for the integration. Section 5 presents the online publication of the integrated resources with two successive release steps. Section 6 discusses the added values for both the dictionary and for the corpus from the point of view of the historical content. In Section 7, we address the remaining issue of how to link back from the corpus to the dictionary, and Section 6 concludes our work.

## 2 The Lexical Resources

The DEAF (Baldinger, Möhren & Städtler 1971–) is a longstanding dictionary compiled under the aegis of the Heidelberg Academy of Sciences and Humanities in Heidelberg, Germany. It researches the Old French language from its first resource 842 AD until ca. 1350. The dictionary is traditionally published as a series of printed books. Since 2010, it is also published as a versatile electronic version with online dictionary entries and elaborate research functions, called DEAF *électronique* (DEAF*él*).<sup>1</sup> The DEAF organizes the Old French lexicon in word families to show the etymological relations between single lexemes. The *main-lemma* of a dictionary entry is the lexeme that is developed or borrowed from a Latin, Greek, etc., origin. The derivations from this lexeme are the *sub-lemmata*.

The online publication DEAF*él* consists of two parts: DEAF*pré* and DEAF*plus*. DEAF*plus* is the online version of the well-known dictionary DEAF; it consists of extensive articles of the scientifically acknowledged lexicographical quality that has characterized the DEAF for more than 30 years (DEAF*plus* features a number of added values compared to the printed book and explaining the ‘plus’, cf. Tittel 2010). DEAF*pré* is not a dictionary in its proper sense, but offers the complete raw material of the DEAF. This material is accessible online in the form of compendious articles that are orthographically and semantically structured in a semi-automated manner. Therefore, it is valuable for all research within our discipline as long as DEAF*plus* does not cover the entire alphabet. Together, DEAF*plus* and DEAF*pré* form DEAF*él* with approximately 83,000 entries.<sup>2</sup>

The dictionary draws its source material from an open textual corpus. This corpus consists of three components: (i) the entirety of accessible scholarly text editions of Old French texts (currently around 3,000 primary texts within around 10,000 manuscripts), (ii) the information published in the secondary literature (monographies, journals), and (iii) the information published in related dictionaries. The 1.5 million handwritten and now digitized *fiches* (slips) lead to 12 million attestations within the corpus. For the most part, the textual corpus of the dictionary cannot be digitally accessed. Even though

1 See <https://deaf-server.adw.uni-heidelberg.de> [accessed 03-28-2018].

2 The lemmata treated in the form of DEAF*plus* (letters D – K) will add up to approximately 9,000 in 2020; approximately 73,000 lemmata will remain as DEAF*pré* (the rest of the alphabet) for the time being. The division of the dictionary into two considerably different parts has been implemented in 2010; it is due to changed monetary conditions affecting the duration of the project.



the editorial process does to some extent integrate corpus queries on a few digital texts, we can say that the DEAF is clearly not a corpus lexicographic endeavor.

In 2014, the DEAF started a cooperation with DocLing (University of Zurich, Glessgen 1998–).<sup>3</sup> DocLing is one of the most significant projects of Old French corpus linguistics. It comprises digital scholarly text editions of 2,185 Medieval French charters (deeds of donation, contracts of purchase, inheritance matter, etc.) dating between 1205 AD and ca. 1450 and with approximately 800,000 word-occurrences.<sup>4</sup> These text editions were created within the framework of the corpus project.<sup>5</sup> They make accessible the important textual genre of legal documents, and are textual witnesses that cover all aspects of human social interaction. For the time being, the DocLing material will be integrated into the part of the DEAF that we call DEAF*pré*.

### 3 Integration Approach

The aim of the DocLing-DEAF cooperation is the full integration of the DocLing data into the DEAF dictionary. The attestations of Old French lexemes from the text editions of DocLing shall find their correct place within the dictionary entries of the DEAF.

According to Asmussen (2013), there exist two prototypical approaches to the retrospective integration of a corpus and a dictionary. The first is to add “deliberately selected and processed text material from a corpus [...] to a dictionary to give more citations for each definition in the dictionary” (Asmussen 2013: 1084). He qualifies this approach as being a tedious and error-prone task that should not be carried out (*ib.*: 1087). Instead, he promotes the second approach, i.e., to establish a virtual interlinking based on orthographical and morphological matches of the lemmatized corpus data with dictionary entries. We consider this second approach insufficient for the following reason: The interlinking focuses only on the graphical realizations of the lexemes. The question of how to establish pointers from the lexical units to the right sense within a dictionary entry is thus not resolved. It is clear that this approach still needs the semantic disambiguation of the corpus data and a correct semantic mapping. As such, the error-prone task remains. Asmussen identifies the question of how to semantically map the corpus data as an important issue for future research (*ibid.*). We will present our solution to this question in this paper.

With the integration, we follow the second approach while we also address the issue of how to perform the semantic mapping. Hence, we have two objectives. The first objective is to integrate the corpus data with respect to the graphical realizations of the lexemes, i.e., within the *apparatus of graphical variants* of the respective dictionary articles.<sup>6</sup> The second objective – and this is our main concern – is to integrate the corpus data with respect to the meaning of the words, i.e., within the

3 See <http://www.rose.uzh.ch/docling/> [accessed 03-28-2018].

4 DocLing also comprises documents of the Romance speaking Suisse regions and the Francoprovençal linguistic area. For the moment, these are not relevant for our purposes.

5 Most other corpora with texts of Old French incorporate already existing text editions, e.g. the *Textes de Français Ancien* – TFA (Kunstmann, Ottawa, <http://artfl-project.uchicago.edu/content/tfa> [accessed 03-27-2018]), the *Base de Français Médiéval* – BFM (Guillot / Heiden / Lavrentiev / Marchello-Nizia / Prévost, Lyon, <http://bfm.ens-lyon.fr> [accessed 03-27-2018]) and the *Corpus représentatif des premiers textes français* – CoRPTeF (Guillot, Lyon, <http://corptef.ens-lyon.fr> [accessed 03-27-2018]), the editions of the two oldest texts have been redone for this purpose, i.e. the *Serments de Strasbourg*, 842 AD, and the *Séquence de sainte Eulalie*, ca. 900 AD).

6 Similar to the medieval stage of other Romance languages, Old French does not have a consistent orthographic norm. Each scribe of a manuscript realized the sound of a word in his own fashion, influenced both by random circumstances and by his dialect that could differ significantly from what we consider the standardized Old French language. Thus, we find a great variety of spellings for the same word that we assign, as graphical variants, to the canonical form that is the lemma of the dictionary entry (*cf.* Möhren 2015).

*semantic tree* of the respective dictionary entries. To achieve this, we have established a workflow for the graphical integration as well as for the semantic mapping to the corresponding sense.

We merge the corpus and the dictionary data – graphically as well as semantically – based on the following assumption: The DEAF dictionary entries constitute a lightweight ontology. Note that this ontology is characterized by a very low degree of axiomatization as opposed to other, more formalized types of ontologies within the ontology spectrum, such as taxonomies and logical languages (*cf.* Grimm et al. 2011: 522–525). In this ontology, we understand the dictionary entries, their hierarchical organization into main- and sub-lemmata, and their respective semantic trees of *main-senses* and *sub-senses* as entities. This ontology constitutes the framework for the integration of the corpus data: We assign the lexical units of DocLing to the concepts of the ontology. These entities are represented in the data format.

## 4 Integration with Semi-Automated Workflow

The integration of the DocLing data into the dictionary is carried out on both the level of the back-end and on the front-end, i.e., on the level of the applications and also on the level of their graphical user interfaces (GUIs). In the following, we refer to the DocLing application as Phoenix2. The DEAF application is the tailor-made dictionary writing system, in the following referred to as DEAF-DWS.

### 4.1 Mapping of the Data Models

On the level of the back-end we have implemented a bidirectional data exchange: Each word occurrence plus a specific set of metadata is imported from Phoenix2 into DEAF-DWS, edited there and then written back to Phoenix2.

Naturally, the data models differ between the two systems. The data models of Phoenix2 and DEAF-DWS represent the structure specific to the textual domain of DocLing and the lexicographical domain of the DEAF, respectively. Fortunately, the parts of the data models relevant to the integration match conceptually to a large degree: The basic data entity to model an attestation of a given lexeme in Phoenix2 is the *occurrence*; this entity corresponds closely to what is called the *fiche* in DEAF-DWS. Moreover, we identified a large overlap of metadata that are associated to DocLing occurrences and DEAF fiches, respectively. These metadata can easily be mapped onto each other: the written representation of the lexeme (called *surface* in Phoenix2, *Zettelwort* in DEAF-DWS), the part-of-speech information, the siglum (the abbreviation used for the text), the text-reference, the date of the text, and the scripta (i.e., the written form of a spoken dialect of Old French). Also, every DocLing occurrence is assigned to a *lemma*, and so is the DEAF fiche. The fiche is the basic data unit of the DEAF. It is the starting point for the editorial process. With the data import, the DEAF-DWS turns a DocLing *occurrence* into a DocLing fiche to make it compliant with the DEAF fiche. Conceptually, the DEAF-DWS treats DocLing fiches as a sibling type to the original DEAF fiche (i.e., they share a common supertype fiche). Each DocLing fiche corresponds to exactly one occurrence in DocLing. To allow for a mapping of a given DocLing fiche to the corresponding occurrence during the write-back to Phoenix2, each DocLing fiche contains the *occurrenceID* of its associated occurrence as an attribute. The *occurrenceID* is a unique identifier for each DocLing occurrence within Phoenix2 that had been imported.

The only integration-relevant difference between the data models, prior to the integration, was the handling of the lemma structure. DEAF-DWS models the lexemes as word families with one main-lemma and one to many sub-lemmata. In contrast, Phoenix2 categorized according to sub-lemmata, only.

To allow for a unique mapping of lemma entities in Phoenix2 to lemma entities in DEAF-DWS, we added the main-lemma to the Phoenix2 data model too.

The prerequisite for the data exchange is a *compliant set of lemmata* for both DocLing and DEAF. Originally, the DocLing data were lemmatized according to the Modern French orthographical norm. Thus, a preparatory step for the integration was to migrate the Modern French lemmata to Old French lemmata. This manual re-lemmatization necessarily followed the standard lemmatization of Old French conducted by the DEAF. This lemmatization is widely accepted as the norm of middle 12<sup>th</sup> century French. During this preparatory step, the 800,000 occurrences in Phoenix2 were attached to approximately 5,300 Old French lemmata.

## 4.2 Import and Export of Data

The applications communicate via a *REpresentational State Transfer* / REST web service (*cf.* Fielding 2000). The compliant set of lemmata of DocLing and DEAF is the foundation for the data exchange: For each lemma, we import the attached occurrences from Phoenix2 into DEAF-DWS. This import includes the following information that is assigned to a given occurrence in Phoenix2: *surface*, *lemmaPOS* (part-of-speech), *sigel* (siglum), *division* (text-reference), *date*, and *scripta*. These information units correlate with the DEAF data structure. In addition to this, the import includes *scriptorium* (where the document was written), *context* (the textual context of the occurrence), *URL* (of the electronic edition within the DocLing website), and, finally, the *occurrenceID*.

After its initial import into the DEAF-DWS, an occurrence ‘exists’ twice, i.e., as an occurrence in Phoenix2 and as a fiche in DEAF-DWS. To prevent data inconsistencies between these two representations of the same occurrence, each metadata property of an occurrence can only be modified by exactly one of the two systems. More specifically, only the DEAF-DWS is allowed to change the lemma assignment of an occurrence (technically, of a fiche that corresponds to an occurrence). All other metadata except the lemma must only be changed by Phoenix2, e.g., the date, scripta and scriptorium. After every data modification, the systems need to be synchronized in order to make the modification visible to both systems.

## 4.3 Workflow and Graphical User Interfaces

The workflow comprises (i) the import (from Phoenix2 into DEAF-DWS), the assignment and write back (to Phoenix2) of DocLing occurrences, and (ii) the graphical and semantic integration of the data into the respective dictionary entries. It consists of automated and manual steps. We have implemented a number of features including the necessary GUIs to the DEAF-DWS to perform these steps.

### 4.3.1 Lemma Assignment

The import of DocLing occurrences into the DEAF-DWS is triggered by hand. Figure. 1 shows the feature implemented for the import: The editor searches for a given lemma (in the respective field, e.g., *dame*) and imports all occurrences attached to the lemma. Technically, the lemma assignment is not carried out on the level of the lemmata, but on the level of the occurrences. However, the GUI displays the lemmata as the unit that is most familiar to the editor. Below the search field, the GUI displays the pending lemmata, i.e., DocLing lemmata that have not been assigned to a DEAF lemma (Fig. 1).

## Import DocLing

**Hauptlemma**  **Lemma**

Bitte das zu importierende Lemma eingeben. Zum Import mehrerer Lemmata Jokerzeichen verwenden. Dabei entspricht "%" einer beliebigen Anzahl von Zeichen. "\_" entspricht genau einem Zeichen. Beispiel: "a%" importiert alle Lemmata, die mit 'a' beginnen.

Bitte nur einmal drücken. Abgleich erfolgt asynchron. Zum Anzeigen des aktuellen Abgleichs [Seite neu laden](#).

---

**Zuzuordnende Lemmata**

Die folgenden DocLing-Lemmata konnten nicht eindeutig zu DEAF-Lemmata zugeordnet werden. Bitte nehmen Sie die Zuordnung manuell vor. (Taucht ein Lemma mehrfach in der Tabelle auf, wurde es in mehreren Importen als ausstehend markiert. Es muss dann auch mehrfach zugeordnet werden.)

<< < > >>

DocLingLemma	
paire	<a href="#">manuell zuordnen</a>
porcol	<a href="#">manuell zuordnen</a>
prouveu	<a href="#">manuell zuordnen</a>
raplegier	<a href="#">manuell zuordnen</a>
raquester	<a href="#">manuell zuordnen</a>
roage	<a href="#">manuell zuordnen</a>
succession	<a href="#">manuell zuordnen</a>
traitier	<a href="#">manuell zuordnen</a>
trecens	<a href="#">manuell zuordnen</a>
tresque1	<a href="#">manuell zuordnen</a>

Figure 1: Import of DocLing occurrences (via a lemma).

By clicking on the button for the lemma assignment ('*manuell zuordnen*'), the interface as shown in Fig. 2 opens up. The assignment needs to be done manually whenever the given lemma is imported for the first time. The DEAF system supports this step by suggesting a main-lemma-sub-lemma combination it finds within the DEAF data where the sub-lemma matches the incoming DocLing lemma. In our example in Fig. 2, the lemma to assign is *succession* (subst. fem.) that is a sub-lemma of the main-lemma *succeder* (verb) within the DEAF-DWS.

## Zuordnung von DocLing-Lemma *succession* zu einem DEAF-Lemma

### Auswahl DEAF-Lemma

Suche mit Jokerzeichen: "%" entspricht einer beliebigen Anzahl von Zeichen. "\_" entspricht genau einem Zeichen.

Showing 1 to 1 of 1 << < > >>

Hauptlemma	Unterlemma	zuordnen
succeder	succession	<a href="#">zuordnen</a>

[Abbrechen](#)

### Okkurrenzen

<< < > >>

#### Wort im Kontext / Metadaten Okkurrenz

dans , à Aville , à Boans , à Montboson , à Fontenoy , à Roches , à Sorans , à Dampierre , à Folains , à Athoisons , à Chonoche , à Chambornay / . à Cromari , à Venise / . et en plusours autres leus / . en homes / . en lour **successions** / . en chans , en prez , en boys , en aygues et en lour decors / . en dismes , en rantes , en censes , en menaides , en justises et en totes autres droitures / . en chesaux , en maisons , en cultis , en hoches , en fourz , en es [↗](#)  
1280/02/06 — chHS111 5 **Red.**: AbbBellevaux! **Scripta**: frcomt. **Texttyp**: donation

eschat , soit d ' eschange / . soit de gaigliere ou de queque autre meniere d ' esquat / . soit de noz fiez et de nos rerefiez / . de noz demenuyres / . de noz mes taillables ou non taillables / . en homes et en lour **successions** / . en chans , en prez , en vignes , en vergiers / . en boys , en aigues et en lour decors / . en pescheries / . en dismes , en rantes en , en menaydes , en justises / . et en autres droitures / . en fourz , en estanz / . en m [↗](#)  
1280/02/06 — chHS111 9 **Red.**: AbbBellevaux! **Scripta**: frcomt. **Texttyp**: donation

a terre / . dois Donperré en lou en alant vers Colopné qui du de part le pere et de part la mere des diz freres / . et Humberes voussit auc ? lou disme de Saint Aignin ? pour ce que il disoit que il n ' estoit pris de lour **succession** de lour pere et de lour mere et por plusours autres raisons . Je prononce en declairant et veul que il diz dismes seilt entierement à dit mon signor Jehan derichief com li diz Humberes en ? de que li freres de la vigne de Bina [↗](#)  
1294/04/00 — chJu093 10 **Red.**: CAuxerre! **Scripta**: bourg. **Texttyp**: arbitrage?

tes sept cenx lb . tom . de rente sanz passer as autres hoirs / . descendenz de eus / . et en soient verai heritier sanz empeeschement , / . non contreitant la coustume de Normandie de la garde des non aalgiez / . et de la **succession** de ainzneesté / . et toutes autres coustumes qui en la succession de la dite rante leur porroit fere prejudice ou empeeschement en aucune maniere , ou as conditions / . et convenances dessus dites . / . Toutes les queles coustum [↗](#)  
1298/10/32 — R 1298 10 32 02 109 **Red.**: ChR **Scripta**: Paris **Texttyp**: Lettre patente en forme de charte

descendenz de eus / . et en soient verai heritier sanz empeeschement , / . non contreitant la coustume de Normandie de la garde des non aalgiez / . et de la succession de ainzneesté / . et toutes autres coustumes qui en la **succession** de la dite rante leur porroit fere prejudice ou empeeschement en aucune maniere , ou as conditions / . et convenances dessus dites . / . Toutes les queles coustumes contraires à ce / . ou aucunes des celes , / . de certeine scien [↗](#)  
1298/10/32 — R 1298 10 32 02 110 **Red.**: ChR **Scripta**: Paris **Texttyp**: Lettre patente en forme de charte

Figure 2: Assignment of DocLing occurrences to a DEAF main-lemma-sub-lemma combination.



In cases when the system finds several homonyms for the given lemma it will display all possibly fitting main-lemma-sub-lemma combinations for the editor to choose from. In cases when none of the suggested combinations is the correct one or the lemma in question is yet unknown to the DEAF system the editor can manually create a main-lemma-sub-lemma combination (Fig. 3).

**Zuordnung von DocLing-Lemma succession zu einem DEAF-Lemma**

**Auswahl DEAF-Lemma**

Suche mit Jokerzeichen: "%" entspricht einer beliebigen Anzahl von Zeichen. "\_" entspricht genau einem Zeichen.

**Filtern** **Aufheben** **Neu** Showing 1 to 1 of 1 << < 1 > >> **Gehe zu Seite**

Hauptlemma	Unterlemma	zuordnen
succeder	succession	zuordnen

**Okkurrenzen**

**Wort im Kontext / Metadaten Okk**

dans , à Aville , à Boans , à Montbos  
Venise /, et en plusours autres leus /,  
censes , en menaldes , en justises et  
1280/02/06 — chHS111 5 **Red.:**A

eschat , soit d ' échange //, soit de  
taillables ou non taillables //, en hom  
pescheries /, en dismes , en rantes e  
1280/02/06 — chHS111 9 **Red.:**A

a terre /, dois Donperré en lou en ala  
pour ce que il disoit que il n ' estoit p  
dismes seit entierement à dit mon sig  
1294/04/00 — chJu093 10 **Red.:**C

tes sept cenx lb . torn . de rente sanz  
de Normandie de la garde des non a  
prejudice ou empeeschement en auc  
1298/10/32 — R 1298 10 32 02 105 **Red.:**ChR **Scripta:** Paris **Texttyp:** Lettre patente en forme de charte

descendenz de eus /, et en soient verai heritier sanz empeeschement , /, non contreitant la coustume de Normandie de la garde des non aalgiez /, et de la  
succession de ainzneesté /, et toutes autres coustumes qui en la **succession** de la dite rante leur porroit fere prejudice ou empeeschement en aucune maniere , ou  
as conditions /, et convenances dessus dites . /, Toutes les queles coustumes contraires à ce /, ou aucunes des celes , /, de certaine scien  
1298/10/32 — R 1298 10 32 02 110 **Red.:**ChR **Scripta:** Paris **Texttyp:** Lettre patente en forme de charte

**Hinzufügen**

**Hauptlemma**

**Unterlemma**

**Speichern** **Abbrechen**

**Abbrechen**

<< < 1 > >>

Figure 3: Creation of a new main-lemma-sub-lemma combination.

The result of the lemma assignment is presented in a second table (not shown). The export of the data to Phoenix2 is again triggered by hand (it can also be done at a later date). Via the REST web service, the export writes back the main-lemma-sub-lemma combination for each DocLing fiche to the corresponding DocLing occurrence in Phoenix2. The lemma assignment needs to be performed only for the first import. When a second import of the same lemma is triggered, the DEAF system will perform the assignment automatically based on the already existing information in DEAF-DB. As mentioned earlier, after the successful import-export procedure each Phoenix2 occurrence now has a corresponding DocLing fiche representation in DEAF-DWS.

To be able to track all imports and write-backs, detect errors, etc., all processes are stored in a third table (not shown).

#### 4.3.2 Integration into a Dictionary Entry

After the data import and lemma assignment, the corpus data need to be merged into the respective dictionary entries. For this purpose, the DEAF-DWS displays the DocLing fiches within two GUIs, one for the editing of the graphical apparatus and one for the editing of the semantic part of the dictionary entry.



Kurzartikel: *mouture*

Sortierung Unterlemma

mouture mouturage mouturance mouturer

Wörterbuchblock Titelblock **Graphien** Bedeutungen Kommentar

## Alle Primärzettel zum Lemma

## Manuelle Sortierung

Nach Zettelwort/chron. sortieren

&lt;&lt; &lt; 1 2 3 &gt;&gt;

	Zettelwort	Wortart	Datum	Sigel	Scripta	Stelle	Varianten	Definition	Rest	Reihenfolge	Graphie	Verstecken?
0	<a href="#">meutire</a>		av. 1227-1265	ChansArtB	art.	XVII 87; XXII 216		taxe perçue pour la mouture du blé; redevance en farine				<input type="checkbox"/>
1	<a href="#">meutire</a>		1270/11/29	chdouai0473a	Nord	30						<input type="checkbox"/>
2	<a href="#">meutire</a>		1270/11/29	chdouai0473a	Nord	30						<input type="checkbox"/>
3	<a href="#">meutire</a>		1270/11/29	chdouai0473b	Nord	5						<input type="checkbox"/>
4	<a href="#">meutire</a>		1270/11/29	chdouai0473c	Nord	4						<input type="checkbox"/>
5	<a href="#">meutire</a>	dates mult.	MontRayn			II 33		mouture				<input type="checkbox"/>
6	<a href="#">meutire</a>	dates mult.	NoomenFabl			t.9, 348a		grain moulu	110 70			<input type="checkbox"/>
7	<a href="#">mo(u)ture</a>		1271/01/01	ChMe001	lorr.	3						<input type="checkbox"/>
8	<a href="#">moeture</a>		1270/11/29	chdouai0473b	Nord	5						<input type="checkbox"/>
9	<a href="#">moeture</a>		1270/11/29	chdouai0473c	Nord	4						<input type="checkbox"/>
10	<a href="#">moltire</a>		1237/01/19	chMM006	lorr.	12						<input type="checkbox"/>
11	<a href="#">moltire</a>		1270/05/00	ChSL 005	bourg.	28						<input type="checkbox"/>

Figure 4: Editing feature for the graphical apparatus of a dictionary entry.

Figure 4 shows the GUI for the editing of the apparatus of graphical variants. The DEAF-DWS classifies the graphical realizations attested by the imported DocLing data within the apparatus. Using *surface* of the DocLing fiche and *Zettelwort* of the DEAF fiche, it arranges all graphical variants in alphabetical order and merges the DocLing data with the original DEAF data. Within the alphabetical order it collates the attestations in chronological order and, as a third assorting step, also in alphabetical order of the sigla (using the DEAF metadata units *Datierung* and *Sigel* / DocLing metadata units *date* and *sigel*). This is done in a fully automated way. Moreover, options for a manual post-processing are also provided. Note that the merged DEAF fiches and DocLing fiches are displayed within the same table for the convenience of the editor. To be distinguishable, DEAF fiches and DocLing fiches are displayed in white/blue and in shades of green, respectively.

The semantic integration of the DocLing material needs to be performed manually with linguistic expertise, and this process has recently been started. The editing process of the entries of DEAF*pré*, on the other hand, were completed in 2017. As such, the starting point for the semantic integration of the DocLing data into an entry is a completed DEAF*pré* entry with the semantic tree already established. Figure 5 shows the semantic tree (*'Bedeutungsbaum'*) for the lexeme *succession* (subst. fem.) with one main-sense and one sub-sense. We can see that the (white and blue) DEAF fiches have already been assigned to the proper sense and they are displayed in a table on the right-hand side of the GUI. The newly added (green) DocLing fiches initially appear in the table on the left-hand side. Each of them needs to be assigned to the correct sense of the semantic tree. The GUI offers several drag-and-drop mechanisms and other features to do this in a time-saving way.

As soon as all DocLing fiches are merged, the updated dictionary entry can be exported as an XHTML file that is used for the online publication.

Kurzartikel: *succeder* Sortierung Unterlemma

succeder succederesse succeneor succeseor successeresse successerle successif **succession** successivement

Wörterbuchblock Titelblock Graphien **Bedeutungen** Kommentar

DocLing-Zettel (5) Alle Zettel

Wort im Kontext / Metadaten Zettel >

dans , à Aville , à Boans , à Montboson , à Fontenoy , à Roches , à Sorans , à Dampierre , à Folaïns , à Athoissons , à Chronoche , à Chambornay , à Cromari , à Venise /, et en plusieurs autres lieux /, en homes /, en leur successions /, en chans , en prez , en boys , en aygues et en leur decors /, en dismes , en rantes , en censes , en menades , en justises et en totes autres droitures /, en chesaux , en maisons , en cultis , en hoches , en fourz , en es ☐

1280/02/06 — chHS111 5 Red.:AbbBelevaux! Scripta:fromt. Texttyp: donation

eschat , soit d' eschange //, soit de gaignere ou de queque autre meniere d' esquat //, soit de noz fiez et de nos renefiez //, de noz demenuyres //, de noz mes taillables ou non taillables //, en homes et en leur successions //, en chans , en prez , en vignes , en vergiers //, en boys , en aigues et en leur decors //, en pescheries //, en dismes , en rantes en , en menaydes , en justises //, et en autres droitures //, en fourz , en estanz //, en m. ☐

1280/02/06 — chHS111 9 Red.:AbbBelevaux! Scripta:fromt. Texttyp: donation

a terre /, dois Donperré en lou en alant vers Colopné qui du de part le pere et de part la mere des diz freres /, et Humberz voussit auc ? lou disme de Saint Aigrin ? pour ce que il disoit que il n' estoit pris de leur succession de leur pere et de leur mere et por plusieurs autres raisons . Je prononce en declarant et veul que il diz diemes sell entièrement à dit mon signor Jehan derchief com il diz Humberz en ? de que il freres de la vigne de Bina ☐

1294/04/00 — chJu093 10 Red.:CAuxerre! Scripta:bourg. Texttyp: arbitrage?

tes sept cenx lb . torn . de rente sanz passer as autres hoirs /, descendenz de eus /, et en soient verai heritier sanz empeeschement . /, non contraitant la coustume de Normandie de la garde des non aaignez /, et de la succession de ainznesté /, et toutes autres coustumes qui en la succession de la cite rante leur porroit fere prejudice ou empeeschement en aucune maniere , ou as conditions /, et convenances dessus dites . /, Toutes les quelies coustum ☐

1298/10/32 — R 1298 10 32 02 109 Red.:ChR Scripta:Paris Texttyp: Lettre patente en forme de charte

descendenz de eus /, et en soient verai heritier sanz empeeschement . /, non contraitant la coustume de Normandie de la garde des non aaignez /, et de la succession de ainznesté /, et

Zettel in Wörterbuch- und Bedeutungsblock sortieren

Bedeutungsbaum

Alle ausklappen — Alle einklappen — Zeige Wurzel

Neue Hauptbedeutung Neue Unterbedeutung Löschen XML Speichern

↳ héritage, biens qu'une personne laisse en mo[...]  
↳ série de personnes ou de choses qui se suiv[...]

<definition>série de personnes ou de choses qui se suivent en se remplaçant l'une l'autre, sans interruption ou à peu d'intervalle  
</definition>

à br case case-ending case-form citation cited-letters cited-word collocation comment compound definition designation evidence flexion flexion-ending flexion-form footnote gram hebr idem link locution manuscript onomasiology pos proverb quotation remark remark-reference sc siglum sup terminology text-reference usage

Primärzettel (4) - Sortieren << >>

≤ Zetteliwort Wortart Definition Datum Sigel Stelle Rest

Figure 5: Editing feature for the semantic part of an entry.

## 5 Online Publication

With the integration of the DocLing material, the entries of DEAF<sup>pré</sup> are enriched with attestations that are integrated into the apparatus of graphical variants and into the semantic part.

As we have shown above, the graphical integration of the corpus material is performed automatically by the DEAF-DWS. However, the manual integration of the DocLing material into the semantic structure of each article is time-consuming, and will be performed gradually in the years to come. To compensate for the long-lasting workflow, we implemented two release steps for the publication. As a first release step, we created a preliminary publication that is the result of a fully automated process. This enables us to give online access to the valuable new material before the task of the semantic integration will be accomplished. We execute the second release step after the manual post-processing has been completed, i.e., after the semantic integration of the DocLing data.

### 5.1 Release Step #1: Automated Processing

For the display of the new material, we have modified the design of the online publication of DEAF<sup>pré</sup>.

The modification of the entry's graphical apparatus was straightforward. We display each DocLing attestation with its siglum and text-reference as we do with the original DEAF attestations (Fig. 6).

#### MOUTURE f.

[FEW 6<sup>3</sup>, 42b lt. \*MOLITURA – TL 6,374,1; TL 6,374,1; Gdf; Gdf; Gdf; GdfC 10, 167a; AND; AND; MED 6,796b; DCCarp 273b; TLF 11, S 1177 b, 1240; FEW; FEW 43b; FEW VI/ 3, 42b; Arnaldi p 264; DiStefLoc 568. – Bev 28; Bev 28; Drüppel 37, 86; Drüppel 86; Morlet 127; Morlet 263; [sigle].]

(*meutire* ChansArtB XVII 87; XXII 216; chdouai0473a 30; chdouai0473a 30; chdouai0473b 5; chdouai0473c 4; (sigles à datations multiples:) MontRayn II 33; NoomenFabl t.9, 348a, *mo[u]ture* ChMe001 3, *moeture* chdouai0473b 5; chdouai0473c 4, *molture* chMM006 12; ChSL 005 28; CensToulO; JurésSQuenD 716A, 717D, *mosture* chMM027 3, *motture* chMe009 11, *motture* DocHMarneG; DocHMarneG; chMe009 4; chHM078 5; chMe219 5; ChHM253 10; ChHM253 11; chHM267 5;

Figure 6: DEAF<sup>pré</sup> entry *mouture* (subst. fem.): (headword, dictionaries, secondary literature, and) part of the graphical apparatus.

To enable the user to recognize the origin of the DocLing material we display it in a color different from the one used for the DEAF attestation. This is important because the DocLing material is of a significantly better quality compared to the original DEAF*pré* material. The attestations given in a DEAF*pré* article are not verified in the sources, i.e., in the editions of the primary texts. The reason for this major flaw is the very limited time that the two-fold concept described above allowed for the editing of the DEAF*pré* articles (and it is the most significant difference to DEAF*plus* where every information is verified). In contrast to this, the DocLing material is sound evidence that deserves to be identifiable as such.

The fact that we import the context of each attestation, the URL and the other metadata from Phoenix2 allows us to make this information accessible to the user. By clicking on any DocLing attestation, the user can display this information (Fig. 7). The button ‘Ouvrir ce passage dans DocLing’ provides the hyperlink to the respective document within the DocLing website.

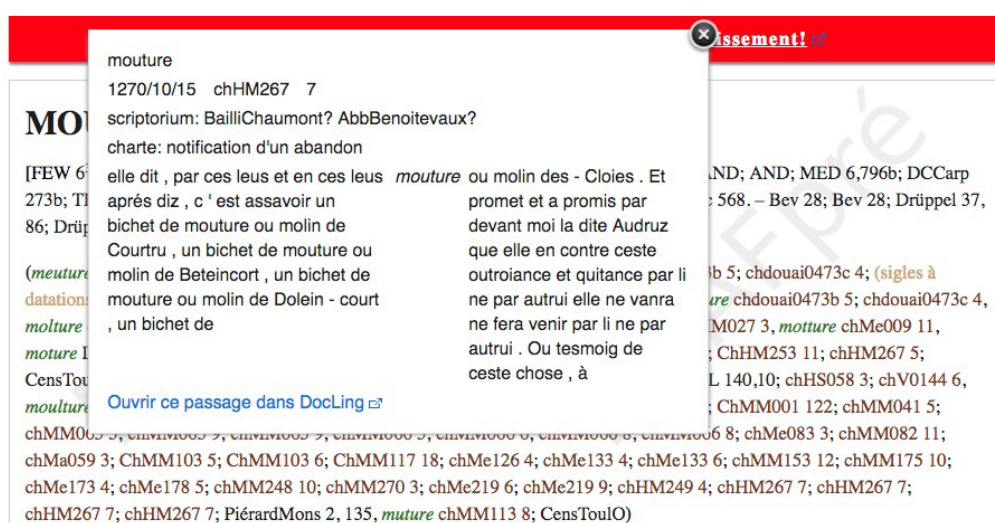


Figure 7: Display of a DocLing fiche within DEAF*pré*.

For the creation of the semantic part of the entry, the export routines of the DEAF-DWS place all DocLing material (again with attestations and text-references) that is not yet properly semantically integrated into a container. In the online publication, we display this container as clearly distinguishable addenda (‘*Identificanda* DocLing’) to the semantic tree of the respective entry (Fig. 8).



Figure 8: DEAF*pré* entry *mouture* (subst. fem.): semantic part with container ‘*Identificanda* DocLing’.



## 5.2 Release Step #2: Manual Post-Processing

The second release step results from the accomplished task of the semantic integration (and thus it does not concern the graphical apparatus). The publication merges all attestations in the semantic tree, as shown for the first main-sense “*travail de moudre du blé et sim.*” of the lexeme *mouture* in Fig. 9.

◆ 1<sup>o</sup> “travail de moudre du blé et sim.” (ChMM001 122; chMe009 4; chMe009 11; chMM066 6; chMM066 8; CensToulO; CensToulO; JurésSOuenD 716A, 717D; PiérardMons 2, 135; (sigles à datations multiples:) MontRayn II 33; NyströmMén VIIIb 107, Bev 28; Bev 28; Drüppel 86; Drüppel 37, 86; Morlet 127; [sigle], TL 6,374,1; TL 6,374,1; Gdf; Gdf; Gdf; GdfC 10, 167a; AND; AND; MED 6,796b; FEW VI/ 3, 42b; FEW)  
 ◆ “grain moulu” (DocAubeC 48-22; DocHMarneG; chMM027 3; chMM041 5; CensToulO 1,25v<sup>o</sup>, 26; 13v<sup>o</sup>; 30v<sup>o</sup>; 25,26,31,34;

Figure 9: DEAF*pré* entry *mouture* (subst. fem.): semantic part with semantically integrated DocLing attestations.

## 6 Added Value for Both the Dictionary and Corpus

Our aim is to merge the data with mutual benefit for the corpus project and for the dictionary. From the dictionary’s point of view, the vocabulary of the DocLing corpus texts enriches the dictionary’s material with the medieval language of law which previously had been widely unregarded by Old French lexicography. It helps to develop the comprehension of the lexemes in a considerable way, producing added value within the historical lexicography of Old French. This clearly extends the limits of the traditional historical dictionary. The added value is also of specific interest for the historical sciences focusing on medieval law and the application of law that is witnessed in the documentary sources. We foresee that the new source material will shed light on the senses of many lexemes of DEAF*pré*, in particular because it represents the discourse tradition of legal documents. Therefore, the DocLing material will add to a better understanding of the semantic scope of these lexemes. As a consequence, the editor’s task while integrating the DocLing material will be to evaluate the semantic tree of the DEAF*pré* entry and to improve and expand it if necessary. This will increase the quality of the DEAF*pré* entries in a significant way.

From the corpus’ point of view, one benefit is the semantic disambiguation of the corpus data. Traditionally, digital text editions – both as a single publication and as a part of a larger corpus – are stand-alone products. The publication usually does not offer an instrument (e.g., a comprehensive glossary) that supports the reader to understand the text. With the integration of the data into the dictionary we create a means that helps the reader to grasp the comprehension of the vocabulary. Also, with the integration, we embed the specific juridical vocabulary of the DocLing texts within the overall system of the Medieval French language as it is established by the DEAF. This reveals the place of the lexical units attested in the corpus within the broader semantic range of the Old French lexicon and the significance of the vocabulary within the history of the language.

## 7 Establishing Links from the Corpus to the Dictionary

A remaining issue is how to establish a link from a given word occurrence within a DocLing text edition (Fig. 10) back to the publication of the respective entry in DEAF*él*.

## chdouaioo05

1225 (n.st.), février.

TYPE DE DOCUMENT: vente.

OBJET: Vente par Rainier de "Gorghechon" de 8 muids de terre à Jean "del Cerf" et Wagon de Saint-Albin.

AUTEUR: Rainier de Gorguechon, chevalier.

DISPOSANT: L'auteur.

SCEAU: non scellé; devise chirographaire: CY-RO-GRA-PHE

BENEFICIAIRE: Jean du Cerf, Wagon de Saint-Albin, bourgeois de Douai.

AUTRES ACTEURS: plèges et otages: (i) chevaliers: Gautier de Jenlain, Gilles de Symion(?), Robert d'Artres, Amand de Rouvignies, Hugues de Markete; (ii) écuyers: Gautier de Monchecourt, Thierry de Douchy, Gilles de Gorchon(?), Rainier de Gorguechon. - échevins de Douai: Heuvin Malet, Simon le Conestable.

SUPPORT: Original parchemin chirographe superposé bipartite - partie supérieure.

LIEU DE CONSERVATION: AM Douai, FF 657/5626.

ÉDITION ANTÉRIEURE: Bonnier, "Etude critique...", n°2, p. 298. - Gysseling, "Les plus anciens textes français non littéraires...", n° 24, p. 205-206.

1 Ce sacent tout cil ki or sunt *et* ki a-venir sunt 2 que jou Rainiers de Gorghechon, chevaliers, ai vendut a Jehan del Cerf *et* a Wagon de Saint Aubin, <sup>13</sup> borgois de Dowai, .VIII. muis de terre en tous preus prendans [dusques] <sup>[1]</sup> <sup>14</sup> a .VII. ans a la mesure de Dowai *et* li tierce pars de ces [...] [VIII.] [muis] [de] <sup>[2]</sup> <sup>15</sup> tere ne doit ne disme ne terage ne rente ne service *et* leus .II. p[ars] [...] <sup>[3]</sup> <sup>16</sup> droite disme *et* ces .VIII. muis de tere doivent li bourgeois keusir [...] <sup>[4]</sup> <sup>17</sup> tere; *et* a cest premerain aoust *que* nos atendons, doivent il prendre .XVI. r. <sup>18</sup> de blet le semure *et* .XVI. r. de marc tout avestit a-prendre en quel liu <sup>19</sup> qu'il volront de toutes mes teres ki sunt semencies; *et* quant cis premiers aous <sup>20</sup> sera passés, il doivent avoir ces .VIII. muis de tere ki devant sunt nomet <sup>21</sup> por faire leur volenté dusques adonc que li termes sera passés, ki ci devant est <sup>22</sup> només; *et* s'il avenoit

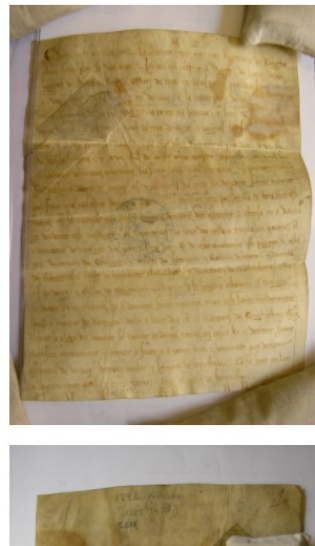


Figure 10: Online edition of a document in DocLing.

Based on the successful integration of the DocLing data into the DEAF database and the semantic mapping of the lexical units, we foresee the possibility to establish an automated interlinking process using the *occurrenceID* of each DocLing fiche and DocLing occurrence, respectively. This needs to be further evaluated in a follow-up study.

Independent from the above-described data integration is an approach that parts from the XML/TEI data of a digital text edition, as described in Tittel, Bermúdez-Sabel and Chiarcos (accepted paper): With the insertion of RDFa compliant attributes (*cf.* Herman et al. 2015) into the existing XML elements, the data of the text edition can automatically be enriched with hyperlinks to the DEAF dictionary. The fact that the DocLing corpus texts are published in an XML/TEI format (TEI Consortium 2017) makes this a promising approach for the creation of references from DocLing to DEAF.

## 8 Conclusion

To the best of our knowledge, this is the only example of a retrospective and successful integration of two voluminous and longstanding projects of a historical (Romance) language both from a graphical and a semantic point of view. We have implemented a semi-automated process resulting in a time-saving editorial workflow. We argue that our approach to fully integrate the corpus data of DocLing is a promising way to solve the problem of semantic mapping. We show how to perform the semantic mapping of the lexical units of DocLing using an existing dictionary writing system. A flexible publication concept with two release steps compensates for the time-consuming semantic integration, as it makes it possible to publish two versions of each dictionary entry: one that is created in a completely automated way, and another that shows the result of the manual post-processing with linguistic expertise. This clearly contradicts Asmussen 2013: 1087: "Combining existing dictionaries with existing corpora will inevitably yield products of second quality".



Also, we conclude that the integration of DocLing and DEAF is a promising pilot project for the integration of other corpus linguistic data into a dictionary. At the same time, it emphasizes the role of the DEAF as a standard reference that can also be used for other single scholarly editions of Old French texts that are digitally published.

## References

- Asmussen, J. (2013). Combined products: Dictionary and Corpus. In R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (Handbücher zur Sprach- und Kommunikationswissenschaft / HSK 5.4). Berlin / Boston: De Gruyter, pp. 1081–1090.
- Baldinger, K. (founder), continued by F. Möhren, published under the direction of T. Städtler (1971–). *Dictionnaire étymologique de l'ancien français – DEAF*. Québec / Tübingen / Berlin: Presses de L'Université Laval / Niemeyer / De Gruyter; electronic version DEAFél accessed at: <https://deaf-server.adw.uni-heidelberg.de> [03-28-2018].
- Fielding, R. T. (2000). Chapter 5: Representational State Transfer (REST). In *Architectural Styles and the Design of Network-based Software Architectures (Ph.D.)*. University of California, Irvine. Accessed at: [http://www.ics.uci.edu/~fielding/pubs/dissertation/rest\\_arch\\_style.htm](http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm) [03-31-2018].
- Glessgen, M.-D. (1998–). *Documents linguistiques galloromans. Édition électronique, Collection founded by Jacques Monfrin, continued by M.-D. Glessgen (with the collaboration of Hélène Carles, Frédéric Duval and Paul Videsott)*. Accessed at: <http://www.rose.uzh.ch/docling/> [03-28-2018].
- Grimm, S., Abecker, A., Völker, J., Studer, R. (2011). Ontologies and the Semantic Web. In J. Domingue, D. Fensel, J. A. Hendler (eds.) *Handbook of Semantic Web Technologies*. Heidelberg: Springer, pp. 507–579.
- Herman, I., Aside, B., McCarron, S., Birbeck, M. (2015). *RDFa Core 1.1 – Third Edition*. Accessed at: <https://www.w3.org/TR/rdfa-core> [03-28-2018].
- Kunstmann, P. (2007–2014). *DÉCT: Dictionnaire Électronique de Chrétien de Troyes*. LFA/Université d'Ottawa – ATILF/CNRS & Université de Lorraine. Accessed at: <http://www.atilf.fr/dect> [03-28-2018].
- Möhren, F. (2015). L'art du glossaire d'édition. In D. Trotter (ed.) *Manuel de la philologie de l'édition*. Berlin: De Gruyter, pp. 97–437.
- Sinclair, J. (1996). *Preliminary recommendations on Corpus Typology. Technical Report*. EAGLES (Expert Advisory Group on Language Engineering Standards). Accessed at: <http://www.ilc.cnr.it/EAGLES96/corpusstyp/corpusstyp.html> [03-30-2018].
- Tittel, S. (2010). Le « DEAF électronique » – un avenir pour la lexicographie. In *Revue de Linguistique Romane*, 74, pp. 301–311.
- Tittel, S., Bermúdez-Sabel, H., and Chiarcos, C. (accepted paper). Using RDFa to link text and dictionary data for Medieval French. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018)*, Miyazaki, Japan, May 2018.
- TEI Consortium (2017). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0. Last updated on 10th July 2017*. Accessed at: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/> [03-28-2018].



## Heritage Dictionaries, Historical Corpora and other Sources: Essential And Negligible Information

**Alina Villalva**

*Faculdade de Letras & Centro de Linguística da Universidade de Lisboa*

*E-mail: alinavillalva@campus.ul.pt*

### Abstract

Contemporary dictionaries are often the result of accumulating the contents of previous dictionaries, namely heritage or patrimonial dictionaries, which are those that set a landmark in the historical lexicography of a given language. This is certainly the case for European Portuguese dictionaries, which offer word lists containing words in usage as well as rarely used words, words that belong to other language variants and phonetic/orthographic and morphological alternates, being apparently unable to distinguish among relevant, puzzling, or simply useless information.

If contemporary dictionaries were to be redesigned, should lexicographic ancestry be ignored, or should it be dealt with otherwise? The discussion on new lexicographic models is beyond the range of this paper, but any lexicographic innovation must be grounded in solid lexicological analyses that cannot ignore the consideration of heritage dictionaries and historical *corpora*. These sources provide a large quantity of information requiring specialized interpretation and critical trimming, which may nevertheless be insufficient to fully grasp a thorough knowledge of the related words. This paradox, which is the focus of the present work, will be addressed by advocating the need to adopt a lexicographic protocol that may help to select the right amount of information and combine it with identical information from other languages.

**Keywords:** heritage dictionaries, historical corpora, European Roots

Studying words is a lengthy task that relies on the collection of information, and on the ability to interpret the relevant bits and to relate them in a meaningful way. Heritage dictionaries, crafted according to coeval word knowledge, convey some of this information, but these dictionaries are the privilege of only some languages and some moments in their history. Iberian languages, for instance, have had bilingual dictionaries since Nebrija's *Dictionarium*, published in 1492, but monolingual dictionaries, such as Bluteau's *Vocabulário Português e Latino*, with ten volumes published between 1712 and 1728, arrived considerably later. Furthermore, heritage dictionaries need to be read in their time. Bluteau's long entry on *molher* 'woman' looks like misogynist libel to our eyes, but it has to be interpreted culturally. Notice that, in this entry, Bluteau invokes several classical 'authorities', like Salomon, Diphilus, Socrates, Democritus, or Tacitus, to substantiate a negative portrait of women, which is crowned by a quotation of Scaliger's definition of *femina* "natura sua infida est, suspicax, inconstans, insidiosa, simulatrix, supersticiosa, cui si potentia adjuncta est, fit intolerabilis". Bluteau's entry shows to posterity how men, and particularly clergyman like himself, saw women, and that in itself is an interesting piece of information.

Therefore, not only do heritage dictionaries fail to cover the whole existence of older words, but they also provide insights biased by what was then the received state of the art. Hence, other possibilities of data collection need to be considered. Historical *corpora*, despite covering only a sample of written production, do help to fill in lexicographic gaps, particularly those that precede the appearance of early dictionaries, and they allow us to crosscheck lexicographic inputs with real usage.

Unfortunately, historical textual databases are not yet as powerful as we would like them to be – the coverage has to be improved, the philological quality needs to be certified, and text annotation requires a heavy workload. Time and academic contributions will probably help to overcome some shortcomings of these databases, but professional and specifically trained manpower allocation is essential to the much needed upgrading.

This said, it becomes self-evident that the consideration of heritage lexicographic and historical textual information raises numerous methodological problems, particularly those that are related to the fact that dealing with these sources requires specialized interpretation and critical trimming. This is indeed a gigantic task, which may benefit from the adoption of a lexicographic model such as the European Roots prototype<sup>1</sup> that I have described on several occasions (cf. Silvestre & Villalva 2014, Villalva & Silvestre 2015). The ER dictionary prototype aims to serve as a principled source of information for general monolingual and multilingual dictionary makers, since it is intended to help to select the right amount of information and to combine it with identical information from other languages.

According to this model, lexicologists of each language must identify the set of heritage and contemporary dictionaries that must be considered. In the case of Portuguese, the list includes eight items: Cardoso (1569), Barbosa (1611), Pereira (1697) and Bluteau (1712-1728), available online at *Corpus Lexicográfico do Português* (CLP), Figueiredo (1913), also available online at *Dicionário Aberto*, and Morais Silva (1813) that can be consulted online at the website of Biblioteca Brasileira Guita e José Mindlin. Two other online dictionaries (*Infopedia* and *Priberam*) embody the contemporary stage. Each of these dictionaries represented a turning point in the Portuguese lexicographic panorama at the moment when they were released – they are not exempt from criticism, but they are the best there is. The same applies to historical textual *corpora*. For Portuguese, we may use *Corpus do Português* (CdP), that includes texts from the 13<sup>th</sup> to the 20<sup>th</sup> centuries, and *Corpus de Referência do Português Contemporâneo* (CRPC), that comprises texts (and some transcripts from oral speech) from the mid-19<sup>th</sup> century to present-day.

Though the information provided by these sources is enormous, it may nevertheless be insufficient to fully grasp a thorough knowledge of the related words. This is probably the most serious issue and deserves further discussion, and it will be the focus of this paper, as demonstrated by the complexity of the analysis of the Portuguese word *laranja* ‘orange’, some cognates like the Castilian *naranja* and the French *orange*, and some etymologically unrelated equivalents like the Greek *πορτοκαλί* [portoka’li]<sup>2</sup>.

## 1 The Complexity of *Laranja*

Mabberley (2004: 483) claims that “the names of both citrus and orange are surrounded by a series of confusions, false etymologies and perhaps puns”. The Portuguese noun *laranja* is equally difficult to grasp, but it is a piece of the general puzzle that has not often been considered by itself. The widespread tendency to look at Portuguese as a variety of Castilian may have obfuscated any relevant specific information.

The etymological information on *laranja* ‘orange’ conveyed by Cunha (1986, s.v. *laranja*), a Portuguese etymological dictionary, refers to an Arabic etymon (e.g. *nāranġa*) that originates in a Sanskrit word (e.g. *nāraṅg*). Contemporary dictionaries, like *Infopedia* and *Priberam*, echo the same

1 This project is published at [sites.google.com/a/campus.ul.pt/european-roots/](http://sites.google.com/a/campus.ul.pt/european-roots/). Landlex, a more recent project devoted to the study of the landscape lexicon, has been testing the ER prototype, particularly in German and Modern Greek, besides Portuguese (cf. [sites.google.com/a/campus.ul.pt/landlex---lisbon-meeting/about-landlex](http://sites.google.com/a/campus.ul.pt/landlex---lisbon-meeting/about-landlex)).

2 I am indebted to Simeon Tsolakidis, from the University of Patras, who provided the information on Modern Greek *πορτοκαλί*.

information, regardless of how they spell these etymons<sup>3</sup>. Cunha also claims that the word first occurred in the 14<sup>th</sup> century, but this *a quo* terminus is not documented<sup>4</sup>. Since Portuguese was never in direct contact with Sanskrit, a vehicular language has indeed to be considered, but Arabic is probably not the best immediate candidate. In fact, language contact between Portuguese and Arabic existed for centuries, but it stopped after the definitive Portuguese conquest of the Algarve, in the mid-13<sup>th</sup> century, and even if we admit later commercial exchanges, we still need to explain how the initial consonant changed from the nasal [n] to the lateral [l], which is not a common phonetic change – this is a major setback for the Arabic loan hypothesis. Etymological research must be plausible, and, in this case, there is, at least, one missing link.

Where does Cunha's hypothesis come from? The etymological dictionaries by Corominas, though devoted to Castilian, were and still are very influential references for both Portuguese and Brazilian etymologists. Therefore, it makes sense reading Cunha's entry on *laranja* as an adaptation of Corominas (1981) entry on *naranja*, but the hypothesis of a parallel path for both words (i.e. *laranja* and *naranja*) is fragile, since it fails to explain the Portuguese phonetic output. Portuguese and Castilian etymologies are often quite similar, but in this case there is a slight distance to the closeness. Hence, the remote etymology of *laranja* may be solved, but there is no insight into how the word came to be what it is.

## 2 *Laranja* in Heritage Dictionaries

A survey of heritage dictionaries regarding the word *laranja* and some of its derivatives shows us that they are part of the entry list of them all, and they also evidence a growing descriptive complexity. Cardoso, Barbosa, Pereira and Bluteau all offer Latin equivalents to *laranja*, which seek to be explanatory, not a translation, since oranges were probably unfamiliar in the Roman Empire. Cardoso (1569) uses the generic noun *malum* (which meant fruit in Latin), and the adjective *medicum* to describe *laranja* and close explanations for the derivatives *laranjada* and *laranjol*. In this entry, he adds the word *arantia(ae)*, unattested elsewhere to our knowledge, but it looks like a Latinized version of the Italian word *arancia*. Another single time, in the entry regarding *malus syria* he presents it as an equivalent to *malus medica*, both of which correspond to the Portuguese *larangeira* 'orange tree'. The reference to Syria invokes what Cardoso probably believed to be the regional origin of the species.

- |  |   |
|--|---|
| <p>(1) Laranja. Malum medicum. arantia(ae).<br/>         Laranjeira. Malus medica.</p> <p>Laranjada. Medicatus(us).<br/>         Laranjal. Medicetum(i).</p> | <p>Malum medicum / Medicum malum. A laranja.<br/>         Malus medica / Medica arbor. a laranieyra.<br/>         Malus syria. siue malus medica. A laranjeyra.</p> |
|--|---|

Forty years later, Barbosa (1611) almost replicates Cardoso's input. *Malum medicum* is the first Latin equivalent that he provides, but he also offers an alternative, i.e. *mala aurantia*, quoting the Italian Renaissance botanist Mattioli (1501-1577). It is worth noting that Pianigiani (1907) relates *arancio* to a Vulgar Latin *arangia* and *aurantia*, and he identifies a folk etymology link with Latin *aurum* 'gold' (cf. "accostato per etimologia popolare al Lat. AURUM"). Possibly, Mattioli inaugurated the adjustment of the Italian word to a Latin 'motivated' etymon, and Barbosa followed him. On the other hand, the mention to Syria disappears, but a close reference, to the *media regione*, shows up.

3 Cf. *Infopedia*: Do persa *narang*, «*laranja*», pelo árabe *naranjâ*, «*idem*»; *Priberam*: (árabe *narandja*).

4 Lorenzo (1968: 220) quotes an anthology of medieval documents where *laranja* occurs, in 1377.



- (2) Laranja. Malum Medicum. Vel, mala aurantia, Mathiolus in lib. 1. Dioscor. c. 131.  
 Lorangeyra. Malus Medica. Arbor Medica. Plin. lib. 12. cap. 3. Dicta est Medica a Media regione.  
 Laranjal. Locus malis Medicis consitus.

Almost a century later, Pereira (1697) also relates *laranja* to *malum medicum*, and he adds a new Latin equivalent (i.e. *malum aureum*), thus making the metaphorical connection with ‘gold’ even more explicit. His Latin-Portuguese dictionary is yet more prolific. *Malum aureum* is presented as a variant of *malum aurantium* (he may well be just trying to ameliorate the Latin form of the expression), but Cardoso’s *arantia* (characterized as a new word) and Barbosa’s *aurantia* are somehow retrieved. Apparently, they were circling the Italian word and seeking for a plausible Latin etymon, a task doomed to failure that even considered the more bizarre word *anatarantium*, marked by Pereira as never found. Then, Pereira lists a series of *malum* variants that partly refer to prior knowledge (i.e. *assyrium*, *medicum*) and partly bring novelty (i.e. *hesperidum*, *citream*). *Malum hesperidum* is probably a name influenced by Ferrari’s recent treaty<sup>5</sup> that located the origin of these fruits in the mythological Garden of the Hesperides. The second innovation brings *citream* into play, and a whole new series of interrogations raised by the equivalence between *laranja*, *cedromelon* and *cidra* that we will consider below, and *narantrium* and *nerantzum* that look like the Latinization of the Sanskrit etymon. Finally, Pereira accounts for two new meanings for *laranjada*. The first one is related to a strike with oranges, and the second meaning identifies an orange preserve. The amount of information found in Pereira is thus quite considerable, but not proportionally as instructive.

- |   |   |
|---|---|
| (3) Laranja. Malum medicum. Malum aureum. | Malum (Aurantium) aureum. A laranja. Jun.<br>Arantium, ii, n. g. A laranja. Verbum novum. 1. b.<br>Anatarantium, ii, n. g. A laranja. Non inveni.<br>Aurantium, ii, n. g. A laranja. Amalth.<br>Malum Hesperidum. Assyrium. Medicum.<br>Citream. A cidra, o limam, a laranja. Jun.<br>*Cedromelon, i, n.g. A laranja, ou cidra.<br>1.c.2.b.3.l. Onom. M.<br>*Narantrium, ii, n.g. A laranja. 1. l. Nicand.<br>*Nerantzum, ii, n.g. A laranja. 1. b. L. G. B<br>Medica malus. Cidreira (a lorangeira) (Medicum<br>malum. A cidra, ou laranja, & c.). 1.l.2.b. Virg.<br>Georg. 2.!! |
|---|---|

Laranjada. Mali medici ictus.

Laranjada, id est, conserva de laranjas. Mala Medica saccharo cocta.

Laranjal. Locus malis medicis cōsitus.

Bluteau’s dictionary began to be published less than twenty years after Pereira’s, but it is substantially different. Though it also makes use of Latin equivalents, the essence of his *Vocabulario* is practically monolingual, and it is tentatively encyclopedic. In the case of *laranja*, Bluteau (1712-28) resists choosing a Latin equivalent. He claims that *laranja* is a ‘known fruit’, then he invokes Virgil and, mostly, he quotes Ferrari, who gives a list of names and justifications, most of which failed to pass later inquiries:

laranja. Fruto conhecido. Alguns lhe chamaõ Malum aureum. Virgilio diz, Aurea mala. Outros dizem Malum citream orbiculatum. Sobre os nomes Latinos, ou alatinados, que os Autores dão à laranja, diz o P. Ferrari nas suas Hesperides, liv. 1. pag. 43. Inter acida postremum poma sagacissimi conjectores, recentiore nomine appellant, vel Arantium ab Arantia, pomorum feracissimo Achaiae oppido;

<sup>5</sup> G. B. Ferrari (1646) *Hesperides siue de malorum aureorum cultura et vsu Libri quator Io*. Rome: Sumptibus Hermanni Scheus.

quo mala Hesperidum primum Hercules tulisse credebatur, vel Aranium, quasi Ararium, id est, Persicum: est enim Aroa, ut ait Hellanicus, allique, Perfidis regio, vel certe Rantium, tamquam Raedum, hoc est, aeris colore fulvum. Vel Neratio inventore Neratium, vel eum veteri Nicondri Scholiaste Necrantzion, vel (quod etiam Hermulao placet) Narantium à Narantia, quae Ptolomei videtur esse naranga, ex qua idem cum Pausania existimat ab Hercule id ponu fuisse in Graeciam asportatum. Vel demum quia, ut modo diximus, relucet auri colore, aurengium, malum aurantium, unaque expuncta littera Arantium, & aureum malum, quod veteres Hespericum etiam vocavere sed nondum potuit malum aurantium auri quod nomine praefert luce, suos satis demonstrare natales.

Neither Moraes (1813) nor Figueiredo (1913) bring anything new to this discussion. Moraes, who lived in Brazil, describes the fruit and acknowledges a number of local subspecies: sweet oranges or from China; bitter tangerines, or seedless, and some other Brazilian varieties. The reference to Chinese oranges had occurred in a previous dictionary<sup>6</sup>, and it will reappear in Figueiredo, who even mentions *china* as a hyponym of *laranja*. This name for sweet oranges seemingly lasted in Portugal until the beginning of the 20<sup>th</sup> century, and in fact neither *Infopedia* nor *Priberam* make any mention of it.

Laranja

Fruto da laranjeira.

[...] Variedade de pêra portuguesa.

(Do ár. naranj)

hiperónimo de: china

LARANJA, s. f. Fruta d'arvore de espinho com casca de cor amarella, e gomos dentro: há laranjas doces, ou da China; azedas; Tangerinas, com embigo em baixo; selectas, ou sem caroço, mui doces: a Tangerina doce no Rio de Janeiro é diversa da Tangerina d'outras Colonias, e de sabor mui delicado. §. Myra — : peso das pendulas dos relógios de parede. *Mechan. de Marie.*

Figure 1: Moraes (1813) and Figueiredo (1913)

Figueiredo further mentions that *laranja* also refers to a variety of pear, which is probably a mistaken understanding of a variety of oranges that have the shape of a pear, which is commonly found in Brazil. *Infopedia* repeats the same information and it also presents the fruit as an hesperidium, which is an outdated botanical classification.

In sum, so far we have a hint about the remote Sanskrit origin and the vehicular Arabic etymon, and a fluctuating list of Latin equivalents that have no existence in classical Latin. However, the equivalence between *laranja* and *cidra* initially established by Pereira opens a new line of research.

### 3 *Cidra, Cidrão and Cidreira* in Heritage Dictionaries

Quite unexpectedly, we find out that Cardoso (1569) assigns to *cidra* the Latin equivalent that Pereira (1697) also used for *laranja* (i.e. *malum citreum*). The same applies to Barbosa (1611), who also

<sup>6</sup> Folqman (1755) “LARANJA, [...] Laranja da china, Malum aureum sinense. § Laranja azeda, Malum aureum acidum”

brings the name *malum Hesperium*, again found in Pereira (1697) for *laranja*, though he only provides the equivalence *malum citreum* for *cidra*. Finally, Bluteau (1712-28) characterizes both *cidra* and *cidrão* as fruits from a tree called *cidreira*, and, as to Latin equivalents, he registers *malum citreum* or *malum medicum*. The contamination between *laranja* and *cidra* becomes evident – semantically, they are quite equivalent.

*Cidra* may eventually be the word of Latin origin that could not be found for *laranja*. Indeed it has a Latin origin (cf. Cunha, s.v. *cidra*), but what can be found in Latin is not what might be expected. Lewis and Short (1879) claim that the Latin word *citrus* is probably “a mutilation of κέδρος, *cedrus*”, and they assign two meanings to it, which suggests that the naming issue is ancestral:

- I. *The citrus*, an African tree (hence Atlantis silva, Luc. 10, 144, and Massyla robora, Stat. S. 3, 3, 94), whose very fragrant wood (v. citrum) was used in making household furniture, and was prized very highly
- II. *The citron-tree* (also called *malus Medica*, *Persica*, etc.), *Citrus Medica*, Linn., whose fruit and leaves were laid between the folds of clothing to preserve it from worms

In all probability, the trees that the Romans knew as *citrus* were *cedars*, coniferous trees in the *pinaceae* family<sup>7</sup>, and they may have later adopted the same name *citrus* to refer another tree that became a member of the *rutaceae* family. Smith (1859) endorses Fée’s claim<sup>8</sup> that “for a long period [...], the *citron* was without any specific name among both the Greeks and Romans”.

\*CITRUS (κιτρία or κιτρία), the Citron-tree. For a long period, as Fée remarks,<sup>4</sup> the Citron was without any specific name among both the Greeks and Romans. Theophrastus merely calls it *μηλέα Μηδική ἢ Περσική*. Pliny<sup>5</sup> styles it the Median or Assyrian Apple-tree, “*Malus Medica sive Assyriaca*.” At a later period, *μηλέα Περσική* became a name appropriated to the Peach-tree, while “*malus Assyriaca*” ceased to be used at all: the designation of the Citron-tree then became more precise, under the appellation of *malus Medica* or *Citrus* (*μηλέα Μηδική, κιτρία*). Of all the species of “*Citrus*,” that which botanists term, *par excellence*, the Citron-tree of Media, was probably the first known in the West. Virgil<sup>6</sup> gives a beautiful description of it, styling the fruit “*felix malum*.” This epithet *felix* is meant to indicate the “happy” employment of the fruit as a means of cure in cases of poisoning, as well as on other occasions; while the *tristes succi* indicate, according to Fée, the bitter savour of the rind, for it is of the rind that the poet here points out, as he thinks, the medical use: he makes no allusion to the refreshing effects of the citron, but only to its tonic action; and this latter could not refer to the juice, the properties of which were not as yet well known. Some commentators think that, when Josephus speaks of the apple of Persia, which in his time served as “*hadar*,” he means the citron. This, however, cannot be correct. It would seem that he merely refers to a remarkable and choice kind of fruit, which was to be an offering to the Lord; so that *hadar* cannot be the Hebrew for

the citron-tree or its produce.<sup>7</sup> Neither is there any ground for the belief that the Jews in the time of Moses were acquainted with this tree.<sup>8</sup>—Virgil<sup>9</sup> says that the fruit of the citron was a specific against poison, and also that the Medes chewed it as a corrective of fetid breaths, and as a remedy for the asthma. Athenæus<sup>10</sup> relates a remarkable story of the use of citrons against poison, which he had from a friend of his who was governor of Egypt. This governor had condemned two malefactors to death by the bite of serpents. As they were being led to execution, a person, taking compassion on them, gave them a citron to eat. The consequence of this was, that though they were exposed to the bite of the most venomous serpents, they received no injury. The governor, being surprised at this extraordinary result, inquired of the soldier who guarded them what they had eaten or drunk that day, and being informed that they had only eaten a citron, he ordered that the next day one of them should eat citron and the other not. He who had not tasted the citron died presently after he was bitten; the other remained unhurt!—Palladius<sup>11</sup> seems to have been the first who cultivated the citron with any success in Italy. He has a whole chapter on the subject of this tree. It seems, by his account that the fruit was acrid, which confirms what Theophrastus and Pliny have said of it, that it was not esculent. It may have been meliorated by culture since his time.<sup>12</sup>

Figure 2: Citrus (Fée 1835)

Both these authors quote Theophrastus<sup>9</sup>, who described the species under the name of *Median or Persian Apple* and lists its properties (a fragrant tree, with thorns and inedible fruits, beard at all seasons, leaves that may be used to keep moths away from linen, a breath freshener, and a remedy for deadly

7 Cf. Smith (1984) “[...] the tree called citrus (a species of cedar or juniper), the wood of which was highly esteemed by the Romans for furniture.”

8 A. L. A. Fée is a French botanist who published *Flore de Virgile, ou Catalogue raisonné des plantes citées dans ses ouvrages*, in 1835.

9 The Greek botanist Theophrastus (3<sup>rd</sup> century BC) is the author of *Historia plantarum*.

poison ingestion). Theophrastus claimed that the tree grows in Media and Persia, but he describes how to propagate it, which seems to indicate its introduction in the Italic peninsula. A few centuries later, Pliny the Elder<sup>10</sup> is probably the first to acknowledge the term *citrus* that he links to the *nata Assyria malus*. He treats it as a ‘really exotic tree’, as he also instructs on how to plant it<sup>11</sup>.

Thus, the much later association between *cidra* and *laranja*, drawn both by Pereira (1697) and Bluteau (1712-28) needs to be interpreted under the light of the terminological confusion induced by the introduction of new species, and we may conclude that the understanding of word meanings may require well more than lexicographic research – hopefully these other sources will be part of historical corpora<sup>12</sup>.

## 4 Laranjas and Cidras in Historical Corpora

Data from *Corpus do Português* show that the first attestations for *laranja* occur in the 15<sup>th</sup> century, and they are quite rare until the 19<sup>th</sup> century and especially the 20<sup>th</sup>. Three of the most remote attestations refer the existence of *laranjas* in distant locations. The first one reports an offer of *laranjas* and *cidrões* by the king of Mombaça to Vasco da Gama, during his inaugural trip to India, by sea, in 1498.

Ao domjnguo de rramos mandou o Rey de Mõbaça ao capitam moor hũ carneiro e mujtas **laranjas** e cidrões e canas daçquar (*Diário da viagem de Vasco da Gama*, 1498)

The second attestation reports the existence in San Tomé of ‘orange trees’ (brought from Portugal) that produce ‘oranges’ apparently bigger than expected:

**Larangeyras** ha muytas e as trouuerom de Portugal e dam fruto **laranja** tam grande como grande çydra de Portugal // Em esta ylha de Sam Thome ha muytas cidras e tammanhas como a barriga de pote de meo almude Limões ha muytos e tam grandes como cidrões em Portugal Limas muytas tam grandes como as cidras de Portugal (*Códice Valentim Fernandes*, 1506-1510)

The third attestation concerns China and the discovery of local sweet oranges:

Ha muitas e muito boas **laranjas**, ha tres generos de **larãjas** doces a quaes milhores, hũas que tem ha casca muito delgada, que quasi sabem a uvas, outras que tem ha casca grossa e crespa tamalaves bicaes mui sabrosas, que lhe comem casca e tudo: outras maiores que as demais que tem ha casca em meo, nem muito grossa nẽ muito delgada: estas sam somenos por serem muito docicadas. (*Enformação das cousas da China*, 1520)

All these attestations value *laranjas* positively – they are a gift from a king, and they are praised for their size and for their taste. Yet the very first attestation brings *laranjas* as ammunition, similar to stones and certainly inedible:

Hũus lançauom pedras, outros **laranjas**, e outros cospiom contra ele (*Crónica de D. Fernando*, 1431-43)

The first attestation of *cidra* also dates from the beginning of the 15<sup>th</sup> century and occurs in a Portuguese culinary treaty. It is the recipe of a sweet preserve, similar to marmalade:

10 Pliny the Elder (AD 23 – 79) is a Roman naturalist, author of the *Naturalis Historia*.

11 Cf. [www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.02.0137%3Abook%3D12%3Achapter%3D7](http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.02.0137%3Abook%3D12%3Achapter%3D7).

12 It should be noted that, in contemporary Portuguese, there are two homophonous words [‘sidrɐ’]: according to *Infopedia*, *cidra* (from the Latin *citrea*) is the name for a fruit that is green and bigger than a lemon, and *sidra* (from the Hebraic *schēchar*) is the name of a beverage obtained from fermented apple juice. French and English equivalents for the Portuguese *sidra* are written with an initial <c> (cf. Fr. *cidre*, Eng. *cider*), though <s> would be more expectable, given the etymon *sicēra*. The graphic contamination can also be documented in Portuguese: *cidra* occurs quite frequently meaning ‘apple cider’.



Pera fazer diacidrão Escolherão muyto boas cidras & bem feytas que nã Sejam quejmadadas da jeada nẽ Verdoemgas. (*Tratado de cozinha portuguesa*, 1400?)

Lorenzo (1968: 220) documents *cidra* in the same text where he found *laranja*, some 30 years sooner than the 15<sup>th</sup> century, and since this is an enumeration, *cidras* and *laranjas* must be different fruits.

Romãas e laranjas e limões e çidras

In sum, Lorenzo (1968) locates both words in the final quarter of the 14<sup>th</sup> century, in the same document and in the same sentence, referring to different entities, and *CdP* allows us to identify two meanings for *laranja* and another closely related meaning for *cidra*, but none of them help us to understand why ‘oranges’ are called *laranjas* in Portuguese.

## 5 Cross-linguistic Evidence

The very essence of the European Roots project is to facilitate cross-linguistic inquiries through similar monolingual descriptions. The Portuguese word *laranja* is formally close to the Galician *laranxa* and the Basque *laranja*, and they all contrast with the Castilian proximity to the Arabic (Sanskrit) source (cf. Cast. *naranja*). The Italian output *arancia*, like the French *orange*, is also diverse – both these languages lack the initial consonant. These phonetic distinctions are almost microscopic, but they may help to understand how such a variety originated and how they all relate to each other.

The dictionary of the Real Academia Española claims that *naranja* comes from the Hispanic Arabic word *naranġa* and the *Corpus del Español* provides evidence for the word in a document from 1418. Since the Nasrid Kingdom of Granada survived until 1492, direct language contact between Castilian and Arabic is quite plausible until the end of the 15<sup>th</sup> century. Therefore, the Arabic origin hypothesis for the Castilian word looks like a sound hypothesis, but it sheds no light as to its equivalents in other Romance languages.

The *Trésor de La Langue Française* claims that the French word *orange* was preceded by the compound *pome (d') oreng*, a calque from the Italian *melarancio*<sup>13</sup>, which may explain the absence of the initial consonant in *orange*, if we succeed in understanding why it is absent from Italian (cf. *mel(a)+arancio*). Pianigiani (1907) offers the useful suggestion that a rebracketing of the source word (i.e. the Arabic *narangi*) took place when it entered a southern Italian dialect: the initial consonant [n] was interpreted as an indefinite article. Since Italian and Arabic were in contact, at least in Sicily, this hypothesis also appears to withstand skepticism. Only Portuguese, Galician and Basque words lack a logical hypothesis:

- (4) Sanskrit *nāraṅga*
  - > Arabic *nāraṅġ*
    - > Hispanic Arabic *naranġa*
      - > Castilian *naranja*
    - > Mediterranean Arabic *narangi*
      - > Southern Italian dialect (u)n *arancia*
        - > French *orange*
  - ? Portuguese *laranja*
  - ? Galician *laranxa*

13 Many French words of Arabic origin arrived through Italy (cf. It. *zucchero* > Fr. *sucre*). The Portuguese word *açúcar* and the Castilian *azúcar* have the same Arabic origin but, as usual in both these languages, they incorporate the Arabic determiner (i.e. *as-sukkar*).



## 6 Oranges from Portugal

At this stage, a further cross-linguistic survey may be worth considering. According to Triantafyllidis' dictionary, the Modern Greek equivalent for *orange* is ποτοκαλί [portokalí], a borrowing from an Italian word, *portogallo*. The Modern Greek word, meaning 'sweet orange'<sup>14</sup>, is first attested in 1669. Similar names can be found in many other languages, mainly located in Northern Africa and in the southern and western vicinity of the Black Sea, as well as in some Italian dialects:

(5) Arabic	إل البرتوكالي (albertuqaliu) <sup>15</sup>
Georgian	ფორთოხალი (p'ort'okhali)
Greek	ποτοκαλί (portokalí)
Macedonian	портокалова (portokalova)
Romanian	portocaliu

So, quite paradoxically, Portuguese has a peculiar word for orange, and the name for orange in a number of languages evokes the name of Portugal. On Reddit, we can find at least two discussions about orange names. The first one<sup>16</sup>, entitled "The word for the fruit orange in various European languages" presents a colorful map (cf. Figure 3) with the distribution of orange names according to their linguistic origin. It certainly contains inaccurate details (the discussion reveals several discrepancies with speakers from many of these languages), but it helps to visualize the geographic distribution of these etymological word families. In Central and Western Europe, the Sanskrit/Arabic word dominates, either in its simple form (cf. Castilian *naranja*, Italian *arancia*, Portuguese *laranja*, French *orange*) or in a compound preceded by a descendant of the Latin word *pomus* (cf. Polish *pomarańczowy*). Then Germany, Northern Europe, Eastern Europe and Asia typically adopt a compound formed by *Apfel* (the German translation of *pomus*) and a modifier that related to China (cf. German *Apfelsine*). Finally, North Africa, Greece and some neighboring languages choose words related to Portugal (cf. 5). The second map<sup>17</sup> draws a hypothesis for the spreading of the words for orange (cf. Figure 4).



Figure 3: The word for the fruit orange in various European languages.



Figure 4: Spread of the word for the fruit orange.

None of these maps help to understand the specific output of the Portuguese word (i.e. *laranja*), but they do help to set a path for the dissemination of the word related to Portugal. According to the second map, the loan started in Venetian and travelled south in Italy, reached Greek and from there

<sup>14</sup> Simeon Tsoladikis, whom again I wish to thank, conveyed this information.

<sup>15</sup> Just for sweet oranges. Sour oranges are called *nāraṅg*.

<sup>16</sup> Accessed at [www.reddit.com/r/linguistics/comments/1cmhsv/the\\_word\\_for\\_the\\_fruit\\_orange\\_in\\_various\\_european/](http://www.reddit.com/r/linguistics/comments/1cmhsv/the_word_for_the_fruit_orange_in_various_european/) [28/03/2018].

<sup>17</sup> Accessed at [www.reddit.com/r/etymology/comments/5mrfao/spread\\_of\\_the\\_word\\_for\\_orange\\_the\\_fruit\\_oc\\_963\\_733/](http://www.reddit.com/r/etymology/comments/5mrfao/spread_of_the_word_for_orange_the_fruit_oc_963_733/) [28/03/2018].

it spread to Romanian and to Arabic. Apart from the fact that *Portugal* is not the name for *orange* in Portuguese, as this map suggests, it is also questionable to set a dissemination route based on no evidence. Is it plausible that a Venetian name for orange (i.e. *portogallo*) moved into Greek and from Greek to Arabic, Macedonian, Romanian, Turkish, Georgian and maybe some more languages or dialects in close-by regions? Though no references are provided, Wikipedia's Italian entry on *citrus sinensis*<sup>18</sup> gathers some dialectal information that is displayed in the following diagram:



Figure 5: Orange in some Italian dialects.

Though I am not a specialist on Italian dialectology and cannot ascertain the reliability of these data, it is interesting to note that words related to Portugal occur in a larger number of dialectal spots, but not near Venice (as hypothesized in Figure 5) nor in Sicily. Though the history is still tentative, Dugo and Di Giacomo (2002: 8-9) present interesting information. They remember that the Arabs occupied the Mediterranean territories left free by the Romans and have introduced into these new plants and agricultural techniques. Sour oranges were used to adorn mosques and gardens, for their scent. Then, the same authors refer that “the first mention of the common or sweet orange was found in a historical book by Hugo Falcando, who lived in Sicily from 1154 to 1169” and he also mentions that “the term ‘arangias’ was used by Blondus Flavius in a description (thirteenth century) of the citrus in Amalfi and Naples”. Dugo and Giacomo (2002:9) also state that “Nicolò Speciale (15<sup>th</sup> century), who wrote a book on the siege of Palermo, reported that sour orange was grown in Sicily and its fruits were called ‘arangias’ by the Sicilians”. It is quite tempting to establish a link between the Sicilian word *arangia* and the Portuguese *laranja*, particularly if we admit the hypothesis of a methanalysis process intervening over the sequence *l’arangia*. After all, other cases of methanalysis are documented for fruit names in Portuguese<sup>19</sup> (cf. Lat. (*pruna*) *damascea* > *d’amascea* > Pt. *ameixa* ‘prune’, by the

18 Cf. [it.wikipedia.org/wiki/Citrus\\_sinensis](https://it.wikipedia.org/wiki/Citrus_sinensis).

19 Cf. A.Nascentesarchive.org/stream/DICIONARIOETIMOLOGICORESUSUMIDODALINGUAPORTUGUESAANTENORNASCENTES/DICION%C3%81RIO%20ETIMOL%C3%93GICO%20RESUMIDO%20DA%20LINGUA%20PORTUGUESA%20%20ANTENOR%20NASCENTES\_djvu.txt.

deglutination of what appeared to be a preposition; Lat. *prunum* ‘plum’ > Pt. *abrunho*, by the agglutination of a putative definite article). So, though we do not have clear evidence of the Sicilian origin of the Portuguese word, we do have an indirect hint about this immediate etymological link, probably made possible by means of commercial relationships between Portuguese and Sicilian merchants before the 14<sup>th</sup> century:

- (6) Sanskrit *nāraṅga*
  - > Arabic *nāraṅġ*
  - > Mediterranean Arabic *narangi*
  - > Sicilian *(u)n arangia*
  - > Portuguese *laranja*

Finally, we need to take into account the contribution of a Portuguese agronomic researcher regarding the introduction of citrus plants in Portugal. Ferrão (1992: 167) claims that citrus coming from the Orient arrived in Europe via the Mediterranean Sea, brought by the Arabs. Consequently, citrus plants reached the Iberian Peninsula well before Portugal existed as a country, and penetrated the territory into the north as far as the ecology allowed them. The same author also claims that the oranges of such origin were not necessarily bitter – oranges that grow in milder climates tend to be sweeter than those that manage to survive in less favorable conditions. Algarve was, and still is, a very propitious environment for the cultivation of sweet oranges. On the other hand, still according to Ferrão (1992: 167), the Arabs spread citrus species down the African Western Coast and this is why the Portuguese navigators found oranges in Gambia, and, as we have seen above, Vasco da Gama also found oranges along the East African Coast in 1498. So apparently the introduction of sweet oranges in Europe by the Portuguese navigators who had brought them from China lacks historical support.

Ferrão (1992) considers that the oranges the Portuguese navigators found in India and China were sweeter than those they knew in Portugal, because of the ecology of these different locations. But they nevertheless brought them to Portugal, grew them, and started large-scale export to Europe, advertising them as a new cultivar. Italy was probably a good market, since Portugal used to buy and sell several goods to Italian cities, such as sugar to Genoa and Florence.

## 7 Conclusion

Words that we know and use have a shared existence – shared by all those that also know and use them or have done so in the past. However, shared existence does not always entail a shared knowledge. Words that are the legacy of each generation to the following may undergo a change in the process, and that change often occurs without explicit notice.

Dictionaries may help to trace semantic changes, because they document words, but dictionary-making processes are not primarily concerned with that aim. Therefore, dictionaries may give us some information on past and present word meanings, but this is generally not enough to fully understand them, and it may even be inaccurate. The body of all sorts of texts that human ancestry has produced until now, if preserved and incorporated in resourceful databases, is crucial to complement the information that heritage dictionaries offer us, and to validate the latter’s contents, or not.

However, even if textual databases, here including lexicographic databases, were as comprehensive as we would like them to be, it would still be necessary to manage the output results of any given search, and the amount of information thus obtained may be, as we have seen in the case of *laranja*, quite difficult to manage. The ancestry of oranges and orange trees in Europe is rather challenging to trace, because the words that were and are used to refer to them conceal more than what they show. In

this paper, I have collected the information that is available in heritage and contemporary dictionaries and in historical corpora as well as other relevant sources, which amounts to a large set of data that still fails to provide convincing explanations for all our questions.

This case study helps to demonstrate that we need to develop a sophisticated descriptive protocol that will allow us to precisely identify the relevant data and to be able to relate what we find productively. The European Roots prototype aims to respond to those needs, as it considers essential evidence and discards negligible information, but it is particularly useful in that it allows cross-linguistic analysis, since all languages may use the same template.

## References

- Barbosa, A. (1611) *Dictionarium Lusitanico Latinum*. Bracharae : typis, & expensis Fructuosi Laurentij de Basto. Accessed at: [clp.dlc.ua.pt](http://clp.dlc.ua.pt) [28/03/2018].
- Cardoso, J. (1569-1570) *Dictionarium Latinolusitani cum & Vice Versa Lusitanico Latinum*. Conimbricæ: Joan Barrerius. Accessed at: [clp.dlc.ua.pt](http://clp.dlc.ua.pt) [28/03/2018].
- Bluteau, R. (1712-1728) *Vocabulario Portuguez e Latino*. Coimbra: Collegio das Artes da Companhia de Jesu. Accessed at: [clp.dlc.ua.pt/DICWeb/default.asp?url=Home](http://clp.dlc.ua.pt/DICWeb/default.asp?url=Home) [28/03/2018].
- CdP = *Corpus do Português*: 45 million words, 1300s-1900s. Accessed at: [www.corpusdoportugues.org/x.asp](http://www.corpusdoportugues.org/x.asp) [28/03/2018].
- CdE = *Corpus del Español*. Accessed at: [www.corpusdelespanol.org/x.asp](http://www.corpusdelespanol.org/x.asp) [28/03/2018].
- CLP = *Corpus Lexicográfico do Português*. Accessed at: [clp.dlc.ua.pt](http://clp.dlc.ua.pt) [28/03/2018].
- Corominas, J., Pascual, J. A. (1981). *Diccionario Crítico Etimológico Castellano e Histórico*. Madrid: Gredos.
- CRPC = *Corpus de Referência do Português Contemporâneo*. Accessed at: [www.clul.ul.pt/pt/recursos/183-crpc#cqp](http://www.clul.ul.pt/pt/recursos/183-crpc#cqp) [28/03/2018].
- Cunha, A. G. (1994) *Dicionário Etimológico Nova Fronteira da Língua Portuguesa*. Rio de Janeiro: Nova Fronteira. 2ª edição.
- Fée, A. L. A. (1835) *Flore de Virgile, ou Catalogue raisonné des plantes citées dans ses ouvrages*. Bibliothèque classique latine ou Collection des auteurs classiques latins. Available at: [gallica.bnf.fr/ark:/12148/bpt6k298102/f108.item.r=citrus](http://gallica.bnf.fr/ark:/12148/bpt6k298102/f108.item.r=citrus) [28/02/2018].
- Figueiredo, C. de (1913) *Novo Dicionário da Língua Portuguesa*. Porto, Typ. da Empr. Litter. e Typographica. Accessed at: *Dicionário Aberto*
- Infopédia = *Dicionário da Língua Portuguesa da Porto Editora*. Accessed at: [www.infopedia.pt/linguaportuguesa/](http://www.infopedia.pt/linguaportuguesa/) [28/03/2018].
- Lewis, C. T. & C. Short (1879) *A Latin Dictionary*. Founded on Andrews' edition of Freund's Latin dictionary. Revised, enlarged, and in great part rewritten edition. Oxford: Clarendon Press. Accessed at: [www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0059](http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0059) [28/03/2018].
- Lorenzo, R. (1968) *Sobre Cronologia do Vocabulário Galego-Português (Anotações ao 'Dicionário Etimológico' de José Pedro Machado)*. Vigo: Editorial Galaxia.
- Mabberley, D. J. (2004) Citrus (Rutaceae): A Review of Recent Advances in Etymology, Systematics and Medical Applications. *Blumea* 49: 481–498. Accessed at: [repository.naturalis.nl/document/565086](http://repository.naturalis.nl/document/565086) [28/03/2018].
- Moraes = Moraes Silva, A. (1789) *Diccionario da lingua portugueza*. Lisboa, na Of. de Simão Thaddeo Ferreira. 2<sup>nd</sup> edition (1813). Accessed at: [www.brasiliana.usp.br/pt-br/dicionario/](http://www.brasiliana.usp.br/pt-br/dicionario/). [28/03/2018].
- Nebrija, E. A. (1492) *Lexicon hoc est Dictionarium ex sermone latino in hispaniensem*. Salamanca.
- Pereira, B. (1697) *Prosodia in Vocabularium Bilingue, Latinum et Lusitanum*. Eborac: Typographia Academiae Raskin. Accessed at: [clp.dlc.ua.pt](http://clp.dlc.ua.pt) [28/03/2018].
- Pianigiani, O. (1907) *Vocabolario Etimologico della Lingua Italiana*. Florence. Accessed at: [www.etimo.it/?pag=hom](http://www.etimo.it/?pag=hom) [28/03/2018].
- Priberam = (2008-2013) *Dicionário Priberam da Língua Portuguesa*. Accessed at: [www.priberam.pt/dlpo/chave](http://www.priberam.pt/dlpo/chave) [28/03/2018].
- Real Academia Española, *Diccionario de la Lengua Española*. Accessed at: [dle.rae.es/?id=DgIqVCc](http://dle.rae.es/?id=DgIqVCc) [28/03/2018].

- Silvestre, J. P. & A. Villalva (2014) A Morphological Historical Root Dictionary for Portuguese. Proceedings of the XVI EURALEX International Congress: The User in Focus, 967-978. Accessed at [http://www.euralex.org/elx\\_proceedings/Euralex2014/euralex\\_2014\\_074\\_p\\_967.pdf](http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_074_p_967.pdf) [29/03/2018].
- Smith, W. (1854) *Dictionary of Greek and Roman Geography*, illustrated by numerous engravings on wood. London: Walton and Maberly.
- Smith, W. (1859) *A Dictionary of Greek and Roman Antiquities*. Little, Brown, and Co. Available at: [archive.org/details/adictionarygree05smitgoog](http://archive.org/details/adictionarygree05smitgoog) [28/2/2018].
- TLFI = *Trésor de la langue Française informatisé*. ATILF - CNRS & Université de Lorraine. Accessed at: [www.atilf.fr/tlfi](http://www.atilf.fr/tlfi) [28/03/2018].
- Triantafylidis, M. *Dictionary of Standard Modern Greek*. Accessed at [www.greek-language.gr/greekLang/modern\\_greek/tools/lexica/triantafyllides/search.html?lq=%CF%80%CE%BF%CF%81%CF%84%CE%B-F%CE%BA%CE%AC%CE%BB%CE%B9&dq=](http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/search.html?lq=%CF%80%CE%BF%CF%81%CF%84%CE%B-F%CE%BA%CE%AC%CE%BB%CE%B9&dq=) [28/03/2018].
- Villalva, A. & J. P. Silvestre (2015) Filling gaps in dictionary typologies: ROOTS – a morphological historical root dictionary. Silvestre, J. P. & A. Villalva, eds. (2015) *Planning non-existent dictionaries*. Centro de Linguística da Universidade de Lisboa & Universidade de Aveiro.





# **Sign Language Lexicography**



# Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How

**Gabriele Langer, Anke Müller, Sabrina Wähl, Julian Bleicken**

*University of Hamburg*

*E-mail: gabriele.langer@uni-hamburg.de, anke.mueller@uni-hamburg.de, sabrina.waehl@uni-hamburg.de, julian.bleicken@uni-hamburg.de*

## Abstract

Within the DGS-Korpus Project, a corpus-based dictionary of German Sign Language (DGS) is compiled. Dictionary entries describe signs, their meanings and uses as they are reflected in the corpus. The dictionary entries include authentic examples taken directly from the original corpus recordings. Without a functional writing system for sign languages (SL), corpus building as well as SL usage examples in dictionaries have to resort to videos as representations of SL use. Examples can either be taken from the original corpus material (authentic examples) without the option of even mild editorial changes, or they can be re-recorded with a different signing model. While the latter allows for editing as well as constructed examples, it also entails very drastic changes to the appearance of the original authentic examples they are based on.

In the article our reasons for inclusion of authentic examples are discussed and criteria for example selection listed. To compensate for the challenges that authentic examples removed from their original contexts entail, translations and context information are added to the entries. The practical steps for example preparation, namely selecting the segment, providing context information and adjusting translations are described.

**Keywords:** authentic examples, sign language dictionary, spoken language (as opposed to written), corpus-based lexicography

## 1 Situation of the DGS Community

Sign languages are visual and spatial languages. They make use of the movement of both hands, the head and the upper body as well as facial expressions, movement of lips, tongue and cheeks, eye gaze and eye blinks. Spatial modifications of the sign form are used for semantic and syntactic purposes. Sign language is predominantly used in face-to-face interaction, that is in dialogic situations of a shared environment and the spatio-temporal co-presence of the interlocutors. Signed languages do not have a widely used functional writing system. This is mainly due to the difficulties of reducing a multi-modal language relying on situative embedding to a few abstract symbols, leaving iconic descriptions, facial gestures and indexicality open to the imagination of the addressee.<sup>1</sup>

Systematic situative embeddedness and the absence of a functional writing system make signed language in many respects comparable to spoken language (as opposed to written language)<sup>2</sup> that is used in conditions of everyday conversations. The similarities concern syntactic and semantic processes

1 There are quite successful initiatives to develop and use an iconic writing system known as Sutton SignWriting (cf. Sutton's SignWriting Site) with a growing international community of signers using it (for Germany see Wöhrmann 2003). This writing system might be an adequate system, but it is not yet commonly known. Therefore, it is not suitable to be used for a description of sign senses in the DGS dictionary.

2 In the field of sign language linguistics, the term spoken language is often used in contrast to sign language, irrespective of its medial appearance, which is of importance in this paper. For that reason we refer to this language type as vocal language. Vocal languages typically occur in two medial forms, spoken or written, which both have different properties.

including vagueness and fewer explicit references, and all aspects of on-line production of utterances, such as disfluencies and dynamic changes of plan (Ebbinghaus 2001, Schwitalla 1997).

Sign languages are used in communities of deaf, hard-of-hearing and associated hearing persons, only a minority being children of deaf parents. The majority of congenitally deaf people have hearing parents and are exposed to sign language only at preschool and school. Sign language users constitute a linguistic and cultural minority within a larger community of a vocal language with its spoken and written forms of communication. This has some impact on the social conditions of sign language use in a structural language contact situation. Signers are functionally bilingual “as a result of growing up in a hearing world and receiving education in (at least) spoken language” (Bank et al. 2016: 1284; cf. Ebbinghaus 2001 for DGS), using the surrounding vocal language for written communication. Also, in DGS and other SL spoken language elements, so-called mouthings are integrated into the signing. Mouthings are visible gestalts of German words, often accompanying content signs for denoting objects, events or concepts (cf. Ebbinghaus & Hessmann 2001). Many signs can be combined with several mouthings and mouth gestures, the denotation thus being specified in a dynamic way. This is facilitated by the underlying iconicity of signs, resulting in a lexicon with many highly polysemous signs.

Sign language can be regarded as a structurally spoken language with respect to the means of communication in the spatio-temporal situatedness of co-presence (cf. Fehrmann & Linz 2009). The structure of a sign language is largely determined by these medial contexts of its use, as is the case with spoken or written language. If texts are defined as the result of a temporal dislocation of utterance production and reception (Ehlich 1983), the textual situation leads to a de-contextualisation of language. The beginning of such a process of possible changes in the way language is used can also be observed within sign languages in planned recordings, addressed to a general audience and with editing techniques available (cf. Krentz 2007, Linz & Fehrmann 2009). Video enables the production of sign recordings as texts, and also the conversational use across places with camera devices in portable phones or the internet (cf. Keating et al. 2008 on medial factors influencing linguistic structure). Though we can talk of signed texts in the above definition, these are not written; and they are bound to bodily appearance.

## 2 Lexicography of Sign Languages

The lexicography of sign languages has been shaped by three major circumstances: First, SL are minority languages surrounded by and embedded in societies with at least one dominant vocal language that is also used for written communication. Second, due to their unique visual-spatial nature there is no functional writing system available to write signing. This directly results in a lack of written sources available to study sign language use, and also has major implications on how to represent sign language in a dictionary as well as to what extent sign language is chosen as the metalanguage in SL dictionaries – or not. Third, being a young field of study, general SL research has still not arrived at a general agreement on the basic categories, structures and properties of the languages.

### 2.1 Sign Dictionaries

Serious sign dictionaries<sup>3</sup> contain sign entries that aim at a description of the sign’s meanings and uses in their own right from a basically monolingual perspective. However, without the option to write sign language texts no sign dictionary – as far as we know – has tried to completely do

3 We do not discuss here sign dictionaries of the type known as sign collections. These are in essence simple bilingual but mostly uni-directional word-to-sign lists aimed at hearing users, and do not contain comprehensive sign entries which describe signs and their uses in detail. They also do not include signed example sentences.



without a vocal language as its metalanguage. At least front and back matter, headings and indexes such as subject fields for a thematic access are given in the surrounding vocal language. Almost always the written vocal language is also used as metalanguage for the description of meaning (dictionary definition), and further information on the signs' usage and grammar.<sup>4</sup> Also, all larger SL dictionaries include translational equivalents of the lemma sign into the vocal language, even if the product is not primarily designed to be a bilingual dictionary. Translational equivalents are either given as an indication of meaning or in addition to the dictionary definition. The rationale here is that not only learners but also native signers – who are functional bilinguals – prefer to have an access via the surrounding vocal language.

Sign dictionaries also tend to provide a way to find signs by searching for their form. As far as we know there are no dictionaries of a signed and vocal language pair that are truly bilingual in the sense that both languages receive the same amount of attention and display the same complexity, thoroughness and depth of description for both languages. As such, existing sign language dictionaries are always some form hybrid dictionary between monolingual (with the focus of description on the signs their meanings and uses in the entry structure) and bilingual (with the inclusion of translational equivalents and bidirectional access). One example of such a hybrid dictionary is the ODT-SL, described by Kristoffersen and Troelsgard (2010: 1550) as follows: “As a result of this decision, the Danish Sign Language Dictionary could be described as monolingual dictionary, which instead of definitions has (searchable) equivalents in another language, Danish, which is also the general metalanguage of the dictionary.”

Because of the lack of a functional writing system it is very difficult to present complete sign utterances on paper (as can be seen looking at example 4). Therefore, printed sign language dictionaries usually do not include real SL example utterances. Sometimes they include information on typical semantic contexts of a sign's use – as a substitute for usage examples – in the written vocal language (cf. for example D-SAS). In electronic or online SL dictionaries example sentences are usually presented in the form of video clips, and almost always an additional translation into the vocal language is provided. Sometimes either a gloss transcription and or a gloss-like literal translation of each sign of the example utterance is added in order to visualize the sequential sign order and to support easy, clickable cross-referencing to the entries of the other signs in the example sentence (see for example ODT-SL for glosses and ONZSL for a gloss-like literal translation of each sign). Normally the usage examples in sign dictionaries are studio recordings, and as far as we know we are the first dictionary project to include example sentences directly taken from the filmed original corpus data.

## 2.2 Corpus-based Lexicography of Sign Language

Without the availability of written texts in sign language, research has to rely on recordings of signing as permanent representations of language use to be investigated and analyzed. Only recently has technological progress in recording and storing signed data, and in annotating and retrieving this data more efficiently by the way of annotation tools, made the collection and use of sign corpora of a considerable size feasible. However, the idea of corpus-based SL lexicography has been around for quite a while. The first, very impressive attempt to collect and use a corpus for SL lexicography was made in the D-ASL despite technical limitations of that time (cf. Stokoe 1993). Other general SL dictionary projects, such as the D-NZSL and the ODT-SL, included data collection sessions. The recordings were reviewed and selectively tagged – in the case of D-NZSL for signs expressing concepts of a pre-defined list (cf. Kennedy 1996), and in the case of the ODT-SL for sentences to be used as models for re-recorded examples (cf. Kristoffersen 2010). Both projects combined the analysis of

4 One exception is the D-LSFB. This online dictionary includes videos with signed texts for etymology explanation and dictionary definitions, and it also includes recorded competence examples.

some corpus data with the intuition of consultant groups and editors for the description of lemma signs in dictionary entries.

The KS-PJM is based on a large corpus of Polish Sign Language (PJM) collected between 2011 and 2016. The corpus was used for lemma selection and corpus data was used in the preparation of the entries. Gaps of signs missing the corpus were filled and additional unattested meanings/uses of corpus signs were added by the editors. Examples were drawn from the corpus and re-recorded in the studio (cf. Linde-Usiekniewicz & Rutkowski 2016: 377-379).

Up to now all SL dictionaries that worked with corpus data also drew extensively on the intuitions of the editors and consultant groups to supplement the findings. With SL corpus data just becoming available in considerable sizes the field of corpus-based lexicography of SL is very new and remains in the process of forming. Methods, processes and presentational formats are largely still experimental or only just developing, and standards have not yet established in the field. As more and larger SL corpora become available, corpus data is increasingly used in dictionary projects for lemma selection, the discovery and description of sign meanings and uses, and as a source of usage examples.

### 3 DGS-Korpus Project

The DGS-Korpus Project is a long-term project (2009-2023) of the Academy of Sciences and Humanities in Hamburg, Germany. Its central aims are the collection of a corpus, making parts of it available as the Public DGS Corpus to the language community and researchers alike, and the development of a corpus-based dictionary for DGS-German.

#### 3.1 Data and Annotation

The filming sessions for the corpus data collection took place between 2010 and 2012 in 12 different German cities. The sample is balanced for region, age and gender. A mobile studio was used to film the 330 informants participating in this project. Informants from the same region were filmed in pairs with a moderator leading through the sessions. The data collection sessions were designed to cover a wide range of different topics. Tasks to initiate signing on different topics included open conversation, narrations of life experiences or events witnessed and retellings of stories (Nishio et al. 2010). The data collected reflects the properties of sign languages in typical face-to-face interactions, and is of structurally spoken nature. In the corpus, there are no discourse or text types implying registers of social distance, as would be seen in public talk, working instructions or legal texts. And there are no signed texts in the sense of spatio-temporal dislocation or signing through media for bridging spatial distance.

Data collection resulted in nearly 560 hours of signed material. The data is tokenized, lemmatized and annotated. For sign languages this has to be done manually. Lemmatization is achieved by matching tokens to types that are identified by unique glosses<sup>5</sup> and HamNoSys notations describing the signs' forms. Lemmatization and annotation is still ongoing. Already available for lexicographic analysis are 66 hours of the material with a completed continuous basic lemmatization<sup>6</sup> and annotation (May 2018). A very large part of the corpus has been translated. Translations are time-aligned with the signing, and thus can be searched for specific concepts even in parts that have not yet been lemmatized. In some cases, relevant signs, chunks or smaller passages have been selectively token-type-matched and annotated in videos that remain unlemmatized (so-called spot transcriptions). The current corpus size,

5 In the underlying database types are identified internally by an unchangeable ID-number. Glosses are bound to that ID. Any change concerning the gloss or the HamNoSys notation of the type will automatically affect all tokens that have been linked to that type.

6 Lemmatization here is token-type-matching and an important part of the basic annotation.

including all tokens, is approximately 480,000 tokens (May 2018). The DGS corpus is the basis for the work on the corpus-based dictionary. Compared to the size of written language corpora containing billions of tokens, the DGS corpus is fairly small. Nevertheless, the token count is sufficient to start with the description of the more frequent signs.

### 3.2 The DGS Dictionary

The DGS dictionary is one of the first corpus-based dictionaries of an SL. The DGS corpus has been collected to serve as the basis for lexicographic description of signs. The corpus and dictionary are to be interconnected. The dictionary is descriptive, with no intention of standardization.

One of the guiding principles at the present time is that dictionary entries contain only sign forms and sign meanings/uses that are attested in the corpus.<sup>7</sup> These are found in the corpus by summarizing views on the corpus data and looking at the original occurrences of a sign in context (cf. Langer et al. 2018). Also, with a few very particular exceptions, all examples shown in the dictionary are taken directly from the corpus. This approach can be described as corpus-based, but at the same time also as corpus-bound. We consider it rewarding to explore the possibilities and limits of this approach to base all essential information on signs on “objective evidence of language in use”, as Atkins and Rundell have put it (2008: 53). However, considering the relatively small size of our corpus we are aware of the limitations of this approach, and might decide to move beyond a strict corpus-bound method at a later stage. If we do, we feel it would be necessary to clearly mark information that has been added from other sources than the corpus.

As a consequence of the corpus-bound approach, we have corresponding corpus evidence for all variant forms and senses of a sign included in the entry, and thus can retrace the abstracted information to the original data. It is planned to link and cross-reference between the dictionary and the Public DGS Corpus, where the authentic usage examples can be viewed in a wider context and further occurrences of the lemma signs can be found.

Just like other sign language dictionaries, the DGS dictionary will also be a hybrid type with some properties of a monolingual and some properties of a bilingual dictionary.<sup>8</sup> The DGS dictionary will have to serve very different user groups, such as native signers, advanced learners and beginning learners. The different information types in the entries have different functions for these groups, and not all information types are addressed to all groups in the same way. It will be a monolingually oriented dictionary in that it primarily describes the language units of only one language, that is DGS. Full entries with a description of forms, meanings, uses and distributions will only be listed for DGS signs. The listing and description of the senses of a sign is in the first step described as independently as possible from the structure of the second language (German). However, written German is used as a metalanguage for any written information, such as the description of the senses. In practice, the dictionary can be used as a bilingual dictionary. For each sense listed there are one or several German equivalents shown serving the direction of use from the source language DGS to target language German. The dictionary does not have detailed information on the German equivalents. However, for bilingual native signers providing a simple equivalent will often be enough to fulfill their informational needs and, if not, it can serve as a starting point for the search in a monolingual German dictionary.

7 In other dictionary projects corpus data has been supplemented by the intuition of the editors and SL consultants for missing signs and senses (cf. for example Stokoe 1993: 132; Kennedy 1996: 37-38; McKee & McKee 2013: 517-518; Linde-Usiekiewicz & Rutkowski 2016: 377). For the user of these dictionaries it remains unclear which parts and information of the resulting dictionary entries are actually derived from corpus data evidence, and which parts are additions made on the basis of intuition.

8 Erlenkamp has mentioned that for an adequate description of a language a monolingual dictionary is the better choice, while for a minority language in a majority society it is useful and politically more supportive to produce a bilingual dictionary (1998: 102). A hybrid form combines both requirements in a useful compromise.

The German equivalents will be searchable. Access by German equivalents is helpful for learners of DGS, but also for bilingual native signers to search for sign entries via vocal language words (source language German to target language DGS). Also, a search function via the sign form will be provided, as well as a thematic access to the sign entries.

Entries will be continuously published until 2023, with revisions published as necessary. At this point in time, structure, labeling and layout of the entries is still experimental and allows users to view and discuss the information types and contents of the first entries.

One important decision on layout function is to not represent lemma signs by glosses.<sup>9</sup> While glosses are very convenient for internal workflow processes, they can also lead to a confusion of languages and misleading assumptions on the signs' properties as inferred from the gloss words' properties by non-expert users. All DGS elements in the dictionary entries are represented as video (or a play button indicating that a video is available), or as a micon (moving icon) in combination with an identification number. The micon represents the lemma, while the number provides a unique sortable and searchable label in cross-references, and supports the discrimination of signs with a similar form.

### 3.3 Structure of Preliminary Entries

Currently a preliminary entry page (see Figure 1) is divided into two main areas.<sup>10</sup> A fixed display window is placed on the left side. It is used to display any kind of signed information located at different places within the entry. On the right side all information on the lemma sign is displayed. At the moment the entry is organized in the form of a table. At the top a unique number identifies the entry and the lemma sign. The head of the entry contains information on the sign form, possible sign variants and further comments on the sign and its grammar. Play buttons in the row labelled "form" can be used to start the video showing the chosen form. The main body of the entries is a list describing the different senses. Here you find sense-related information on, for example, mouthings, definitions, German translational equivalents, signed usage examples and possible synonyms and antonyms and collocational patterns.

Authentic examples taken from the DGS corpus illustrate each sense. By using the play button the DGS utterances may be viewed in the display window. Next to the play button, context information (text in square brackets) and a German translation of the example is given (with the sign's translational equivalent in bold face).

All cross-reference signs (e.g. synonyms and antonyms) are represented by micons and identification numbers. When the mouse is placed over an icon it starts to move. A click onto the micon will show the sign in the display window. If the entry for the cross-reference already exists, the number background is red and a click on the number changes the screen to the corresponding entry. Information on sign combinations and related or similar signs is given at the bottom of each entry below the sense listing.

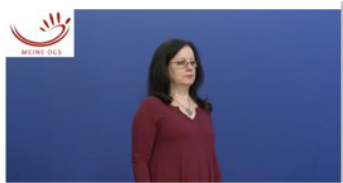
In the preliminary entries, information given on DGS in DGS is the lemma sign and its variants, all synonyms, antonyms and other cross-references, collocational patterns and the signed examples. Usually two examples per sense are given. Thus the user is given the possibility to draw conclusions

9 A gloss in the world of sign language research and teaching is a vocal language word used as a label for a sign. Such labels – usually written in capitalized letters – make it easy to speak about signs, represent signed utterances in a written form, to sort and search for sign entries electronically, and to order them alphabetically according to their gloss. Often numbers or other markers are added to the gloss word in order to distinguish different signs and/or to mark morphological features. Usually the gloss name is a word that corresponds in one of its meanings to a core meaning of the sign, and can therefore be mistaken for a translation. In annotation sign glosses are used as unique labels for signs in the token-type-matching.

10 The entry in Figure 1 does not include all of the following information types.



from the examples. All these information types support the differentiation of senses. Translational equivalents and the translations of the signed examples are bilingual elements of the dictionary, and are in the target language, German. German is used as metalanguage for the context on examples as well as the definitions.<sup>11</sup>













FORM	351.1 	
KOMMENTAR	Kann ein- oder zweihändig ausgeführt werden.	
GRAMMATIK	Richtungsgebärde	
LESART 1	...	
LESART 2	...	
LESART 3	MUNDBILD	zurück
	DEFINITION	etwas, das man bekommen, mitgenommen oder sich ausgeliehen hat, zurückgeben, zurückbringen oder zurückschicken
	DEUTSCHE ÜBERSETZUNGEN	zurückgeben; zurückbringen; zurückschicken; zurückzahlen (Geld)
	BEISPIELE	<p>1  [Thema: Karte zum Einkaufen in der Metro] Ich habe noch die Karte. Ich sollte sie eigentlich schon <b>zurückgeben</b>, aber ich habe sie noch behalten.</p> <p>2  [Thema: Umweltschutz und Pfandflaschen. Die Erzählerin kauft lieber Getränke in Glasflaschen als in Tetrapackverpackungen.] Ich nehme meistens Glasflaschen, die kann ich wieder <b>zurückbringen</b>.</p>
	SYNONYME	   05188 08133 07275
	SEM.-SYNT.	bei Ausrichtung: [Person, die zurückgibt] → [Person, die zurückbekommt]; [Gegenstand, der zurückgegeben wird] → [Ort, an den er zurückgebracht wird]
	SACHGRUPPEN	Transfer von Gegenständen; Geld, Geldwesen, Bankwesen
VERWANDT	  014044 027819	
FORMÄHNLICH	  014403 018091	

Figure 1: Structure of a preliminary entry

## 4 Authentic Examples

Atkins and Rundell (2008: 453-355) list the following most basic functions of examples in a dictionary entry: to attest the existence of a word or meaning, to elucidate the meaning by complementing the definition and clarifying the sense, and to illustrate contextual features such as syntactical patterns, collocation, and the like.

One of the basic decisions when planning a dictionary is whether to work with constructed or authentic examples. Prinsloo and Gouw (2000) discuss advantages of each of these and sketch a continuum between constructed editorial examples, heavily or slightly altered examples from a corpus, and unaltered authentic examples. Atkins and Rundell (2008: 457) note that in practice examples are usually abstractions from recurrent patterns found in a large number of corpus data, with adjustments to fulfill

<sup>11</sup> Though definitions could be given in sign as videos, this will not be done in the DGS dictionary. Signed definitions as a text type are not established yet, and developing and recording them would require a lot of resources. Second, as videos progress in time they vanish quickly and cannot easily be surveyed. The mnemonic function of writing is not fulfilled by filmed definitions. Moreover, video clips take up a lot of space. The same considerations hold for the description of example contexts.



the criteria of good examples. For written languages constructing or editing examples is therefore more a matter of degree than a clear-cut choice.

In the case of sign language examples, even a small editing adjustment results in a full re-recording, and thus a completely new appearance of the example and the loss of much of its original character. So for sign languages the question of authentic vs. edited or constructed examples is a fundamental decision. Here sign language and spoken language lexicography face similar challenges with respect to some problems and difficulties arising from their structural characteristics as face-to-face communication systems (Verdonik & Sepesy Maučec 2017). There are initiatives in spoken-language research to build corpus-based dictionaries or to add spoken-language specifics to existing dictionaries that use written texts as their base (cf. Verdonik & Sepesy Maučec 2017: 147-148). Auditory examples along with a written transcription and contextual information can be found occasionally in the ODT-VL on spoken language elements (Hansen & Hansen 2012: 930). Möhrs et al. (2017: 282) intend the integration of “multimodal information, such as corpus-based audio-examples and transcriptions for each entry” in their planned lexicographic resource of spoken German. However, these cases remain exceptional. Trap-Jensen (2004) also argues for taking authentic examples of spoken language, even if they are transformed into the written modality as is the case with the Danish dictionary. Though it may be “impossible to excerpt readable quotations in unaltered form from the spoken language corpus” (Trap-Jensen 2004: 316), because features intrinsically bound to the spoken medium are lost by that transformation, it may be worthwhile including authentic spoken examples. To convey the original quality as far as possible, all editorial changes to “authentic quotations with the source of the origin” like omissions, additions and reformulations shall be indicated as in a critical edition (Trap-Jensen 2004: 317).

Editing examples is useful for the reason of intelligibility, and distracting performance factors in particular can be eliminated with this. However, the DGS dictionary will include authentic examples taken from the corpus. As our data is filmed signed interactions, the options for presenting usage examples are, in principle, to use a segment of the original data recording, to re-record examples with a signing model, or to produce an animation to render the corpus example. Re-recording and animation<sup>12</sup> both allow for editing – with respect to content as well as to form aspects, while anonymization is also facilitated. However, we have good reasons to give priority to authentic examples as signed in the corpus, as outlined below:

- Re-recording an example – with or without editing – is very labor-intensive and the result is prone to lose the performative aspects of the original. Signed language production is not detachable from the body. Gestural and expressive aspects form integral parts not only of the utterance, but also of the language system (Fehrmann & Linz 2009). Performativity is an essential feature of SL and is, as such, not fully reproducible. Re-recordings are thus bound to lack important information.
- It is not entirely clear which aspects of signed utterances are to be preserved in reproduction; see for example the difficulty in separating affective and grammatical functions of facial expressions (cf. De Vos et al. 2009 on eyebrow movement).
- DGS exhibits a lot of regional variance. When re-producing examples from different regions a signing model would have to reproduce the signing in the examples very closely to ensure consistency with regard to regionality. The model would thus have to execute many signs that they might not use themselves, a task which is very difficult to do naturally, especially in a studio setting.
- The dictionary will be closely linked to the freely accessible Public DGS Corpus (cf. Jahn et al.

<sup>12</sup> Animations that can be generated automatically from signed input data may be a future device to compensate for the lack of a writing system, for granting anonymity and the possibility of editing, but are to date not yet available. We do not dig further into this issue, as all arguments listed here are independent of the techniques of how occurrence examples are reproduced.

2018), which is a valuable language and cultural resource of DGS. Using corpus examples in the dictionary is a way of appreciating this resource as well as appreciating the contribution of the informants to the project, and exploring further the possibilities of its use. The dictionary can thus contribute to the further recognition of DGS within the surrounding German society (cf. Erlenkamp 1998: 99). From the entries there will be a direct access to the source data. We regard it as an advantage that, in general, examples can be traced back to their source conversation with all background context to be viewed. Because the informants have given their informed consent, there is no general need to re-record all examples for reasons of anonymization.

- Including authentic examples signed by many members of the signing community may boost the interest in and the acceptance of the dictionary, as people can browse through the product and recognize some of their relatives, friends or other people they might know. Authentic examples of many signers are certainly livelier and show a larger range of signing styles than plain studio recordings of only a few signing models.
- In a poly-functional dictionary different user groups should be served. Authentic examples are directed more at native signers and advanced learners than novices.

We want to stress the central argument for including authentic examples. The DGS-Korpus project aims at language documentation, and this should also especially apply to the choice and properties of the examples. Editing examples means to basically change the language found in the corpus, which is shaped by face-to-face interaction. This means making utterances normally embedded in a discourse context self-contained, making the arguments of the verb explicit, avoiding digressions and – for the sake of having as clear example – removing superfluous information and redundancy. Although the reception of examples may be improved, the editing process results in the well-known text type called an example sentence, not a product of natural conversations – and it is natural conversation we want to document.<sup>13</sup>

## 5 Authentic Examples in the DGS Dictionary: Practical Experiences

Several dictionary projects include editorial examples based on corpus examples. Their editing process follows guidelines that have been created to produce examples suitable for the dictionary user (e.g. McKee & McKee 2013, Kristoffersen 2010). Editing includes actions such as changing the corpus utterance into a self-contained sentence by clarifying unresolved references, removing distracting elements and modifying the content for reasons of political correctness or balancing subjects.

In the DGS dictionary we include authentic examples from the original recording. This entails a number of consequences, because editing is not possible at all. Perceived shortcomings in the original recording have to be balanced by additional information, e.g. context information to ensure understanding of the example even in isolation from its context. Disturbing elements of performance (e.g. lax, disrupted, too fast signing) have either to be tolerated or lead to non-selection. As the choice of possible examples is limited to the corpus, occurrences using authentic examples is always a compromise between naturalness and accepting perceived imperfections. However, by means of careful selection, choosing the most suitable segment of the utterance and adding information (context, translation), it is possible to provide the user with useful and interesting examples.

13 One of our elicitation tasks asked for isolated signs known for regional variation, the use of which should be exemplified. When asked, the informants produced sentence-like, self-contained examples, often also useful evidence of signs other than the target sign. These sentences may fulfil the properties of a good example, but they may also appear artificial and similar to constructed examples.

## 5.1 Selection of Examples

Authentic examples are selected to best support the dictionary users' understanding of the senses. The following criteria have proven useful to achieve a good selection from the possible example candidates.

### 5.1.1 *Clear Illustration of Sense*

Atkins and Rundell (2008: 454) state that one important function of examples is to “illustrate usage”, and thus complement the more abstract definitions and help “clarify sense distinctions in a polysemous word”. Sometimes one utterance may cover more than one sense – depending on the interpretation – while another example illustrates only one specific sense clearly in contrast to the others. Examples that unambiguously illustrate one specific sense should be chosen. Moreover, examples should complement and support the definition by providing at least some semantic information on the sign's meaning in its surrounding context. The user should be able to infer from the examples some information about the meaning of the sign and typical contexts of its use in this specific sense. Synonyms or antonyms in the example utterance provide important clues on the sign's contextual meaning, and help to distinguish the sense more clearly. Utterances containing synonyms or antonyms are preferred to be chosen as examples, if available.

### 5.1.2 *Clarity of Sign Execution*

For the dictionary user the examples need to be understandable, and should be easy to perceive. Examples taken directly from a corpus may contain unclear and partially superfluous signing, such as false starts, or signs that are only indicated or executed in an unclear or idiosyncratic way. There may also be side remarks that distract from the relevant parts of the example. This and disfluencies in signing, e.g. breaks or searching for signs, can make it difficult to understand the content easily. Moreover, signing that is too quick can disturb the dictionary user, as the target sign may be seen too briefly to perceive it easily. All the previously mentioned difficulties should be avoided as much as possible, and examples with clear and undisturbed signing are preferred.

### 5.1.3 *Inclusion of a Wide Variety of Informants: Balancing for Region, Age, and Gender*

Since the corpus contains a lot of different peoples' signing this diversity should be reflected in the selection of examples. In the dictionary, the overall selection of examples is aimed to include all regions, age groups, both genders and as many informants as possible. When a particular sign's use is evidenced only for one region or mainly for a certain age group, the example selected for that sign will reflect this restricted use. No group should be overrepresented without need, as the dictionary is to be seen as a mirror of the data collected. Showing original informants avoids having to consider which sign model can best represent a certain region or group.<sup>14</sup>

### 5.1.4 *No Stereotypes, Discriminating Content and Avoiding Personal Information*

The contents of an authentic example should not be discriminating or reflecting stereotypes or prejudices. The non-selection of an example could also be a way of supporting a stereotype; for example, when only selecting examples occurring in the context of heterosexual partnerships for the sense being described as a “male partner in a relationship” when there is also a good example in the context of a homosexual relationship, like example (1).

<sup>14</sup> The ONZSL dictionary project used a group of eight sign models. They took special care to match the model's social factors to the language use of the original example (McKee & McKee 2013).

- (1) [Über Gunther Trube] Seine Mutter und sein **Lebensgefährte** haben einige Leute eingeladen.

Translation: [About Gunther Trube, a Deaf celebrity] His mom and his **husband** had invited some people.

Some informants reveal personal information about their lives or talked about people they know. Even though the informants have consented to the publication of their signing, we protect their and other people's privacy when needed. Examples containing very personal information are usually not selected. If it is not possible to do without a particular example with personal information, we actually re-record the example to achieve anonymization. This is one of the very rare cases where we do not show the original material. In example (2) we also replaced the real name with a very common German family name in the re-recording. This is a different case to that shown in example (1) with Gunther Trube, who is a well-known and even famous figure within the deaf community.

- (2) [Über eine Lehrerin] Ich mochte **Frau** Meyer, weil sie gebärden kann.

Translation: [About a teacher] I liked **Mrs.** Meyer, because she could sign.

### 5.1.5 Handling of Citation Form vs. Contextually Modified Forms

In use, signs do not always appear in their citation form, but are modified according to syntactic and morphologic rules, and thus occur as different word forms of the sign. Many signs can also be modified according to their iconic characteristics – a type of modification where the forms the sign can possibly take are not fully predictable. The sign form shown in the examples need not necessarily be of the citation form, on the contrary – a variety of forms that are typical of the lemma sign are preferred, as that shows the patterns of its use. This is valuable information for L2-learners of the language, to see how a sign can be spatially or iconically modified. Sign 7 in example (3) is a directional verb indicating the arguments of the verb by its directional execution. In citation form, the sign is signed from the signer's front towards themselves (in the sense of “someone gives back to me”); in example (3) it shows a modified form meaning “I give it back to *previous location* [*here: the shop*]”.

(3)



contextual meaning	mostly	I	glas	bottle
mouththing	-----meist-----		-----glas-----	



contextual meaning	<i>pointing</i>	can	return	can	<i>gesture</i>
mouththing		glas kann	wieder zurück	-----kann-----	

- ▣ [Thema: Umweltschutz und Pfandflaschen. Die Erzählerin kauft lieber Getränke in Glasflaschen als in Tetrapackverpackungen.] Ich nehme meistens Glasflaschen, die kann ich wieder **zurückbringen**.

Translation: [Subject: environmental protection and returnable bottles. The signer prefers glass bottles to Tetrapack carton packages when buying beverages.] Usually I take glass bottles because I can **return** them.

### 5.1.6 Treatment of Variants

DGS is an SL that shows many phonological variants. Usually such variants are described in one entry, thus one variant needs to be chosen to function as the lemma. This does not mean that examples containing other variants are avoided. It is instead the case that we aim to show a balanced selection of examples according to the variants' frequency. Sometimes authentic examples containing less frequent variants show the sense more clearly than possible examples containing the main variant of the lemma sign. In this case they are preferred over examples containing the main variant. Showing different variants also corresponds to the aim of documenting DGS and its range of variation.

## 5.2 Preparing Examples

After selecting a particular example, it is prepared for the inclusion in the dictionary entry. Preparation includes selecting the segment to be shown, adjusting the translation and adding a written context, if necessary.

### 5.2.1 Selecting the sequence to be shown

When deciding on which part of the utterance is to be shown it is important to consider the structures of DGS. Prosodic units should be completely included, hence utterance boundaries have to be considered. In example (3), theoretically sign 9 does not contribute to the meaning, and might have been cut off as it is a gesture mostly used to indicate that an utterance is finished. However, the prosodic phrase is not finished without sign 9, and the mouthing 'kann' stretches over signs 8 and 9. As such, leaving it out is not appropriate.

As for the content, a set of rules were established. The sequence needs to be long enough to properly illustrate the sign sense, but still short enough to be easily understood and perceived. When the shown segment is too long the user might lose focus, not recognize the sign and be distracted by other information that is not central in the signing.

### 5.2.2 Context

It is well known that genuine discourse examples taken out of context and shown in isolation are often hard to understand and not ideal in form and content, especially if they are taken from a spoken – or signed – conversation. Such authentic examples are difficult to comprehend in isolation, and need additional context information.

- (4) ▣ Dort waren nur wenige **Besucher**.

Translation: There were only few **guests**.

In example (4) is not clear what 'dort' (pointing sign) refers to. Here, context information is needed to understand the reference of the pointing sign and to comprehend the sense-specific semantic conditions of the sign's use. Thus the example would occur with a context (within square brackets) in the dictionary, as follows:



(5) [Thema: Festveranstaltung eines Basketballvereins] Dort waren nur wenige **Besucher**.

[Subject: festivity of a basketball club] There were only few **guests**.

Also without a larger context some examples would create a false impression of the informant (see example (6), being released from prison vs. being released from political imprisonment):

(6) [Der Erzähler wurde nach einem missglückten Fluchtversuch aus der DDR inhaftiert.] Als ich **aus** dem Gefängnis **kam**, war es mir wichtig, meine Ausbildung zum Tischler weiterzumachen. Die Prüfung habe ich dann auch bestanden.

Translation: [The signer was detained after an attempted escape from the GDR.] When I **got out** of prison, I concentrated on my training to become a carpenter. I passed the final exam.

For most examples we provide written information to promote the understanding of the selected examples. Necessary information is given, but should be as short and precise as possible. Whether a context is necessary or not may also depend on the user group. Native signers usually need less additional information than learners.

### 5.2.3 Translation

Each authentic example is presented in the entry along with a translation. For large parts of the corpus, translations are already available and time-aligned to the videos. These translations aid annotations and are also shown in the Public DGS Corpus. For the examples these translations can be used as the starting point for the translation shown with the example in the entry. However, as we mostly focus on a small segment it is necessary to check whether the translation of that part corresponds to the signing in the chosen segment. Additionally we adjust translations to be understandable in isolation. The translational equivalent of the sign is highlighted in bold face in the translation (cf. Figure 1).

## 6 Conclusion

In the special circumstances SL lexicography has to deal with, we found from our practical work that using authentic examples taken from original corpus recordings can be done and is a choice worthy of further consideration in corpus-based SL lexicography. This option was made possible by a large lemmatized and annotated corpus of SL with good technical quality of the original recordings and the informed consent of the informants for public use of their data. We consider the authentic examples to be valuable information to illustrate the usage of SL in a very typical, natural and lively – that is authentic – way. We are aware of the drawbacks of authentic examples, but also found that these can largely be compensated for by careful selection, preparation and additional information given in the related entries. As corpus-based SL lexicography is still in its first stages of development, and thus it remains to be seen whether other SL dictionary projects follow along this line, and how the best practices and standards will be developed in the future.

## References

### Dictionaries

- (D-ASL) Stokoe, W.C., Casterline, D.C. & Croneberg, C.G. (1965). *A dictionary of American Sign Language on linguistic principles*. Washington, DC: Gallaudet College Press.
- (D-LSFB) *Dictionnaire de LSFB en ligne et la grammaire de LSFB*. Accessed at: <http://dicto.lsfb.be> [29/03/2018].

- (D-NZSL) Kennedy, G., Arnold, R., Dugdale, P. & Moskovitz, D. (1998). *A dictionary of New Zealand Sign Language*. Auckland: Auckland University Press.
- (D-SAS) Penn, C. (1992-1994). *Dictionary of Southern African signs for communicating with the deaf*. Pretoria: Joint Project of the Human Sciences Research Council and the South African National Council for the Deaf.
- (KS-PJM) Łacheta, J., Czajkowska-Kisil, M., Linde-Usiekniewicz, J. & Rutkowski, P. (2016). *Korpusowy słownik polskiego języka migowego/Corpus-based Dictionary of Polish Sign Language*. Warsaw: Faculty of Polish Studies, University of Warsaw. Accessed at: <http://www.slownikpjm.uw.edu.pl/en> [29/03/2018].
- (ODT-SL) *Ordbog over Dansk Tegnsprog*. Center for Tegnsprog (2008-2016). Accessed at: <http://www.tegnsprok.dk> [29/03/2018].
- (ODT-VL) *Ordbog over Dansk Talesprog*. Københavns Universitet. Accessed at: <http://odt.hum.ku.dk> [14/05/2018].
- (ONZSL) McKee, D., McKee, R., Pivac Alexander, S., Pivac, L. & Vale, M. (2011). *Online dictionary of New Zealand Sign Language*. Wellington: Deaf Studies Research Unit, Victoria University of Wellington. Accessed at: <https://nzsl.vuw.ac.nz> [29/03/2018].

## Literature

- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bank, R., Crasborn, O. & van Hout, R. (2016) The prominence of spoken language elements in a sign language. In *Linguistics*, 54 (6), pp. 1281-1305.
- De Vos, C., Van der Kooij, E. & Crasborn, O. (2009). Mixed Signals: Combining Linguistic and Affective Functions of Eyebrows in Questions in Sign Language of the Netherlands. In *Language and Speech*, 52 (2-3), pp. 315-339.
- Ebbinghaus, H. (2001). Wird sich die Sprache der Gehörlosen 'vertikalisieren'? Möglichkeiten und Grenzen der Registerdifferenzierung in der Deutschen Gebärdensprache. In G. List, G. List (eds.) *Quersprachigkeit. Zum transkulturellen Registergebrauch in Laut- und Gebärdensprachen*. Tübingen: Stauffenburg Verlag, pp. 187-199.
- Ebbinghaus, H., Hessmann, J. (2001). Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different things. In P. Boyes Braem, R. Sutton-Spence (eds.) *The hands are the head of the mouth. The mouth as articulator in sign languages*. Hamburg: Signum-Verlag, pp. 133-151.
- Ehlich, K. (1983). Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung. In A. Assmann, J. Assmann & C. Hardmeier (eds.) *Schrift und Gedächtnis. Beiträge zur Archäologie der literarischen Kommunikation*. München: Wilhelm Fink Verlag, pp. 24-43.
- Erlenkamp, S. (1998). Lexikographie als Teil einer Minderheitenpolitik – Minderheitenpolitik als Teil einer Lexikographie. In *Das Zeichen*, 12 (43), pp. 98-105.
- Fehrmann, G., Linz, E. (2009). Eine Medientheorie ohne Medien? Zur Unterscheidung von konzeptioneller und medialer Mündlichkeit und Schriftlichkeit. In E. Birk, J.G. Schneider (eds.) *Philosophie der Schrift*. Tübingen: Max Niemeyer, pp. 123-143.
- Hansen, C., Hansen, M.H. (2012). A Dictionary of spoken Danish. In R. Vatvedt Fjeld, J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress, 7.-11. August 2012*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929-935.
- Jahn, E., Konrad, R., Langer, G., Wagner, S. & Hanke, T. (2018). DGS-Korpus Project: Different Formats for Different Needs. In M. Bono, E. Efthimiou, S.E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch & Y. Osugi (eds.) *Involving the Language Community. Proceedings of the 8th Workshop on the Representation and Processing of Sign Language. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*. Paris: Elra, pp. 83-90.
- Keating, E., Edwards, T. & Mirus, G. (2008). Cybersign: Impacts of New Communication Technologies on Space and Language. In *Journal of Pragmatics*, 40 (6), pp. 1067-1081.
- Kennedy, G. (1996). Designing a dictionary of New Zealand Sign Language. In G. Kennedy (ed.) *Topics in New Zealand Sign Language Studies*. Wellington: Victoria University of Wellington, pp. 27-48.
- Kristoffersen, J.H. (2010). *From utterance to example sentence – excerpting and adapting corpus sentences for use in the Danish Sign Language Dictionary*. [Poster from the Sign Linguistics Corpora Network. 4. Workshop: Exploitation, Berlin, 3.-4. December 2010.] Accessed at: [http://www.tegnsprog.dk/hovedside/litteratur/SLCN\\_Berlin\\_2010\\_poster.pdf](http://www.tegnsprog.dk/hovedside/litteratur/SLCN_Berlin_2010_poster.pdf) [29/03/2018].

- Kristoffersen, J.H., Troelsgård, T. (2010). The Danish Sign Language Dictionary. In A. Dykstra, T. Schoonheim, (eds.) *Proceedings of the 14th EURALEX International Congress, 6.-10. July 2010*. Leeuwarden/Ljouwert: Fryske Akademy, pp. 1549-1554.
- Krentz, C.B. (2007). The Camera as Printing Press: How Film has Influenced ASL Literature. In H-D.L. Bauman, J.L. Nelson & H.M. Rose (eds.) *Signing the Body Poetic*. Berkeley: University of California Press, pp. 51-70.
- Langer, G., Müller, A., Wähl, S. (2018). Queries and Views in iLex to Support Corpus-based Lexicographic Work on German Sign Language (DGS). In M. Bono, E. Efthimiou, S.E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch & Y. Osugi (eds.) *Involving the Language Community. Proceedings of the 8th Workshop on the Representation and Processing of Sign Language. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*. Paris: Elra, pp. 107-114.
- Linde-Usiekniewicz, J., Rutkowski, P. (2016). The division into parts of speech in the Corpus-based Dictionary of Polish Sign Language. In T. Margalitadze, G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress, 6-10 September 2016*. Tbilisi: Ivane Javakhishvili Tbilisi University Press, pp. 375-388.
- McKee, R., McKee, D. (2013). Making an Online Dictionary of New Zealand Sign Language. In *Lexikos*, 23, pp. 500-531. Accessed at: <http://dx.doi.org/10.5788/23-1-1227> [16/05/2018].
- Möhrs, C., Meliss, M. & Batinić, D. (2017). LeGeDe - towards a corpus-based lexical resource of spoken German. In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19.-21. September 2017*. Brno: Lexical Computing CZ s.r.o., pp. 281-298.
- Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T. & Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) Corpus Project. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz & A. Schembri (eds.) *Corpora and Sign Language Technologies. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta*. Paris: ELRA, pp. 178-185.
- Prinsloo, D.J., Gouws, R.H. (2000). The Use of Examples in Polyfunctional Dictionaries. In *Lexikos*, 10, pp. 138-156. Accessed at: <http://dx.doi.org/10.5788/10-0-891> [29/03/2018].
- Schwitalla, J. (1997). *Gesprochenes Deutsch. Eine Einführung*. Berlin: Schmidt.
- Stokoe, W.C. (1993). Dictionary making, then and now. In *Sign Language Studies*, 22 (79), pp. 126-146. *Sutton's SignWriting Site*. Accessed at: <http://www.signwriting.org> [29/03/2018].
- Trap-Jensen, L. (2004). Spoken language in dictionaries: Does it really matter? In G. Williams, S. Vessiers (eds.) *Proceedings of the Eleventh EURALEX International Congress 2004*. Lorient: Faculté des lettres et des sciences humaines, Université de Bretagne-Sud, pp. 311-318.
- Verdonik, D., Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. In *International Journal of Lexicography*, 30 (2), pp. 143-166.
- Wöhrmann, S. (2003). GebärdenSchrift lesen lernen. In *Das Zeichen*, 17 (65), pp. 364-374.

## Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Program, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Program is coordinated by the Union of the German Academies of Sciences and Humanities.



# Multimodal Corpus Lexicography: Compiling a Corpus-based Bilingual Modern Greek – Greek Sign Language Dictionary

*Anna Vacalopoulou, Eleni Efthimiou, Kiki Vasilaki*

*ILSP-Institute for Language and Speech Processing / Athena RC*

*E-mail: avacalop@ilsp.gr, eleni\_e@ilsp.gr, kvasilaki@ilsp.gr*

## Abstract

This paper describes the process of compiling NOEMA+, a bilingual dictionary of approximately 12,000 entries for the pair Greek Sign Language (GSL) - Modern Greek (MG) and of making it available openly online (<http://sign.ilsp.gr/signilsp-site/index.php/el/noima/>). The dictionary was based on several corpora that have been collected over the years, including information on compounding, GSL synonyms, classifiers, various lemma-related senses, semantic relationships, etc. These different corpora have been joined, normalized and translated into MG to form a parallel corpus of the language pair in question. In turn, this parallel corpus acted as the basis for the compilation of the bilingual dictionary described in this paper.

More specifically, among the issues to be discussed here are lemma identification, which proved far from intuitive for this particular language pair, lemma categorization, dictionary contents and structure, relations between entries as well as the corpus which was used for dictionary compilation. Finally, there will be a description of the different search choices offered, which cater for different user profiles and needs.

**Keywords:** bilingual lexicography, corpus-based lexicography, multimodal parallel corpus, sign language lexicography

## 1 Introduction and Background

In recent years, awareness raising efforts have placed the issue of equal rights at the top of national, European and international policy agendas. From the Convention on the Rights of Persons with Disabilities in 2006 (CRPD) to the EU Disability Strategy 2010-2020 (European Disability Strategy 2010-2020) there is an international movement towards providing all persons with the same rights, meaning, among other things, total accessibility to knowledge and information. In the case of deaf and hard-of-hearing (HoH) persons, several attempts have been made towards this goal. Among the many systematic attempts to compile European Sign Language (SL) corpora, are the British Sign Language Corpus (Schembri 2008), the Corpus NGT (Crasborn, O. & Zwitterlood 2008) of the Sign Language of the Netherlands, the e-LIS corpus of the Italian Sign Language (Vettori 2008), the German Sign Language Corpus (Konrad 2009), and the Swedish Sign Language Corpus (Mesch J. & Wallin 2015). At the same time, a number of lexicographic attempts have also been made. Indicative projects include, among others, dictionaries of Spanish SL (Fuertes et al. 2006), of Danish SL (Kristoffersen, & Troelsgård 2010), of Polish SL (Linde-Usiekiewicz et al. 2014), and of British SL (Fenlon 2014).

The dictionary described in this paper was first created as part of a specially designed workbench bringing together several Language Technology (LT) tools and technologies. The purpose of this venture was to provide better accessibility to an official educational content platform of the Greek Ministry of Education to deaf and HoH users, i.e. in their native language (Efthimiou et al. 2016). To this end, the workbench needed to provide all the necessary tools for the target group to make



full use of both the content in the platform in question and the Graphical User Interface (GUI) features.<sup>1</sup>

In order to approach this diverse material, which was designed for primary and secondary education levels, we brought together a combination of LT tools (Efthimiou et al. 2017) to handle both SL as well as written language, including a tagger and a lemmatizer. As far as resources are concerned, we made full use of already available monolingual GSL corpora as well as bilingual MG-GSL dictionaries. The result of that work was a fully functioning bilingual service, offering to the target audience bilingual and monolingual dictionary look-up, fingerspelling facilities as well as a dynamic SL synthesis environment.

As part of this project, NOEMA+, the dictionary described in this paper, underwent several stages of evaluation (*ibid.*) both internally and externally by professionals, GSL experts (some of whom were native signers), and actual end users. The results of this process, along with the availability of an updated version of the GSL corpus in greater detail, presented us with the opportunity to update and improve NOEMA+ as a separate lexical resource.

Briefly speaking, the tools and technologies that were incorporated in the workbench fall under those created for vocal or written language and those created for sign language.

As far as the first category is concerned, we used a language tool suite (Prokopidis et al. 2011) developed and continually undergoing adjustments by the Institute for Language and Speech Processing (ILSP). This includes a tool that segments character strings within text, a tagger that goes through the segmented units adding to them labels with grammatical information (i.e. morphological and syntactical categories), and a lemmatizer that links each of the segmented and tagged items with a particular lemma in the dictionary. The integration of an improved version of this suite proved to be especially valuable, as morphological complexity is one of the main characteristics of MG, which is a particularly inflectional language (Holton et al. 1997). Indeed, the numerous forms of lemmas that differ considerably or even completely from each other (i.e. irregular types) have been known to present serious barriers to the literacy of deaf users of around the age of the target group (Breadmore 2007). In the case of this platform, such cases would pose problems not only for text comprehension, but also for looking up lexical items in the dictionary.

As regards the SL part of the content, the tools and technologies that were incorporated in the workbench include the following: (a) the bilingual dictionary in question, which was linked with content uploaded in the educational platform, namely textbooks, offering different search possibilities; (b) a keyboard allowing virtual fingerspelling, which consists of gestural equivalents of the Greek alphabet characters and the digits 0-9, used mainly for visually spelling proper names and numbers to be found across all the subjects of the curriculum; and (c) a tool enabling the dynamic synthesis of sign phrases, in which a signing avatar performs GSL phrases typed in by users in real time (Efthimiou et al. 2017).

Finally, the workbench offers a series of tools enabling the use of the GUI features of the platform by deaf and HoH users. These comprise appropriately employed color code conventions and pop-up windows across the interface, help buttons of suitable shape and size, and extra help in the form of a video tooltip (Efthimiou et al. 2016).

1 The research leading to this output received funding from the POLYTROPON project (KRIPIS-GSRT, MIS: 448306) and was based on insights, technologies and language resources initially developed within the Dicta-Sign project (FP7-ICT, grant agreement n°: 231135).

## 2 Corpus Design and Contents

The POLYTROPON corpus is a bilingual parallel corpus for the language pair GSL-MG, which was designed with multiple applications in mind, as is often the case with parallel and comparable corpora (McEnery & Xiao 2007). More specifically, it was meant to serve as a “golden” corpus for the creation of both other lexical and terminological resources and sign language technologies, such as machine learning and machine translation from and into sign language. The creation of the corpus was based on data derived from several corpora that have been collected over the years by means of HD and Kinect cameras within the context of different projects, and it incorporates various data. The two main pre-existing resources were the list of entries of the bilingual GSL-MG multimedia dictionary project NOEMA (Efthimiou & Katsoyannou 2001) and a set of 2,000 lemmas extracted from the segmentation procedure of Dicta-Sign Corpus (Matthes et al. 2012). These two different corpora resulted in a rich set of GSL data, which includes information on signs, synonyms, compounds, classifier constructions,<sup>2</sup> different senses of each single lemma form, and so on.

The methodology followed to create the corpus was divided into three stages, as explicitly described in Efthimiou et al. (2018). In the first stage of the process, all lemmas found in the two pre-existing corpora (approximately 2,000), were reviewed by a working group of SL experts. During this phase, the status of each selected GSL entity had to be evaluated and validated as a simple sign, compound or classifier construction. In the second stage, examples of use for each GSL discussed lemma (classifiers excluded) were created and recorded in three repetitions by means of HD and Kinect cameras. This was mainly decided in order to make the lexicon resource exploitable in machine learning experiments targeting machine translation and sign recognition to illustrate articulation variation, as there can be no two identical video captures of the same token. This is also a key difference between sign language corpora and written language corpora, small or larger chunks of which can and will be repeated in the corpus numerous times. Moreover, the discussions among GSL experts added some 1,600 new lemmas to the corpus, which had not initially been included in the lexicon. This procedure led to the generation of new sentences which were connected to the new lemmas. In the third stage of the procedure, one out of every three recordings of the examples of use was annotated in iLex, a software tool for linguistics analysis of sign language data (Hanke & Storz 2008; Efthimiou et al. 2016). For that reason the annotation procedure of the example of use includes the following set of tiers:

- Clause: determines the clause boundaries of the signed clauses.
- Gloss: assigns an MG equivalent to each GSL identified entity.
- Greek equivalent clause: an MG translation is provided in each GSL example of use. The translation of the annotated clauses was completed in two phases: First, a strongly GSL-like translation was provided by the corpus annotator and, then, each example was reviewed in terms of naturalness and grammaticality by an expert in MG.
- Classifier: analyses classifiers occurring in the examples with respect to their morpho-phonological and semantic status. For dictionary compilation purposes, we were particularly interested in classifier constructions that combine with single signs to create compounds (Efthimiou 2010).
- S-type: marks GSL examples as main or subordinate constructions.
- S-category: marks each sentence type as declarative, negative, interrogative, imperative, or exclamatory.

For reasons of quality assurance and consistency across annotators, a peer-to-peer crosscheck procedure was followed along the lines of the earlier Dicta-Sign Corpus procedure (Dimou et al., 2011).

2 Classifiers are found in almost all SLs. They are morphemes expressed by meaningful hand configurations and often occur with verbs that express: i) motion through space, ii) the location or existence of a referent or iii) a referent that is being held (Efthimiou 2004).

This was essential during every new addition to the corpus in order to ensure that the kappa score for inter-annotator agreement was kept at a very high level.

### 3 Dictionary Compilation

The largest part of entries in NOEMA+ was based on an analysis of the annotated GSL corpus. As implied in the previous section, the identification of separate lemmas in such a newly developed lexical resource as a video SL corpus is far from intuitive. Lemma identification for this particular language pair proved to be a challenging task, as lexicographers had to select autonomous tokens representing separate concepts on their own. There were also frequent cases when a GSL lemma would not have a one-to-one equivalence with a lemma in MG or the other way around. This, of course, is far from surprising in bilingual lexicography, with an abundance of articles (Zgusta 1971, Snell-Hornby 1984, Piotrowski 1994) on similar problems having been published, proving that there is no one-to-one correspondence between lexical structures. As in this dictionary one of the languages presented is not an oral/written one, but rather a three-dimensional language, the problem of equivalence proved to be even more challenging.

In an attempt to get round this problem, lexicographers decided that the best solution would be to start by translating the signed examples of use first before getting to each lemma and its respective sense in the example. As expected, the lemmas were much easier to translate in the context of the examples, from which the lemma translations were then derived. For practical as well as illustrative reasons, some of the examples extracted from the corpus had to be shortened or simplified, as is the case in many lexicographic projects (Kilgarriff 2013). We soon discovered that this cannot be performed by processing an already existing caption, as this would result in a lack of naturalness in the GSL examples. This created the need for even more video captions. At the end of this process, a two-column table of equivalent sentences was created, which served as the basis for finalizing the list of lemmas/senses. Apart from classifier constructions, further items were excluded from this list based on the fact that they were not considered by lexicographers as an integral part of the core MG vocabulary. Such items were:

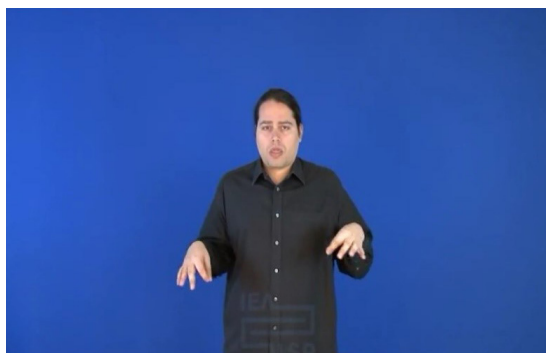


Figure 1: GSL expression “EMPTY-POCKETS” meaning “broke”.

- Signs representing brand names, football teams or other organizations, such as Ikea, Olympiacos, and Facebook.
- A set of GSL-specific expressions that have no direct equivalent in MG. For instance, the sign “EARS-DOWN”, which expresses the meaning of “obey” and that of “EMPTY-POCKETS” (Figure 1), which means “broke”.
- A set of gestures with semantic value equivalent to embodied signals occurring in oral communication in MG. These meaningful gestures add extra-linguistic information to utterances such as “I don’t know” (Figure 2) or “What can I say”, which also occur in GSL.

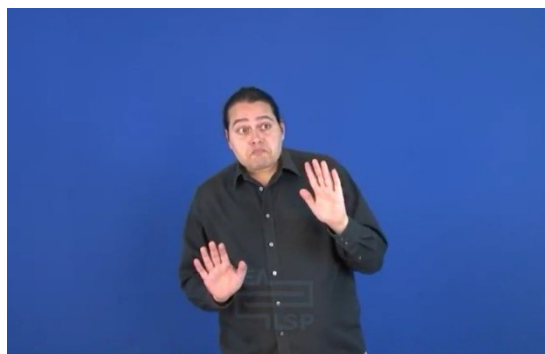


Figure 2: Embodied gestures found both in GSL and MG.

The outcome of this first compilation step was compared to existing MG dictionaries and glossaries, so that basic lemmas from different closed categories would not be left out. During this step, lexical items such as numbers, days of the week, months, seasons, units of measurement, geographical locations, etc., were included for reasons of comprehensiveness. As research on the basic vocabulary of Greek is still scarce (Vacalopoulou & Efthimiou 2015) lexicographers decided to include in it a number of lexical items that are particularly frequent in standard MG corpora. To this end, a frequency list from the Hellenic National Corpus (HNC, <http://hnc.ilsp.gr/>), a large, POS-tagged and lemmatized general-language MG corpus (Hatzigeorgiu et al. 2000), was cross-checked against the list of entries to locate more potential entries. However, a large number of top frequency items, among the first 2,000, had to be excluded automatically, as these typically are content-free function words, such as prepositions, articles or conjunctions, which do not correspond to any concrete sense but help in the formation of grammatical phrases and sentences in most vocal languages. Finally, extra care was taken in order to ensure that every sign (excluding those falling into the categories mentioned above) as well as every written word (excluding proper names and function words) that appear in the NOE-MA+ examples is also a lemma in itself, following standard lexicographic practice. Finally, alternate orthographies of MG entries were also added to the list for reasons of comprehensiveness.

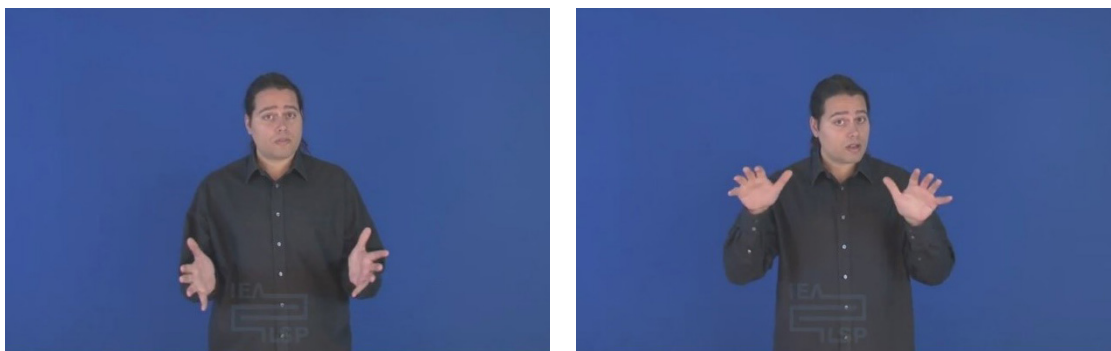


Figure 3: Alternate sign formations for the literal sense of the word “barrel”.

All these enhancement steps resulted in a new output, almost double the size of the original dictionary. As mentioned above, the new contents of the dictionary were piloted with different groups of informants, who identified possible adjustments and improvements so that the final product would have the widest possible acceptance of end users. As the source language of this bilingual dictionary is GSL, it was only natural that alternate sign formations (Figure 3) were also included at this stage by GSL experts so that there could be a balanced representation of variants in both languages. It is worth noting that this process helped, in turn, the enhancement of the original GSL corpus, as more lemmas and respective examples of use were being added.

The form of the GSL entries varies according to morphology. In particular, apart from basic, monomorphemic signs, several types of compounds are also formed as in other SLs (Brown 2010). In particular, the compounding options found as dictionary entries include the following combinations (Efthimiou et al. 2018):

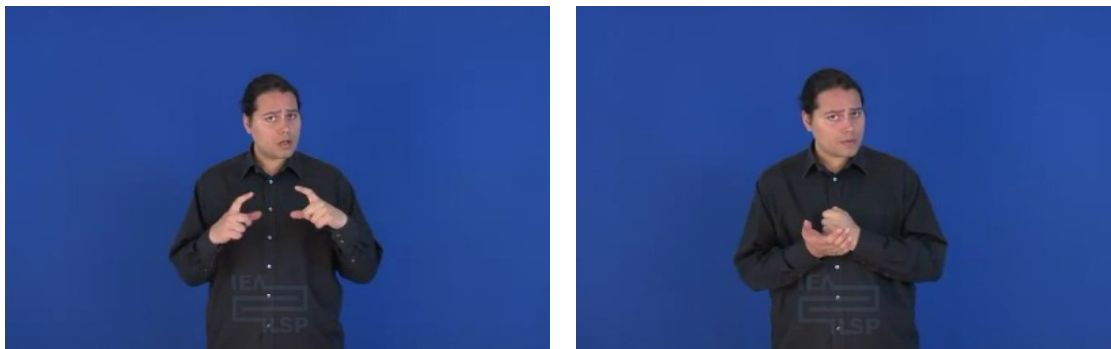


Figure 4: The compound “letter” formed by C1 (indicating shape) and the sign for “seal”.

- sign+sign: compounds such as “FLIGHT ATTENDANT” formed by two individual signs (i.e. AIRPLANE+ACCOMPANY).
- classifier+sign: compounds consisting of a classifier construction and a monomorphemic sign, such as C1+SEAL to form the equivalent of the word “letter” (Figure 4).
- classifier+classifier: compounds formed by the combination of two classifier constructions, such as C5+5 to represent the word “lighthouse” in the literal sense (Figure 5).

As far as dictionary structure is concerned, each GSL entry is accompanied by one or more MG equivalents for each sense it represents, by synonyms (if any) in either language, and by simple examples of use. When applicable, multi-word MG entries are linked to their respective single-word ones (excluding functional words) via cross-references. Apart from facilitating easy reference, this feature also has pedagogical added value, considering that most of the words which form these phrases are inflected types of other entries. It thus becomes easier for users to link each inflected type to the base form of the entry. A deliberate decision to exclude any metalinguistic information at this point was made in order to make NOEMA+ more user-friendly to the primary target audience, i.e. native signers. As the dictionary was based on the aforementioned corpus, the vast majority of the examples of use are authentic as opposed to constructed ones, whereas a small number of them were created ad hoc. As there was more than one video capture for each lemma and each example in the corpus, careful consideration was needed in order to select the best candidates for inclusion in the NOEMA+, meaning the most ‘representative’ ones. Again, the final decision for the most common and natural capture was made by groups of GSL experts, some of whom native signers.

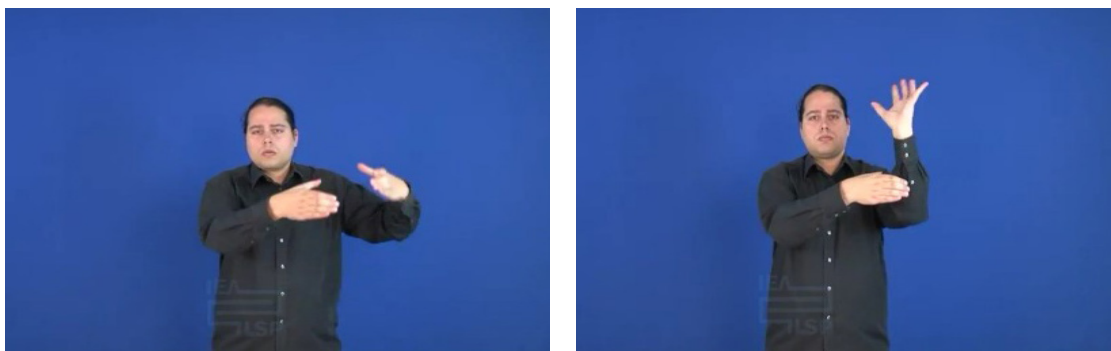


Figure 5: The compound “lighthouse” formed by two classifiers.



Finally, another convention adopted by lexicographers was to translate all GSL entries in the form they would be found as entries themselves in MG dictionaries for reasons of consistency and easy reference. Therefore, verbs appear in the first person singular present in the active voice; nouns appear in the singular nominative; adjectives and past participles appear in the nominative positive (in this case, in the masculine, feminine and neutral); adverbs appear in the positive form. As in MG dictionaries, the only exceptions occur when what is considered as the base form is either ungrammatical or particularly infrequent in the language. Indeed, this convention has proved particularly effective with regard to dictionary look-up, as explained in the next section of this paper.

## 4 Dictionary Look-up

One of the greatest challenges in this project relates to the presentation of the content rather than its compilation. This is based on the paradox that, although the source language of this electronic reference work is GSL, no search options are available in this language due to its three-dimensional nature. In other words, users who are native in the source language will have to perform searches in the target language. This barrier was also the reason behind several choices in the dictionary-making process. For instance, contrary to standard lexicographic practice, the addition of entries in NOEMA+ was made taking into account both the source language and the search language, which are not identical.

Although great care was taken to double-check every entry both ways, it was obvious that users, be they either beginner learners or regular users of GSL, would need to be presented with several alternative options, in order to make successful searches. Therefore, they are provided with three choices (Figure 6), i.e. type in search items, use a virtual fingerspelling keyboard, or select their search item from an alphabetical list of entries.



Figure 6: Looking up a word in NOEMA+.

## 5 Summary and Results

In this paper we present lexicographic work targeted at the development of a bilingual dictionary of approximately 12,000 entries for the pair GSL-MG. This reference work was initially developed as part of an official educational content platform of the Greek Ministry of Education offering open access to its content by deaf and HoH users in their native language. In addition to that, it has now grown to be a much richer dictionary, offered to end users as a standalone service.

The processes of dictionary compilation as well as the way users look up lexical items were particularly challenging, based on the paradox of the source language not coinciding with the search

language. We found, however, that apart from posing practical obstacles, this reality enabled us to apply two-way checks from source to target language and vice versa; not only did this help us avoid errors and misunderstandings (given that a large part of the GSL content had never been recorded in a dictionary before), but it also resulted in easier and quicker enhancement of the contents of both the dictionary and GSL corpus. In fact, because of the limited lexicographic work in GSL, we dare say that NOEMA+ is one among few such projects that cannot escape being corpus-driven. In the words of Tognini-Bonelli (2001: 84), the “corpus [...] is seen as more than a repository of examples to back pre-existing theories [...]”

## References

- Breadmore, H.L. (2008). Inflectional morphology in the literacy of deaf children, (PhD thesis, University of Birmingham, Birmingham, UK). Accessed at: <http://etheses.bham.ac.uk/591/> [25/03/2018].
- Brown, K. & Ogilvie, S. (eds.) (2006) *Concise Encyclopedia of Languages of the World*. P. 324. Elsevier, Amsterdam. 14 vols.
- Crasborn, O. & Zwitserlood, I. (2008). The Corpus NGT: an online corpus for professionals and laymen. In O. A. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, E. Thoutenhooft (eds.) *Construction and Exploitation of Sign Language Corpora*. (pp. 44-49). ELDA, Paris.
- CRPD – *The Convention on the Rights of Persons with Disabilities*. Accessed at: <https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html> [26/03/2018]
- Dimou, A.L., Goulas T. & Efthimiou E. (2011) Grammar/Prosody Modelling in Greek Sign Language: Towards the definition of built-in sign synthesis rules. In E. Efthimiou, G. Kouroupetroglou, S.E. Fotinea (eds.) *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, vol.7206, pp. 183-193. Berlin, Heidelberg: Springer LNCS 7206.
- Efthimiou, E., Vasilaki, K., Fotinea, SE., Vacalopoulou, A., Goulas, T. & Dimou, A.L. (2018) The POLYTROPON Parallel Corpus. In proceedings of the *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community at LREC2018*, Miyazaki [to be published].
- Efthimiou, E., Fotinea, S.E., Kakoulidis, P., Goulas, T., Dimou, A.L. & Vacalopoulou, A. (2017) Sign Search and Sign Synthesis Made Easy to End User: The Paradigm of Building a SL Oriented Interface for Accessing and Managing Educational Content. In: Antona M., Stephanidis C. (eds.) *Universal Access in Human-Computer Interaction. Designing Novel Interactions*. UAHCI 2017. Lecture Notes in Computer Science, vol. 10278. Springer, Cham.
- Efthimiou, E., Fotinea, S.E., Goulas, T., Kakoulidis, P., Dimou, A.L. & Vacalopoulou, A. (2016) A Complete Environment for Deaf Learner Support in the Context of Mainstream Education. In *Proceedings of the Conference Universal Learning Design*, vol. 5. ISSN 1805-3947 Linz, 13-15 July, pp. 35-44.
- Efthimiou, E., Fotinea, SE., Dimou, AL. & Kalimeris, C. (2010) Towards decoding Classifier function in GSL. In *the Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Satellite workshop of the LREC-2010 Conference*, pp.76-79, Valetta.
- Efthimiou, E. & Katsoyannou M. (2001) Research issues on GSL: a study of vocabulary and lexicon creation. In *Studies in Greek Linguistics, Computational Linguistics 2*:42-50 (in Greek).
- European Disability Strategy 2010-2020*. Accessed at: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0636:FIN:en:PDF> [26/03/2018].
- Fenlon, J., Cormier, K.A., Rentelis, R., Schembri, A., Rowley, K., Adam, R. & Woll, B. (2014) BSL SignBank: A lexical database and dictionary of British Sign Language (1st Edition). Accessed at: <http://bslsignbank.ucl.ac.uk/> [26/03/2018].
- Fuertes, J.L., González, Á.L., Mariscal, G. & Ruiz, C. (2006) Bilingual Sign Language Dictionary. In K. Miesenberger, J. Klaus, W.L. Zagler, A.I. Karshmer (eds.) *Computers Helping People with Special Needs. ICCHP 2006. Lecture Notes in Computer Science*, vol 4061. Springer, Berlin, Heidelberg
- Hanke, T. & Storz, J. (2008). “iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography”. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhooft (eds.) *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, pp. 64-67 ELRA, Paris.

- Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H. & Demiros, I. (2000). Design and implementation of the online ILSP Greek Corpus. In *Proceedings of the second international conference of language resources and evaluation (LREC) (vol. 3, pp. 1737-1740)*, Athens.
- Holton, D., Mackridge P. & Philippaki-Warbuton I. (1997). *Greek: A Comprehensive Grammar of the Modern Language*. London: Routledge.
- Kilgariff, A. (2013) Using corpora [and the web] as data sources for dictionaries. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*, pp. 77-96. Bloomsbury Publishing, London.
- Konrad, R. & Langer, G. (2009) Synergies between transcription and lexical database building: The case of German Sign Language (DGS). In M. Mahlberg, V. González-Díaz, C. Smith (eds.). *Proceedings of the Corpus Linguistics Conference (CL2009)*. University of Liverpool, UK, 20-23 July 2009. Accessed at: <http://ucrel.lancs.ac.uk/publications/cl2009/#papers;%20Article%20#346> [25/03/2018].
- Kristoffersen, J.H. & Troelsgård T. (2010). In Dykstra, A. & Schoonheim T. (eds.) *Proceedings of the 14th EURALEX International Congress*, pp. 1549-1554. Fryske Akademy, Leeuwarden/Ljouwert.
- Linde-Usiekniewicz J., Czajkowska-Kisil M., Łacheta J. & Rutkowski P. (2014) A corpus-based dictionary of Polish Sign Language (PJM). In A. Abel, C. Vettori, N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The user in focus*, pp. 365-376. EURAC research, Bolzano/Bozen.
- Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, A.L., Braffort, A., Glauert, J. & Safar, E. (2012). Dicta-Sign – Building a Multilingual Sign Language Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*, Istanbul.
- McEnery A.M. & Xiao R.Z. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist*. Translating Europe. Multilingual Matters, Clevedon, UK.
- Mesch J. & Wallin L. (2015) Gloss annotations in the Swedish Sign Language corpus. In *International Journal of Corpus Linguistics*, 20(1), 102–120.
- Piotrowski T. 1994. *Problems in Bilingual Lexicography*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.
- Prokopidis, P., Georgandopoulos, B. & Papageorgiou, Ch. (2011) A suite of NLP tools for Greek. In *The 10th International Conference of Greek Linguistics*. Komotini, Greece.
- Schembri, A. (2008). British Sign Language corpus project: open access archives and the observer's paradox. In *the 6th International Conference on Language Resources and Evaluation, LREC, May 26-June 1, 2008*, Marrakech.
- Snell-Hornby, M. (1986) The bilingual dictionary – Victim of its own tradition? In R.R.K. Harmann (ed.) *The History of Lexicography*. John Benjamins, Amsterdam.
- Tognini-Bonelli, E. (2001) The corpus-driven approach. In W. Teubert, R. Krishnamurthy (eds.) *Corpus Linguistics at Work*. John Benjamins, Amsterdam/Philadelphia.
- Vacalopoulou, A. & Efthimiou, E. (2015). Multilingual lexicography for adult immigrant groups: bringing strange bedfellows together. In I. Kosem, M. Jakubíček, J. Kallas, S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, Herstmonceux Castle, 11-13 August 2015*. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., Ljubljana/Brighton.
- Vettori, C. & Felice, M. (2008) e-LIS: Electronic Bilingual Dictionary Italian Sign Language-Italian. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*, pp. 791-796. Universitat Pompeu Fabra, Barcelona.
- Zgusta, L. (1971) *Manual of Lexicography*. Janua Linguarum. Series Maior 39. Mouton, The Hague/Paris.



# **Phraseology and Collocation**





# Bilingual Corpus Lexicography: New English-Russian Dictionary of Idioms

**Guzel Gizatova**

*Department of European Languages and Cultures, Kazan Federal University*

*E-mail: guzelgizatova@hotmail.com*

## Abstract

The paper deals with the principles of constructing the first printed and on-line English-Russian dictionary of idioms based on corpus data. The need for a new dictionary of idioms is motivated by the fact that there is presently no corpus-based dictionary of English-Russian idioms built on authentic examples. Existing traditional bilingual dictionaries do not meet modern requirements of the present-day lexicography with respect to vocabulary and illustrative examples, which are often out of date. This is definitely connected with the fact that traditional English-Russian idiomatic dictionaries were constructed in the ‘pre-corpus era’.

The purpose of the present research is thus to introduce a methodology for generating a comprehensive idiom list of the dictionary, to consider linguistic issues presenting difficulties in bilingual lexicography related to the concept of equivalence in idioms, and analyze the semantic asymmetry between English and Russian idioms.

**Keywords:** idioms, corpora, bilingual lexicography

## 1 Introduction

Bilingual lexicography encounters certain problems in connection with the treatment of idioms in dictionaries. In many cases, the generally accepted equivalent of an idiom cannot be used to translate authentic texts, which is why our research strategy is to analyze cross-linguistic correlations between English and Russian idioms that have strong semantic resemblance, as well as to study semantic asymmetry in idioms.

Using relevant lexicographic information, text corpora, and parallel corpora, we shall study the frequency and semantic qualities of idioms by empirical methods in order to identify additional specific features that need to be included in the lexicographic description of an idiom.

The need for a new English-Russian Dictionary of Idioms is indicated by the fact that there is presently no corpus-based bilingual dictionary of these languages. Due to corpus data, the dictionary presents a range of syntactic patterns, polysemous idioms, and unexpected variants, which cannot be retrieved from the existing bilingual and monolingual dictionaries of the English and Russian languages. Many dictionaries fail to register all meanings of idioms. The corpora help to reveal the specific character of their functional correlations and non-trivial semantic preferences of English idioms that do not have standard Russian equivalences.

The primary goal of the research is to conduct a thorough contrastive analysis with the purpose of discovering the unique properties of each idiom and thus enhance the lexicographical description of phraseological studies. The author analyzes one of the most curious cases of semantic asymmetry – phraseological ‘false friends’ (Piirainen 1997). Compare an English idiom (1) and its Russian pseudo-equivalent (2).

- (1) English *to twist (turn, wrap) somebody (a)round one's finger*  
 'to have the ability to persuade (a person) to do exactly as one wants (usually used to describe wives and daughters who persuade their husbands and fathers)' (Longman 1979: 113).
- (2) Russian *обвести вокруг пальца*  
 "to twist somebody (a)round one's finger"  
 'to deceive somebody skillfully' (Lubensky 2004: 446).

These two idioms are basically equivalent, since they are identical with respect to both their lexicalized meaning and image component. However, it is actually wrong to translate the English idiom *to twist (turn, wrap) somebody (a)round one's finger* by the Russian idiom *обвести вокруг пальца*. Analysis of authentic texts in corpora with the idiom *to twist (turn, wrap) somebody (a)round one's finger* shows that this idiom is translated into Russian by the idiom *вить веревки* (3).

- (3) Russian *вить веревки (из кого-либо)*  
 "to twist the ropes (from someone)"  
 'compel someone to your will and force him to act your way' (Birikh 2005: 89).

The usage of idioms in authentic texts from the corpus query system Sketch Engine is illustrated in the following examples.

He has worked there for 38 years and is planning to retire soon. His family includes his wife Linda, a son Jeff and a granddaughter 'who absolutely has me *twisted around her little finger*'. (enTenTen13).

Он работает там уже 38 лет и скоро планирует выйти на пенсию. У него есть жена Линда, сын Джефф и внучка, которая 'беспощадно *вьет из меня веревки*' (enTenTen13).

Thus, despite the formal similarity of the idioms *to twist (turn, wrap) somebody (a)round one's finger* and *обвести вокруг пальца*, this similarity cannot be considered complete. For the lexicographer interested in the maximally precise description of the material, such instances are problematic. The problem is that some dictionaries present these phraseological 'false friends' as full equivalents (cf. Kveselyevich 2002: 350), not taking into consideration that between basically similar idioms in a source language and in a target language, there are practically always certain semantic, pragmatic, and syntactic differences. Our goal is to discover and describe these linguistic-specific differences in English and Russian idioms.

Apart from its theoretic relevance as an instrument for describing idioms of the English and Russian languages, a new dictionary could be used for the purposes of translation and language acquisition.

The paper is structured as follows: after an introduction, the author presents the methodology, theoretical framework and data used in the research. Next, the article gives an overview of results of the study, followed by a discussion. The paper is then summed up in the conclusion section.

## 2 Methodology, Theoretical Framework and Data

The purpose of this project is to explore English and Russian idioms along two lines of inquiry. The first introduces a methodology of generating a comprehensive idiom list for a new dictionary. In order to achieve the purpose of the study, the following steps were undertaken:

- analyzing and retrieving idioms from bilingual and monolingual dictionaries of idioms;
- determining the degree of frequency of selected idioms;
- registering varied forms of the same or similar idioms; and
- determining the keywords for the dictionary entries.

The second line of inquiry considers linguistic issues presenting difficulties in bilingual lexicography related to the concept of equivalence in idioms, and the author analyzes semantic asymmetry between English and Russian idioms. To deal with the second line, the author applies the theoretical concept introduced by D. Dobrovol'skij in his *Studien zur Deutschen Lexik* with respect to cross-linguistic correspondence of idioms: "What is important for cross-linguistic correspondence... is not 'phraseologicalness', but functional equivalence. It is this type of equivalence that is most interesting from the perspective of bilingual lexicography" (2013: 212). The author cannot but agree with this approach, since functional equivalents are parallels which can be used in similar situations "without any information loss" (2013: 212).

The idiom list contains units from *English-Russian Phraseological Dictionary* (Kunin 1984), *Cambridge International Dictionary of Idioms* (1998), *Collins Cobuild Dictionary of Idioms* (Sinclair 1995), *A Dictionary of American Idioms* (Makkai 2004), *NTC's American Idioms Dictionary* (Spears 2000), *Thesaurus of Present-day Russian Idioms* (Baranov & Dobrovol'skij 2007) and the English-language Internet. Presently the idiom list contains about 1,400 idioms and their variants, with the prospect of expanding up to 3,500 idioms. This list forms the baseline for the current study. The present work involves combining variant forms and determining the keywords for the dictionary entries.

The authentic data were collected from the corpus query system Sketch Engine, using the subcorpus enTenTen13 (19.7 billion tokens), the ruTenTen11 subcorpus (14.5 billion tokens), the English-Russian parallel subcorpus OPUS-2, the Russian National Corpus (RNC) and its parallel subcorpora. This made it possible to find instances of English idioms and their Russian equivalents and obtain statistically representative data.

During the first stage of work on organizing the idiom list, 2,500 idioms were picked out from the abovementioned dictionaries, and their frequency was checked in enTenTen13 and the RNC. About 50% of them were excluded because they are seldom, if ever, used today, according to empirical material retrieved from corpora. For instance, there is not a single example of the idioms *to wear a brick in one's hat*, *as fat as an alderman*, *to wear the queen's coat*, *to run like a lamp-lighter*, *stiffen the lizards!* and many other archaic idioms in Sketch Engine.

These examples mainly come from the *English-Russian Phraseological Dictionary* (Kunin 1984). It has been the only comprehensive English-Russian phraseological dictionary since its first publication in 1955 in the Soviet Union to the present day. There are no dictionaries equal to it in the field of English-Russian phraseography. It contains 20,000 idioms, and many generations of students and scholars were 'brought up' on it. Many smaller English-Russian phraseological dictionaries have been published in recent years (Karpova 2004; Shitova & Bruskina 2012; Solodushkina 2016; Vinokurov 2016). However, they have the following drawbacks: either they do not have illustrative examples at all, or if they have them, their examples are self-made, often arbitrary and unpersuasive. Besides, in many cases their vocabulary is out of date; they do not consider polysemous idioms and variants of idioms in dictionary entries; and, as a rule, there is no information about the pragmatic and/or syntactic properties of idioms. Our goal is thus to eliminate the drawbacks of the existing dictionaries, to expand the idiom list with new present-day idioms and illustrate them with authentic data from corpora.

### 3 Results and Discussion

The main important difference between the present dictionary and the traditional dictionaries is in the present one's orientation towards authentic modern data, illustrating the idiom usage drawn from the text corpora and, in some cases, from the English-language Internet. Only those idioms which were found in academic works, spoken usage, fiction, newspaper and magazine texts from the 1950s up to

the present were included in the dictionary. The paper discusses idioms in a variety of contexts, using large text corpora with the aim of clarifying the semantic properties of idioms, providing a new vision of contextual behaviour of idioms and restrictions of their usage.

### 3.1 Cross-linguistic Equivalence of Idioms in the Dictionary

This article analyzes the non-equivalence of English and Russian idioms that are characterized by relative similarity of outer structure, but with significant differences in their actual meaning. The author presents the English idiom *to be born with a silver spoon in one's mouth* (4) and its Russian pseudo-equivalents *родиться в сорочке* (5) and *родиться под счастливой звездой* (6) in the dictionary. This article argues that these idioms are semantically asymmetrical and illustrates their actual functioning in authentic texts in corpora.

In the majority of English-Russian dictionaries (Apresyan 1994; Kunin 1984; Vinokurov 2016; Solodushkina 2016; Shitova & Bruskina 2015) and Russian-English phraseological dictionaries (Kuzmin 2001; Kveselyevich 2002) the idioms (4) and (5), (6) are presented as full equivalents. However, empirical data obtained from corpora suggests that these idioms are semantically asymmetrical. The imprecise treatment of idiomatic meaning in dictionaries happens often, due to their associative similarity in the source and target languages. For this reason, it is important to keep in mind that some figurative language units can have different meanings in spite of their formal similarity. Let us take these examples:

- (4) *to be born with a silver spoon in one's mouth*  
 'to have wealthy parents; be born into a rich family <referring to a child of rich parents who is fed with a silver spoon> (Longman 1979: 310)
- (4) *родиться в рубашке (сорочке)*  
 "to be born with a caul" 'about a person who is blessed with good luck'
- (6) *родиться под счастливой звездой*  
 "to be born under a lucky star" 'extraordinary lucky and successful' (Makkai 2004: 38); 'it is understood that a person succeeds in life easily, he is distinguished by the ability to avoid and get out of serious dangerous and difficult situations' (Teliya 2006: 586).

The analysis of dictionary definitions shows that like the English idiom (4), the Russian idioms (5) and (6) relate to the same idea of being lucky. However, this is the *only* similarity between them. The inner form of each of these idioms contains a culture-specific component, definite knowledge that motivates the actual meaning of the idiom. This knowledge places combinatorial and contextual restrictions on idiom usage and specifies the concept 'luck' differently in relation to English and Russian idioms. The idiom *to be born with a silver spoon in one's mouth* refers to material wealth; that can be proved by empirical data retrieved from the search system Sketch Engine enTenTen13.

- (7) A rich, affluent family's only son, Rahul *was born with a silver spoon in mouth*. Growing up with tremendous amount of disposable money, busy parents and no goal in life, Rahul had no major complaints in life.
- (8) Of course, by "millionaire" we do not refer to those who *are born with a silver spoon* in a royal family or who have grown wealthy overnight by virtue of some propitious inheritance. A good number of millionaires are usually those who have risen by virtue of hard work, following one's dreams and determination.

The hit rate for the idiom (4) amounts to 761 texts in Sketch Engine, they express one and the same idea: 'to be born in a rich family'. The origin of the idiom is treated in the following way: 'In Europe



during the “Dark Ages”, silver utensils, cups and bowls were utilized to aid in protecting the wealthy from the full brunt of pandemics. The expression *born with a silver spoon in their mouth* comes from these “Dark Ages”, when the wealthy gave their children silver spoons to suck on to ward off diseases’ (enTenTen2013).

However, the Russian idioms (5) and (6) do not often express the idea of material wealth. They express either an idea of a lucky escape from imminent danger (5) or an idea of extraordinary luck or quick success in life (6).

The literal translation of the Russian idiom *родиться в рубашке (сорочке)* is “to be born with a caul”. One of the meanings of *рубашка (сорочка)* – ‘caul’ in the Russian language is ‘the inner fetal membrane of higher vertebrates esp. when covering the head at birth’ (NMWD: 129). According to superstition, the appearance of a caul on a newborn baby was seen as a sign of good luck and in medieval times it was viewed as a magical symbol of protection.

The idiom *родиться под счастливой звездой* “to be born under a lucky star” goes back to an ancient mythological form of understanding of the world around us. ‘There is a metaphor underlying the phraseological image component. The metaphor correlates luck and success that accompany a person all his life with his birth *under a lucky star* – under a *star* that is in the middle of the sky at the moment of his birth’ (Teliya 2006: 586-587). It is an ancient idea that this star influences the person’s life and predetermines his destiny.

The hit rate for *родиться в рубашке* amounts to 1185 texts, *родиться в сорочке* – to 156, *родиться под счастливой звездой* to 590 texts. Some examples of Russian idioms (5) and (6) in authentic texts are given below.

- (9) Американец Ламар Лакейз точно *родился в рубашке*. Остаться живым после укусов 1200 пчел – это просто чудо. Даже врачи удивляются, ведь мужчина уже не молодой! Слава Богу, сердце оказалось крепким... (enTenTen11)

An American Lamar Luckeys *was born lucky* for sure. It is a real miracle to be alive after 1,200 bee stings. Even doctors are surprised, for he is not a young man. Thank God, his heart is strong...<sup>1</sup>

- (10) Монахи отправили делегатов в Рим, непосредственно к папе Анастасию четвертому, с жалобой на аббата, обвинив его в чудовищных преступлениях. Туда же явился и сам аббат; вероятно он и впрямь *родился в сорочке*, потому что сразу покори́л папу, и тот не только прогнал монахов, но даже оставил аббата при себе. (enTenTen11)

The monks sent delegates to Rome, to the Pope Anastasius IV directly with a complaint on an abbot, accusing him in monstrous crimes. The abbot himself also was also invited there; he might *have really been born lucky*, because he immediately charmed the Pope who not only sent the monks back, but kept the abbot.

- (11) Александр называет себя счастливым человеком. И это поистине правда, он *родился под счастливой звездой*, потому что у него хорошая работа, есть внуки, отличный коллектив. (enTenTen11)

Alexander calls himself a happy man. And it’s really the truth, he was *born under a lucky star*, because he has a good job, grandchildren, excellent colleagues.

- (12) Дело в том, что я, наверное, *родился под счастливой звездой*; мне очень везет в жизни. У меня были прекрасные родители – добрые, веселые, талантливые. Я был на войне и остался жив. Я с детства хотел стать писателем – стал им... Женщины, которых я любил, любили меня: и о каждой я думаю с нежностью и благодарностью. (enTenTen11)

The fact is that I must have probably *been born under a lucky star*: I am a lucky person. I had

<sup>1</sup> The English translation is provided for the sake of understanding.

wonderful parents – kind cheerful, talented. I was in the war and survived. From childhood I wanted to be a writer, and I have become it become it... The women, whom I loved, loved me too: I think with tenderness and gratitude about each of them.

These texts from corpora clarify that the idioms (5) and (6) have roughly one and the same meaning ‘to be lucky’, but there are still some semantic and pragmatic differences between them. The idioms cannot be substituted for each other in all contexts, because they differ in respect of their individual characteristics. There are pragmatic differences: the idiom (5) *родиться в рубашке (сорочке)* ‘to be born with a caul’ belongs to a colloquial style, whereas the idiom (6) *родиться под счастливой звездой* ‘to be born under a lucky star’ corresponds to a literary style. There are also semantic differences: the idiom (5) is focused on the sense of security, feeling of being protected from strokes of bad destiny and misfortune; the idiom (6) is focused on destiny and on its being successful. All differences of pragmatic, semantic or syntactic differences are reflected in the dictionary entry of the new dictionary.

Again, consider the idiom (4) *to be born with a silver spoon in one’s mouth*. The author argues that the English idiom (4) and the Russian idioms (5) and (6) are different idioms and should be placed in separate entries of the dictionary, unlike the entries of the abovementioned dictionaries. The results of corpus analysis and a detailed study of authentic texts illustrate that idioms (4) (5) and (6) are semantically asymmetrical, and they are not interchangeable.

### 3.2. Idiom Variation in the Dictionary

Fixedness is definitely a vital feature of idioms, and many idioms do not vary at all, but as Moon observed when studying the fixed expressions and idioms included in her database, around 40% of these items have lexical variations or strongly institutionalized transformations, and around 14% have two or more variations on their canonical forms (Moon 1998: 120-121). Idioms can undergo different kinds of variations: lexical, morphological, syntactic and so on. Since the purpose of this study was generating a comprehensive idiom list of the dictionary, the next stage of research was determining the keywords and the lemmas of the dictionary entries. The idioms in the dictionary are organized alphabetically by the keyword. In selection of the keyword, we used the methodology suggested by Kunin: ‘The headword is selected on a purely formal basis, taking into account only the principle of constancy without reference to any grammatical or semantic center the expression may have’ (Kunin 1984: 15).

We will observe an idiomatic group with the meaning: ‘an important person’: *big shot, big cheese, big brass, big wheel, big wig* etc. The majority of idiomatic bilingual and monolingual dictionaries, referred to earlier (see 3.1.), usually treat these idioms as variants of one and the same idiom.

The article examines four English-Russian phraseological dictionaries in respect to organization of their entries, the determination of lemmas and the format of presenting variants in respect to an idiomatic group with the meaning: ‘an important person’:

1. *English-Russian Phraseological Dictionary* (Kunin 1984), 20,000 idioms. The lemma is *big pot*, followed by its variants, arranged in this sequence: *big bug, big gun, big shot, big wig*[sic]; American English: *big cheese, big dog, big fish, big number, big wheel*.
2. *Anglo-Russian Karmanny Slovar’ Idiom* (Vinokurov 2016), 5,500 idioms. The lemma is *big bug*, followed by variants: *big cheese, big gun, big shot, big wheel*.
3. *Dictionary of American Idioms* (Karpova 2004), 1,000 idioms. The lemma is *big gun*, variants: *big shot, big wig, big cheese, big wheel*.
4. *Dictionary of Idioms and Set Phrases* (Solodushkina 2016). The lemma – *big cheese*, no variants.

We referred to the enTenTen13 English corpus to reveal the most frequently used variants of idioms presented in these dictionaries. The frequency graphs were processed manually to avoid information noise.

Table 1: The total amount in the corpus.

Idiom	Quantity	Frequency
big shot	14,448 (0.64)	33.1%
big gun	14,085 (0.62)	32.3%
big wheel	6,352 (0.28)	14.6%
big cheese	3,359 (0.15)	7.8%
big wig	3,244 (0.14)	7.4%
big bug	1,231 (0.05)	2.8%
big brass	859 (0.04)	2%
	43,578	100%

A comparison of the results of the dictionary and corpus analysis in the Table 1 indicates that the lemmas presented in the entries of the four dictionaries are not the most frequent ones. The exception is the idiom *big gun* in Karpova (2004).

The results of the statistical analysis in Table 1 show the total amount of seven most frequently used English idioms with the meaning ‘an important person’ in the enTenTen13 corpus. We need this data to determine the lemma in the dictionary entry. The graph indicates that the most frequently used idiom in this group in contemporary authentic texts is *big shot*, which amounts to 33.1% of idioms with the meaning ‘an important person’ presented in the graph.

Since the idiom *big shot* is the most frequently used idiom in this group, it is selected as the lemma of the dictionary entry. Its variants are *big cheese*, *big wig*, and *big bug*. We regard them as equivalents because semantic differences were not revealed and they all refer to ‘a person of consequence’.

The idiom *big gun* is the next most frequently used idiom in the list. Almost all the analyzed dictionaries rank the idiom *big gun* together with the idioms *big shot*, *big cheese*, *big wig*. But evidence suggests that the idiom *big gun* refers not only to people, but to a broad variety of notions, such as huge companies, corporations, online shopping services, internet providers and so on. This places combinatorial restrictions on the usage of the idiom *big shot* and its variants. Consequently, the idiom *big gun* does not fit in a single entry with the idiom *big shot* in our dictionary. *Big wheel*, as well as *big gun*, does not only refer to people, so it qualifies as a variant of the idiom *big gun*. Cf.:

- (13) Plus of course you can examine at the *big guns* in shopping on the internet Amazon. (enTenTen13)
- (14) Can a national *big gun* deliver highly specialized service that is finely tuned and focused on true art over the generalized consistency necessary to maintain its market share? (enTenTen13)
- (15) Ever since Prada made black nylon chic in the early 1990s, Ugg boots, handbags have been fashion’s *big wheel*, Ugg boots churning out profits and profit margins exceeding 40 percent. (enTenTen13)

The idiom *big brass* also belongs to a different entry, because it usually refers to people having important high level military ranks, for example:

- (16) But there will be resistance: a coup such as the one Gore wants to pull off requires the unconditional and united support not only of the political establishment, but also of the military: and

I am not talking about the *big brass*, the boys in the Pentagon, or even the mid-level officers... (enTenTen13).

As a result, we divided the idiomatic group with the meaning ‘an important person’ into three groups, arranging them into three entries with the lemmas *big shot*, *big gun* and *big brass*. There are comments about semantic and combinatorial properties of the idioms in the entries. Such comments are very important for understanding of semantic, pragmatic, and other properties of idioms, and consequently their adequate translation into the Russian language.

Using the corpus query system Sketch Engine and its English-Russian and Russian-English parallel subcorpora OPUS-2, and the Russian National Corpus, we compared idioms from the idiomatic group ‘an important person’ in relation to the ways of their translation into the Russian language. The results of statistical analysis are presented in Table 2.

Table 2: Results of the parallel corpora analysis.

big shot Russian equivalents	Quantity	Frequency	большая шишка English equivalents	Quantity	Frequency
большая шишка “big cone”	23	31%	big shot	19	26%
большая рыба “big fish”	1	1.3%	big deal	3	4.1%
большой туз “big ace”	1	1.3%	big guy	2	2.7%
			big man	2	2.7%
zero equivalent	49	66.4%	zero equivalent	47	64.3%
	74	100%		73	100%

The results of statistical analysis in Table 2 show 74 Russian correlates of the idiom *big shot*: *большая шишка* (23), *большая рыба* (1), *большой туз* (1), “0” equivalent <49>. Table 2 also presents 73 English correlates of the Russian idiom *большая шишка*: *big shot* <19>, *big deal* <3>, *big guy* <2>, *big man* <2>, zero equivalent <49>.

The evidence in Table 2 suggests that the most frequently used English idiom with the meaning ‘an important person’ in the corpora is *big shot*. Its most frequent correlate is the Russian idiom *большая шишка*.

The idioms *big shot*, *big cheese*, *big brass*, *big wheel*, *big wig*, and so on are identical with regard to syntactic properties and partially with regard to lexical structure. In some cases, their meanings are identical as well. However, the purpose of the study is to discover the unique properties of each idiom, their semantic and pragmatic properties. And it is here that the corpora help a lexicographer to determine the form of the lemma and its variants, and to determine the proper equivalent of the lemma in the target language. The author interprets the idiomatic group as parallel idioms rather than variations of a single idiom, since most of them have their specific semantic features that differentiate them from the other idioms in the group.

Sample entries of an idiomatic group with the meaning ‘an important person’ in two recently published English-Russian phraseological dictionaries are given below. The empirical data of both dictionaries are not based on corpus data.

Table 3: The entry from the dictionary [Shitova &amp; Bruskina (2012: 6)]

a big cheese/shot (*Am inf*)  
 важная шишка; большая птица  
*Today we're being visited by some big shots from the head office.*  
*She loved being the big cheese of her company.*

Table 4: The entry from the dictionary [Vinokurov (2016: 33)]

(a) Big bug (cheese, gun, shot, wheel)  
 разг. важная особа, 'шишка'  
*Bill had been a big shot in high school.*

A sample entry of the idiom *a big shot* from the New English-Russian Dictionary of Idioms is given below and contrasted with those in Table 3 and Table 4:

## Big

*big shot (big cheese, big wig)*

важная особа, большая (важная) шишка, большой туз

📖 Should the law give special treatment to a VIP or *big shot* while denying that to everyone else? (enTenTen13)

Неужели правительство будут оказывать особые услуги в медицинском обслуживании VIP-персонам или *большим шишкам*, а простым людям будет отказывать?

Everyone in business knows the importance of a professional image. Of course, professionalism in itself is important; but the image of being a corporate *big shot* has the tendency to make people take you more seriously. (enTenTen13)

Все знают о важности профессионального имиджа в бизнесе. Конечно, профессионализм сам по себе очень важен; но в имидже корпоративного туза Вас гораздо быстрее воспримут более серьезно.

If there was only one dog in the home before, then the new dog will quickly learn that the older dog is the *big cheese* around here. (enTenTen13)

Если в доме раньше была только одна собака, тогда новая собака быстро поймет, 'кто в доме хозяин'.

He saw Mr. Pogue in the midst of a bunch of oil company *big wigs*. They were heads of Texaco, Shell, Esso, etc... (enTenTen13)

Он увидел мистера Пога среди группы воротил нефтяного бизнеса. Это были хозяева таких корпораций, как Техасо, Шелл, Эссо и др.

📖 Именной компонент в составе идиомы может употребляться во множественном числе.



☞ CAM is scientifically proven to be helpful, beneficial and curative. The ‘traditional medicine’ *big shots* just lie about it and say that there isn’t any evidence supporting it. Big Lie! (enTenTen13)

Научно доказано, что нетрадиционная медицина целесообразна и целительна. Однако шарлатаны от ‘традиционной медицины’ просто лгут, утверждая, что нет доказательств о ее целебных свойствах. Это ужасная ложь!

☞ Возможна атрибутивная модификация.

☞ I sent emails to the cruise line but I never heard back. They are aware of what goes on there. However, because they have *big shot* lawyers they can afford to turn their heads and laugh it away. (enTenTen13)

Я послал уведомления в круизные компании, но ответа от них не получил. Они прекрасно понимают, о чем идет речь. Но так на них них работают видные юристы, компании могут позволить себе отмахнуться от нас и просто посмеяться.

☞ Выступая в функции атрибута, идиома *big shot* может характеризовать не только людей, но и солидные фирмы, крупные корпорации.

☞ We hear about social media all the time. It seems like everyone uses it in some way, shape or form from teenagers, to coffee shops, to *big shot* companies. (enTenTen13)

Мы все время слышим о социальных сетях. Такое впечатление, что в какой-то мере все пользуются ими, начиная от подростков и кофеен до крупных компаний.

☞ Идиома часто может иметь оттенок презрения или недоверия, свидетельствующих о недостойных методах достижения занимаемого положения.

☞ They knew he was a liar the first time, but in the appeal they believed that rotten lie and now Joe is a *big shot* and your father is the patsy. (enTenTen13)

## 4 Conclusion

A thorough analysis of authentic texts made it possible to discover the unique semantic properties of each idiom and find proper Russian equivalents. The author argues that not all of these idioms should be placed in one dictionary entry. However, if they are placed in one entry then there should be detailed explanations of their semantic, pragmatic or syntactic differences.

The results of the study demonstrate obvious advantages of using corpora in bilingual lexicography. Applying corpora in constructing the English-Russian dictionary of idioms provides the following possibilities:

- to single out a more precise and accurate choice of idioms to be included in the dictionary;
- to determine the most significant variants of a particular idiom;
- to determine the degree of frequency of the selected idioms;
- to find the most suitable Russian equivalents of the English idioms; and
- to illustrate the idiom usage with modern authentic data and provide an appropriate translation of concrete examples.

Many idiomatic dictionaries fail to register all meanings of idioms, sometimes introducing their insufficient semantic description and usage, because in some cases reference materials on idioms are based on the intuition of their authors, and the intuition of even the most competent and experienced

authors is not always consistent. This is why lexicographers have to develop better ways of recording, describing and presenting idioms in dictionaries.

## References

- Apresyan, Yu., et al. (1993). *New English-Russian Dictionary* (in three volumes). Moscow: Russky Yazyk Publishers.
- Baranov, A., Dobrovol'skiy D. (2007). *Thesaurus of Present-day Russian Idioms*. Moscow: Avanta.
- Birikh, A., Mokienko V. & Stepanova L. (2005): *Russian Phraseology: historical etymological dictionary*. Moscow: Astrel AST.
- Cambridge International Dictionary of Idioms* (1998). Cambridge: Cambridge University Press.
- Collins Cobuild Dictionary of Idioms* (1995). London: Harper Collins Publishers.
- Dobrovol'skiy, D. (2013). German-Russian idioms online: on a new corpus-based dictionary. In *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International "Dialogue" Conference* (Bekasovo, May 29-June 2, 2013). Issue 12 (19) 210-217. Moscow: RGGU.
- Karpova, M. (2004). *Dictionary of American Idioms*. Moscow: Astrel AST.
- Kveselevich, D. (2002). *Sovremennyy russko-anglijskiy frazeologicheskiy slovar*. Moscow: Astrel.
- Kunin, A. (1984). *English-Russian Phraseological Dictionary*. Moscow: Russky Yazyk.
- Kuzmin, S. (2001). *Translators' Russian-English Phraseological Dictionary*. Moscow: Flinta Nauka.
- Longman Dictionary of English Idioms* (1979). Harlow and London: Longman Group Limited.
- Lubenskaja, S. (2004). *Bol'shoj russko-anglijskiy frazeologicheskiy slovar'*. M.: Ast-PressKniga.
- Makkai, A. (2004). *A Dictionary of American Idioms*. NY: Barron's.
- Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.
- NMWD – The New Merriam-Webster Dictionary. Springfield, Ma: Merriam-Webster Inc., Publishers.
- Piirainen, E. (1997). ‚Da kann man nur die Hände in den Schloß legen‘. Zur Problematik der Falschen Freunde in niederländischen und Deutschen Phraseologismen. In: *Nominationsforschung im Deutschen. Festschrift für Wolfgang Fleischer zum 75 Geburtstag*. Frankfurt/M: Peter Lang.
- Shitova T. & Bruskina T. (2012). *English Idioms and Phrasal Verbs*. St. Petersburg. Anthology.
- Solodushkina, K. (2016). *Dictionary of idioms and set expressions*. Moscow: Infra M.
- Spears, R. (2000). *NTC's American Idioms Dictionary*. NY: NTC Publishing Group.
- Teliya, V. (2006). *Bol'shoj Frazeologicheskiy Slovar' Russkogo Yazyka*. Moskva: AST Press.
- Vinokurov A. (2016). *Anglo-Russky Karmanny Slovar' Idiom*. Moskva: Martin.



# Computer-aided Analysis of Idiom Modifications in German

**Elena Krotova**

*Institute of Linguistics, Russian Academy of Sciences*

*E-mail: elena\_krotova@inbox.ru*

## Abstract

This paper deals with corpus approaches to the study of modifications of idiomatic expressions in German. It concentrates on one group of phraseological units, idioms. In spite of a high degree of stability, idioms still undergo different modifications. To get reliable results about idiom modifications, a large number of modified target structures is crucial. Therefore, a Python-program has been created that obtains information about the usage of idioms and about their possible modifications from a corpus. It also summarizes the data in the form of graphs. The report will look further into the program's opportunities to acquire information about idiom usage, how idiom modifications correspond to the syntactic behavior of their paraphrases or free phrases containing the same verb as the idiom under discussion and in what ways such data can facilitate the work of a phraseologist.

**Keywords:** corpus linguistics, phraseography, modifications of idiomatic expressions

## 1 Introduction

The compilation of dictionary entries for phraseological units, especially for idiomatic expressions, is difficult for many reasons.<sup>1</sup> Idioms have a complex semantic structure and a number of syntactic peculiarities that affect their usage in speech. Only with a detailed description of idioms' semantics and syntax can non-native speakers learn to use phraseological units in an appropriate, native-like way. Before large machine-readable corpora appeared, makers of phraseological dictionaries had to rely largely on their own language intuition, whereas now it is possible to verify information using corpora. For frequent idioms we can find thousands of occurrences in large corpora. This amount of material is sufficient to describe in detail the behavior of idioms in written language. The problem however is that the analysis of such an amount of data takes a lot of time, especially if it is not a detailed study of one particular idiom, but the compilation of hundreds of articles for a phraseological dictionary. This time could be reduced through the use of automatic methods that analyze corpora data, search for different idiom modifications and summarize the obtained information. In the following, we will discuss the development of such a program for frequent idioms of the German language and present its first results.

## 2 Idiom Modifications

Idioms possess a high degree of stability, but they still undergo different modifications. In Figure 1 you can see two graphs<sup>2</sup> with modifications of German idioms:

- 1 I use the term idiomatic expressions, or idioms, to refer to "phrasemes with a high degree of idiomaticity and stability. In other words, idioms must be fixed in their lexical structure (however, this does not exclude a certain variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly but via a source concept) and/or semantically opaque" (Dobrovol'skij 2006; for detailed explanation of the term *idiomaticity* see Baranov & Dobrovol'skij 2008: 50).
- 2 I use the term idiomatic expressions, or idioms, to refer to "phrasemes with a high degree of idiomaticity and stability. In other words, idioms must be fixed in their lexical structure (however, this does not exclude a certain variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly but via a source concept) and/or semantically opaque" (Dobrovol'skij 2006; for detailed explanation of the term *idiomaticity* see Baranov & Dobrovol'skij 2008: 50).

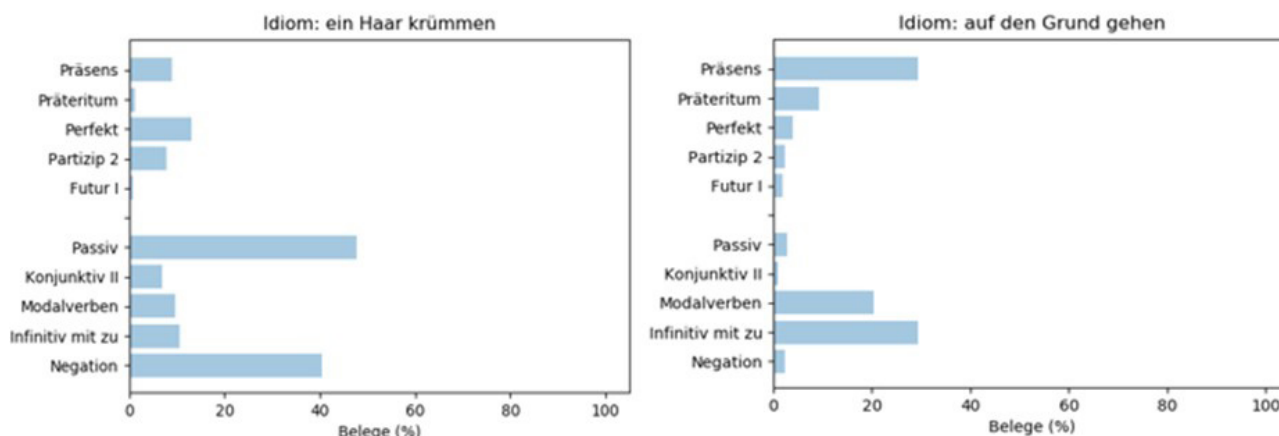


Figure 1: Idiomatic modifications *ein Haar krümmen jmdm.*, *auf den Grund gehen etw.*

The first idiom *ein Haar krümmen jmdm.* ‘to lay a finger on sb.’ is mostly used in the passive voice and with negation, while for the second idiom *auf den Grund gehen* ‘to drill down on sth.’ these modifications are not frequent. The second idiom is mostly used in infinitive constructions, in present and with modal verbs.

We distinguish between the following kinds of modifications: morphological, lexical, lexical-syntactic and syntactic (see Dobrovol’skij 2008: 308-309). Here are some examples from German:

- Morphological modifications include article variation (*auf dem [einem] absteigenden Ast sein* ‘to be on the downgrade’).
- Lexical modifications include changes in component structure (omission or addition of an element, its substitution by another word). Example: *von allen guten Geistern verlassen sein* ‘to have taken leave of one’s senses’ and its lexical variants *von guten Geistern verlassen sein*, *von allen Geistern verlassen sein*, seldom *wie von allen guten Geistern verlassen sein*, seldom *von allen guten Göttern verlassen sein*.

Lexical modifications are complicated to detect, because a researcher does not always know about all the lexical changes an idiom can undergo<sup>3</sup>.

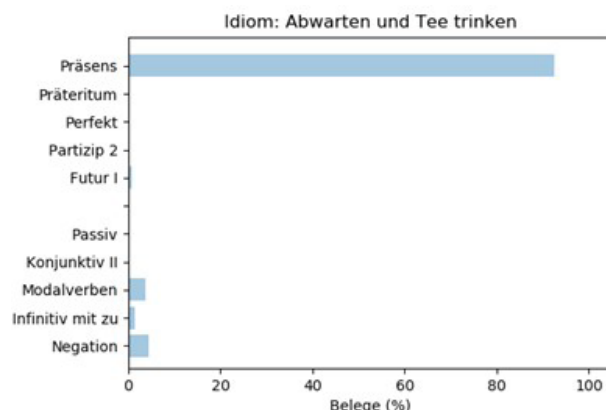
- The lexical-syntactic type covers modifications which affect both syntactic and lexical idiom structure (e.g. putting an adjective modifier between an article and a noun). It can refer to the adverbial (e.g. *sein Herz ausschütten jmdm.* ‘to open one’s heart’ modified by *so richtig* ‘really’: *so richtig das Herz ausschütten*), the attributive (e.g. *einen Haken haben: (etw.) hat einen Haken* ‘There is a hitch somewhere’ modified by *klein* ‘small’: *einen kleinen Haken haben*) and to the so-called metalinguistic type (e.g. *sich fühlen wie ein [der] Fisch im Wasser* ‘to be in one’s element’ modified by *viel zitiert* ‘much-cited’: *wie der viel zitierte Fisch*).
- Syntactic modifications contain the use of idioms with negation, in passive voice, questions, infinitive constructions and so on.

This study deals mostly with syntactic modifications as well as tenses the verbal component of the idiom is used in. The extracted data is presented on the graphs. Other types of modifications are analyzed only to some extent (see Section 4).

There are idioms that hardly ever show modification, such as *abwarten und Tee trinken* ‘to wait and see’:

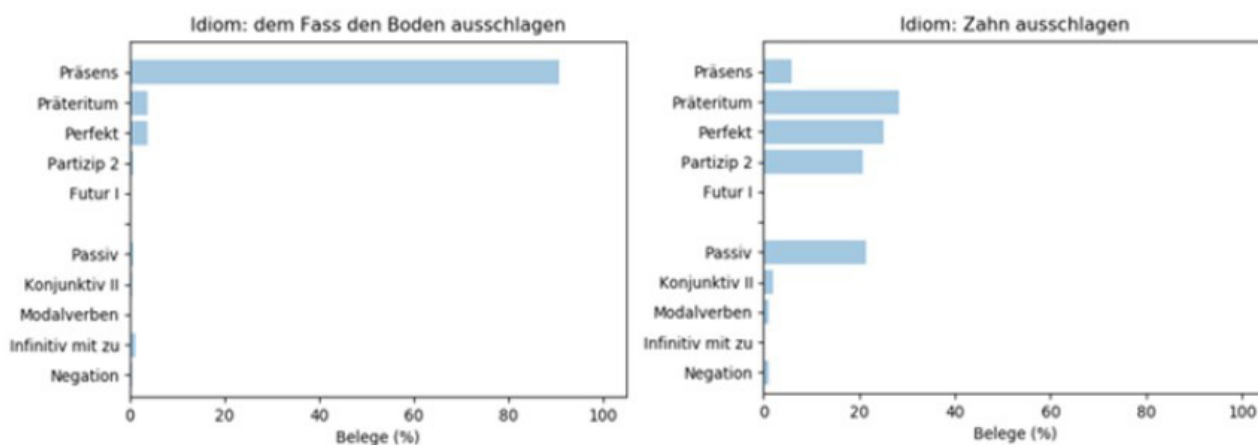
<sup>3</sup> See Section 4 for further elaboration on the graphs.



Figure 2: Idiom *Abwarten und Tee trinken*

However, it is not possible to say that this idiom has an absolutely fixed form. There are occurrences where it is used with negation, with modal verbs and in infinitive constructions. The verbal component of the idiom *trinken* is used in 91.3% of texts in this particular word form, in 3% with a modal verb and in 2.4% in infinitive constructions with *zu*.

This case is rather rare. Among the 100 idioms analyzed, such a small number of changes has been found only for the idiom *dem Fass den Boden ausschlagen*: *Das schlägt dem Fass den Boden aus* ‘That’s outrageous’. This idiom, unlike the previous one, can be used in other tenses, although such occurrences are rare, and the verb can change its form in the present (though the word forms *schlägt aus* and *ausschlägt* comprise up to 86% of occurrences). To the right of the idiom there is a graph for a free phrase containing the verb that is a part of the idiom:

Figure 3: Comparison of the idiom *dem Fass den Boden ausschlagen* and the free phrase *jmdm. einen Zahn ausschlagen*

As you can see, although the idiom is seldom used in the passive, its verbal component in other free phrases is used in the passive quite frequently. The structure of the idiom on its own does not restrict such a modification, e.g. the sentence *Dem Fass wurde der Boden ausgeschlagen* would be correct from a purely grammatical point of view. However, such a modification of the idiom hardly ever occurs in speech. Moreover the free phrase *den Zahn ausschlagen* seldom occurs in the present, which can be connected with pragmatics: in the prototypical situation the phrase describes a result of a physical action, so the usage in past tenses is more probable.

Therefore, a lexicographer should provide language learners with information about the modifications an idiom undergoes, because this information cannot be inferred from the syntactic behavior of the idiom's components in free phrases.

### 3 Material

To get reliable results about idiom modifications a researcher needs to acquire and analyze as many text examples as possible. It is clear that the biggest corpora are collections of texts from the Internet that can be searched by a search engine. However, this approach has its limitations, because a search engine is focused on information retrieval and not on the extraction of linguistic information. That is why linguistically annotated text corpora have been chosen for the study of idiom modifications. The biggest corpus of the German language is *Deutsches Referenzkorpus* (DeReKo), which contains more than 42 billion tokens (03.02.2018), but it is unbalanced and consists mainly of newspaper texts. This has obvious disadvantages: 1) not all idioms are used in written speech and particularly in newspaper texts, even if they are frequent in spoken language, 2) we often have wordplay with idioms in newspapers that does not always reflect their usage in speech. However, the fact that the corpus is large and we can find hundreds of each phraseme's occurrences in modern texts outweighs this drawback.

In order to have more homogenous text material a virtual corpus has been created for research purposes. The corpus contains only newspaper articles published in Germany after 1980 and excludes texts published in other German-speaking countries for the following reasons: Austrian and Swiss texts can contain some idiom modifications which can be typical only for the particular national variant of German but not for other variants; since the modern usage of idioms is investigated, only texts published since 1980 are considered; the corpora contain very few fiction texts published since 1980 (less than 0.01% of the whole text archive), so they are excluded to make it clear that the analysis is based only on newspaper texts.

At the starting stage of the project, 100 widely used German idioms were chosen. Most of them have the structure verb + preposition + noun. The surveyed phraseological units come from the dictionary of frequent German idioms (Dobrovolskij 1997); their frequency was checked against DeReKo. The information given in the dictionary about possible lexical, morphological and lexical-syntactic modifications of the idioms was obtained through the analysis of different lexicographic resources and text corpora.

### 4 Methodology

DeReKo queries have been formulated manually for all of the selected idioms. The queries are made in such a way that possible modifications are not excluded from the results<sup>4</sup>. E.g. the idiom *auf den Grund gehen* 'to get to the bottom of sth.' During the work on the dictionary of frequent German idioms it turned out that the idiom under discussion can be used in the passive form, but it does not seem

4 Lexical modifications can be explored in different ways. Besides consulting different lexicographic resources a researcher can omit some components of an idiom and try to find out what other words can fill the gap. In DeReKo one can use a tool for co-occurrence analysis (*Ko-Occurrenzanalyse*). E.g. when we search for co-occurrences for the phrase *auf den Grund*, we see that the most statistically significant co-occurrence partner will be *gehen*. But if we search for co-occurrences for the verb *gehen*, we won't find the phrase *auf den Grund*, may be because of the very high frequency of the verb. Another example for the idiom *von allen guten Geistern verlassen sein*: if we search for *von allen guten Geistern* the most significant co-occurrence partner will be *verlassen*. If we search for *von allen guten* and *verlassen*, we will find the lexical modification *von allen guten Fußballgeistern verlassen sein*. It is not frequent (only ten occurrences from more than 1,800 sentences containing the idiom), but statistically significant, because this word occurs very seldom, yet in about 30% of sentences together with *verlassen*.

to have any frequent morphological or lexical modifications. As a result, the query looks as follows: (*auf* /+w3,s0 *Grund*) /s0 & *gehen*. It means that *gehen* can be used in any form (operator &) and have any position in a sentence (operator /s0 means that the verb must be in the same sentence as the prepositional group). Besides, a maximum of two tokens can appear between *auf* and *Grund* (operator w), *auf* is followed by *Grund* (operator +), and both tokens must occur within the same sentence (operator /s0). The paragraphs where the idioms occur can be exported. DeReKo doesn't allow you to export over 10,000 examples, but this number is enough for studying the usage of idioms in written texts.

The queries have not been automated for the following reasons: in every case the researcher should decide how to formulate queries so that the data obtained contain a minimum amount of sentences where the idiom's components are used literally and not idiomatically, but at the same time do not exclude possible modifications. For instance, the idiom *sich (D) ins Fäustchen lachen* 'to laugh in one's sleeve' contains the noun *Fäustchen*, which is not that frequent. That is why it will be enough if we just search for sentences where both components occur together. The situation is different for the idiom *mit den Wölfen heulen* 'to do in Rome as the Romans do'. If we search for the lemmas<sup>5</sup> *Wolf* 'wolf' and *heulen* 'to howl' occurring in the same sentence, we will find a lot of targets where there is no idiom. Therefore, the query should be restricted. For example, (*mit* /+w3,s0 *Wölfen*) /s0 & *heulen*, which means that the lemma *heulen* occurs in one sentence with word forms *mit* and *Wölfen*, the maximal distance between them makes two words.

The specially designed program is used to analyze the extracted data. The obtained data is provided at [bitbucket.org](http://bitbucket.org) (Deutsche Idiomatik). It targets the following information:

- tenses and word forms the verb of the idiom is used in (*Präsens*, *Präteritum*, *Perfekt*, *Futur I*)<sup>6</sup>
- syntactic modifications the idiom undergoes, such as usage in passive voice; in *Konjunktiv II*; accompanied by modal verbs and in infinitive constructions. Verbs (*werden* in passive voice, verbs in conjunctive mood, modal verbs) can be used in any tense. Such contexts do not overlap with the first group. For instance, if there is an idiom in the passive voice present tense, it will only be counted as an idiom in passive voice. The program also searches the contexts with negation *nicht* or *kein*.

Here is an example: 805 contexts have been found for the idiom *jmdm. ein Armutszeugnis ausstellen* 'to show sb.'s incapacity / incompetence / shortcomings', among them 56% in *Präsens*, 14.66% in *Perfekt*, 6.58% with *Partizip II* (auxiliary verbs *sein* or *haben* have not been found), 1.61% in passive voice, 4.98% in *Konjunktiv II*, 4.59% with modal verbs, 1.5% in infinitive constructions. This totals 99.72%. In addition, future tense *Futur I* takes up about 0.3%.

The most frequent verb form is *stellt aus* (28.9%). The idiom seldom occurs with negation (4.59%). The most frequent modal verb is *können* (64% of all occurrences with modal verbs).

Moreover, the tokens preceding the nouns in the nominal component are analyzed. E.g. the noun *Armutszeugnis* is preceded by the article *ein* in 75% of all examples. Other variants are as follows: definite article *das* (1.49%), lexical and syntactic modifications: such words as *solches*, *dieses* (about 0.8% each) and such adjectives as *politisches* (1.4%), *größeres*, *großes* (0.8% each), *geistiges*, *eigenes* (0.6% each). By these means the researcher acquires information about possible morphological (*ein* or *das*), lexical and syntactic modifications (*solches*, *dieses politisches*, *größeres*, *großes*).

The program application can go further to search for sentences with questions and those where an idiom is used in the first person and with metalinguistic constructions (see metalinguistic type in Section 2).

5 In order not to exclude syntactical or morphological modifications the queries normally don't contain articles before the nominal component and contain all possible verb forms. Information about lexical modifications was obtained through the previous research. If an idiom occurs in several lexical variants, separate searches are made.

6 Lemma is understood as a set of words.

In the end, the program creates files that contain an idiom's occurrences in different tenses and with the modifications mentioned above. For all modifications there are separate text documents, so the researcher can analyze each of them in detail. The program also summarizes the data in the form of graphs.

## 5 Results

A total of 100 German idioms have been analyzed. The number of idioms is not very high now, but still the attempt has been made to find idioms with similar modification profiles. The following results have been obtained:

- Almost a third of idioms (related lists and graphs can be found in the project folder) are used in more than 15% of occurrences with modal verbs.
- 10% of idioms are used in infinitive constructions in more than 20% of texts.
- 5% of idioms are used with negation in over 25% of examples.
- 15% of idioms appear in present tense in more than 60% of occurrences.
- 6% of idioms are used in *Präteritum* more often than in *Präsens* and *Perfekt*.

The groups were established manually because the number of analyzed idioms is rather small at the moment, and as this number increases in the future appropriate statistics will be obtained.

Let us now take a closer look at the idioms used mostly in *Präsens*. In addition to the idiom mentioned in the introduction *abwarten und Tee trinken* 'just wait and see', this group includes several idioms with similar semantics ('to get on sb.'s nerves'):

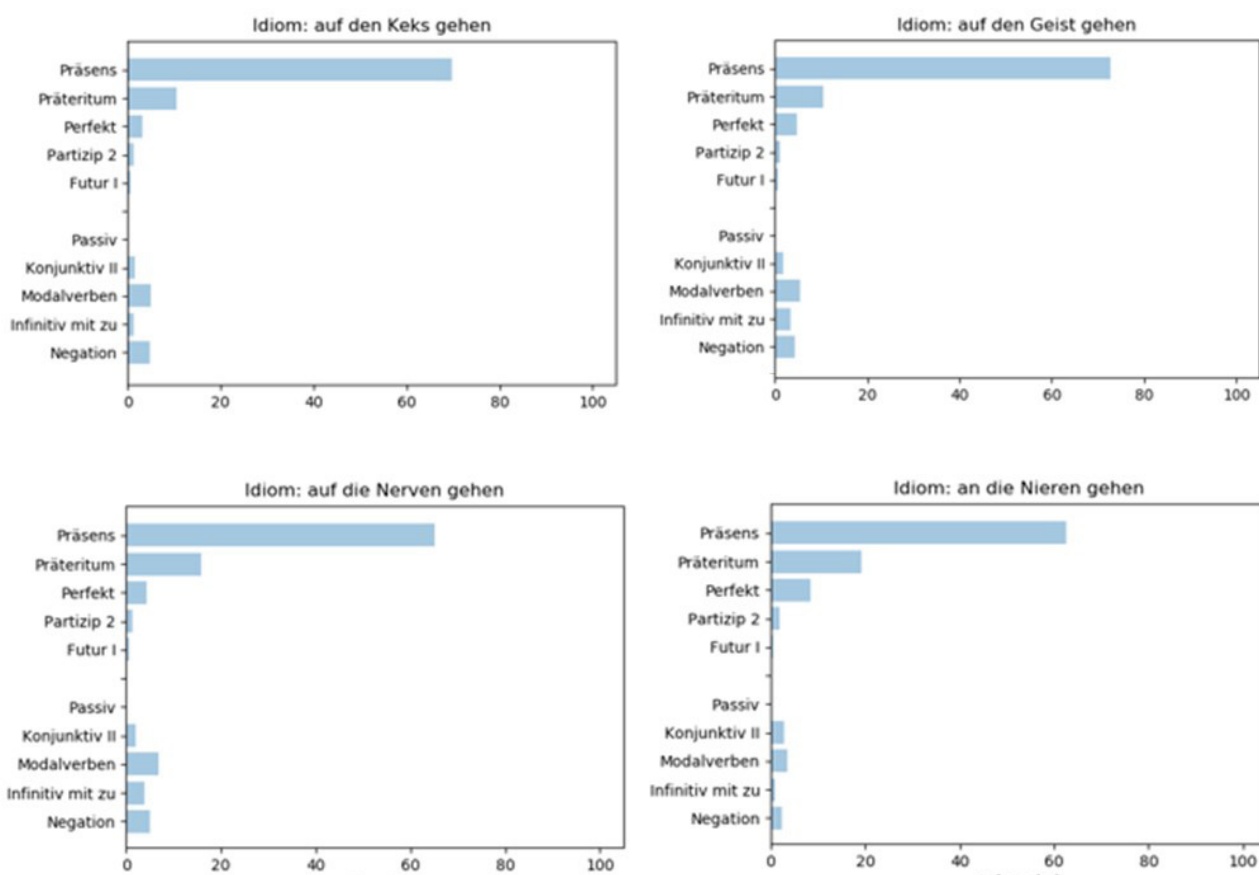


Figure 4: Comparison of graphs for *auf den Keks gehen*, *auf den Geist gehen*, *auf die Nerven gehen*, *an die Nieren gehen*

The group includes three idioms *auf den Keks gehen*, *auf den Geist gehen*, *auf die Nerven gehen* meaning ‘to get on sb.’s nerves’, as well *an die Nieren gehen* ‘to sadden someone’. They are mostly used in *Präsens* and less frequently in *Präteritum*. The usage in *Perfekt* compared to other tenses is rather rare. Though they have similar semantics and the identical verbal component *gehen*, there are still some differences in the frequency of usage with modal verbs, infinitive constructions and *Konjunktiv II*.

Other idioms belonging to the same group have quite different semantics: *auf der Stelle treten* ‘not to make any progress’, *auf Kohlen sitzen* ‘to be like a cat on hot bricks’, *ins eigene Fleisch schneiden* ‘to shoot oneself in the foot’, *dem Fass den Boden ausschlagen: Das schlägt dem Fass den Boden aus* ‘That’s outrageous’, *gegen den Strich gehen* ‘sth. rubs sb. the wrong way’, *Bude einrennen jmdm.* ‘to be overrun’, *ins Fäustchen lachen* ‘to laugh in one’s sleeve’, *aus der Reihe tanzen* ‘to break ranks’, *Armutszeugnis ausstellen* ‘to show sb.’s incapacity / incompetence / shortcomings’, *aus dem Sinn gehen*, *aus dem Kopf gehen: Es geht mir nicht aus dem Sinn [Kopf]* ‘It is always on my mind.’, *im Schilde führen* ‘to scheme sth.’, *vom Hocker reißen* ‘to knock sb.’s socks off’. A lot of them have negative connotations, but this is not unusual for the phraseological system generally, at least for the German language (Raykhshteyn 1980: 61). At the moment it is not clear why these particular idioms mostly occur in *Präsens*, and more idioms should be analyzed to draw more solid conclusions. The hypothesis is that the usage in a particular tense depends more on the pragmatics than on the semantics or the syntactical characteristics of the verbal components.

There is also the question whether graphs such as the ones used here provide the information about the difference between an idiom and free phrases. Let us compare the idiom *auf den Grund gehen* ‘to drill down on sth.’ with its two paraphrases *die Ursache finden*, *den Sachverhalt klären* and a free phrase with the same verb as in the idiom, *ins Bett gehen* ‘to go to bed’.

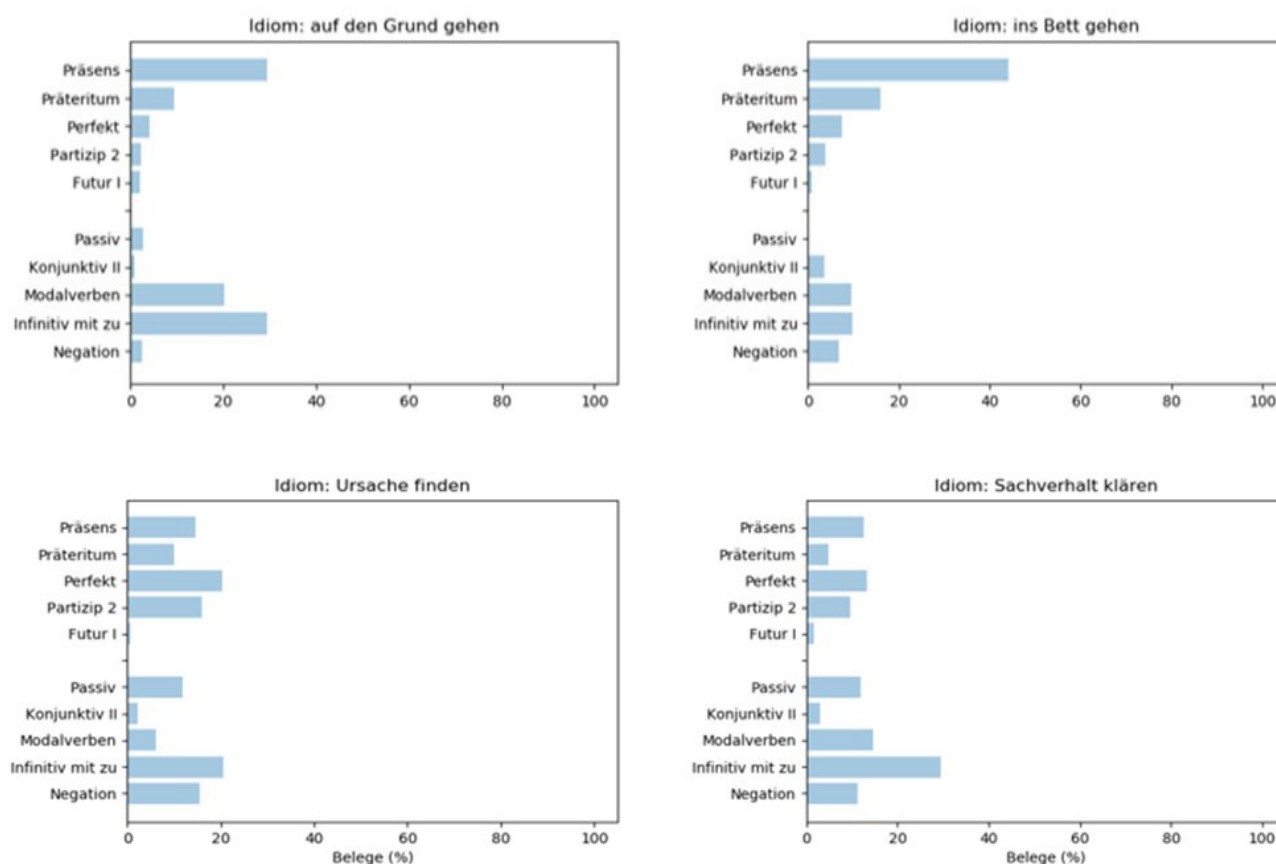


Figure 5: Comparison of graphs for *auf den Grund gehen*, *die Ursache finden*, *den Sachverhalt klären*, *ins Bett gehen*.



Unlike paraphrases and a free phrase containing the verb *gehen*, the idiom is used less often with negation. Compared with the paraphrases the idiom is also less often used in the passive, while the free phrase with the verb *gehen* almost never occurs in the passive, because the verb is normally used in active voice. Unlike paraphrases the idiom is more often used in *Präsens* and less often in *Perfekt*. Thus, the idiom's speech behavior corresponds neither with its paraphrases nor with the syntactic behavior of the verb used in it. On the one hand, the speech behavior of an idiom is determined by its semantics. On the other hand, the syntactic structure imposes its limitations.

## 6 Discussion

### 6.1 Comparison of Sketch Engine and DeReKo

The DeReKo is used in the project as a source of language material and not as a tool to search for all possible modifications. This is for the following reasons: only one DeReKo query should be made, and its processing takes up to several minutes, depending on the frequency of words in the query, the number of word forms and the complexity of the query. To obtain all possible information about modifications the researcher should make many queries and spend proportionally more time. The usage of the developed program for phraseological studies seems to be a more flexible option: it takes less time; the researcher does not have to formulate many queries; and frequency lists for all verb forms and for tokens preceding the nominal component can be easily obtained. Besides, it can be defined what tokens cannot appear between the verb and the prepositional phrase (e.g. commas for some modifications). This makes sense particularly for German, with its long sentences. Otherwise, the researcher gets a lot of contexts that do not contain the idiom, but only the components it consists of.

The reasons why Sketch Engine is not used are mostly the same. Though the biggest German corpora at Sketch Engine, German Web 2013, is half the size of DeReKo, it seems to be well-suited for studying idiomatic expressions, possibly because of its more colloquial character (e.g. for the idiom *nicht alle Tassen im Schrank haben* 'to not have all one's marbles' there are 1,595 occurrences in DeReKo<sup>7</sup> and 4,620 in German Web<sup>8</sup>). However, it does not allow users to export occurrences, and its query language is more difficult and less flexible than that of DeReKo, particularly if the target structures are phraseological units.

### 6.2 Application in Lexicographic Research

The results of the program can be applied in lexicographic research in the following ways. They can help to:

- write comments on acceptable modifications. For instance, if an idiom is a negative polarity item, we can explain in what type of contexts it can be used without negation;
- select illustrative material. The modifications profiles can help to find examples illustrating the usage of an idiom and do not contain rare modifications;
- specify the form of the dictionary entry, for example by answering the following questions:
  - Should a modal verb be a part of a dictionary entry? If so, which one?
  - Which article should be used in the dictionary entry? Example: *Den Ausschlag geben* 'to tip the balance'. In total, the program has found 8,215 idiom occurrences. Among these, 92% contain the definite article *den*. Other frequent lexical and syntactic modifications have also been found, e.g. *letzten* (1.1%) and *entscheidenden* (0.54%). Other tokens preceding the nominal component are less common, like *einen* (0.23%) and *keinen* (0.19%).

7 The query: Tassen /+w2 Schrank

8 The query: "Tassen" []{0,2} "Schrank" within <s/>

- Should the negation be a part of the dictionary entry?

Example 1: *nicht aus dem Sinn gehen* ‘not to go out of mind’, 86% of occurrences contain the negation *nicht*.

Example 2: *jmdm. kein Haar krümmen* ‘not to lay a finger on sb.’, 47% in the passive voice, 39% with the negation *kein*. Below the context from the DeReKo is provided, where the idiom is used without negation:

(1) Wir hatten noch Respekt vor den Lehrern, den meisten jedenfalls. Selbstverständlich gab es auch Lehrer, die wir nicht mochten – trotzdem hätten wir es nie gewagt, dem Lehrer auch nur ein Haar zu krümmen. (Braunschweiger Zeitung, 02.01.2006)?

## 7 Conclusion

Even if idioms possess comparable structures (verb plus prepositional phrase), they all have their own profiles of modifications. Such profiles cannot be inferred only from the semantics of an idiom, from the syntactical behavior of its verbal component or idioms’ paraphrases. Ideally each idiom in a dictionary should be provided with a detailed description of its usage and modifications, as well as corresponding text examples. However, due to lack of space such dictionary articles would only be possible in electronic resources, but not in a printed dictionary.

To reduce the amount of work needed, a program developed as part of this project analyses and summarizes the acquired corpora data. There is a plan to expand the list of analyzed idioms to several thousands. This can be done after the first results have been thoroughly analyzed, and the program improved if needed.

## References

- Baranov, A.N., Dobrovol’skij, D.O. (2008). *Aspekty teorii frazeologii*. Moskva: Znak.
- Deutsche Idiomatik*. Accessed at: [https://bitbucket.org/elena\\_krotova/deutsche\\_idiomatik](https://bitbucket.org/elena_krotova/deutsche_idiomatik) [31/03/2018]
- Deutsches Referenzkorpus*. Accessed at: <https://cosmas2.ids-mannheim.de/cosmas2-web/> [31/03/2018]
- Dobrovol’skij, D. (2008). Idiom-Modifikationen aus kognitiver Perspektive. In Kamper, H., Eichinger, L.M. (Hrsg.) *Sprache - Kognition -Kultur. Sprache zwischen mentaler Struktur und kultureller Prägung*. Berlin / New York: de Gruyter, 2008, pp. 302-322.
- Dobrovol’skij, D. (2006). Idiom dictionaries. In: Keith Brown, (Editor-in-Chief). *Encyclopedia of Language and Linguistics*. Second edition. Volume 5. Oxford: Elsevier, 2006, pp. 514-518.
- Dobrovol’skij, D.O. (1997). *Nemetsko-russkiy slovar’ zhivyykh idiom*. Moskva: Metatekst.
- Raykhshteyn, A.D. (1980) *Sopostavitel’nyy analiz nemetskoy i russkoy frazeologii*. Moskva: Vysshaya shkola.
- Sketch Engine*. Accessed at: <https://www.sketchengine.co.uk/> [31/03/2018]



# Neologisms





# On the Detection of Neologism Candidates as a Basis for Language Observation and Lexicographic Endeavors: the STyrLogism Project

**Andrea Abel, Egon W. Stemle**

*Institute for Applied Linguistics, Eurac Research*

*E-mail: andrea.abel@eurac.edu, egon.stemle@eurac.edu*

## Abstract

The goal of the project STyrLogisms is to semi-automatically extract candidate neologisms (new lexemes) for the German standard variety used in South Tyrol. We use a list of manually vetted URLs from news, magazines and blog websites of South Tyrol, and regularly crawl their data, clean and process it. We compare this new data to reference corpora, additional regional word lists and all the formerly crawled data sets. Our reference corpora are DECOW14, with around 60 million word forms, and the South Tyrolean Web Corpus, with around 2.4 million word forms; the additional word lists consist of named entities, terminological terms from the region and specific terms of the German standard variety used in South Tyrol (altogether around 53,000 word forms). Here, we will report on the method employed, the first round of candidate extraction with an approach for a classification schema for the selected candidates, and some remarks on the second extraction round.

**Keywords:** neologism, web corpus, dictionary of variants

## 1 Research Goals and Motivation

The goal of the STyrLogism project is to semi-automatically extract neologism candidates for the German standard variety used in South Tyrol, a province in Northern Italy where German is an official language. Immediate use-cases for these neologisms include, for example, consideration for future editions of the *Variantenwörterbuch des Deutschen* (*Dictionary of variants of the German language*, abbr. *VWB*) (Ammon, Bickel, & Lenz 2016) and other dictionaries. More generally, the project is to be used as an empirical basis for the long-term observation and evaluation of trends of the local standard variety of the German language, which makes it interesting for language policy and language planning measures.

The research on the German standard variety used in South Tyrol is based on the concept of pluricentricity of the German language (Clyne 1992, Ammon 1995). According to this, differing standard varieties are being used in German speaking areas. Among the crucial aspects for considering a variety a standard and not only a dialectal variety we can mention the official status of the language in a specific area, school instruction in the language, the existence of own codices, etc. South Tyrol is a particularly interesting object of linguistic studies, is due to its role as “national semi-center” (i.e. not having own language codices) from a pluricentric perspective, its marginal position within the German speaking area and the language contact situation (above all concerning the German and the Italian languages) (cf. Ammon, Bickel, & Lenz 2016). In 2016, the entirely revised second edition of the *VWB* was published 12 years after the first edition. But for this new edition it was not possible to analyze the South Tyrolean German variety to the same extent as the varieties of the “full centers” (Germany, Austria, Switzerland) and recent developments are less represented (cf. Abel 2018). We are aware that it is not among the aims of a dictionary of variants to record neologisms, but the

example shows that, in general, a constant, comprehensive language observation and documentation of the German language in South Tyrol over time is still missing, including, of course, the emergence of new words. However, the nexus between the research on language change and neologisms is evident, the former being the superordinate subject area (cf. Kinne 1998: 76).

## 2 Definitions of Key Concepts

Usually, investigations of lexical innovation use the following categories: neologisms, occasionalisms and other innovations. The aim of our project is the detection of neologism candidates. Neologisms are usually divided into at least the following two categories: one category for words used in a new meaning (*Neubedeutung*) but without any change of form, the other category for new lexemes (*Neulexem*) with an unseen graphical representation, for example compounds or derived forms (Kinne 1998: 83 ff.) (cf. Figure 1). According to Kinne (1998: 85) a defining feature of a neologism is that, initially, it is not included in any dictionary; it emerges from a communication need in a communication community and it passes through different phases, such as becoming common usage practice, acceptance, and lexicalization, as well as the perception of its newness from a majority of the language users.



Figure 1: Lexical innovation (Kinne 1998: 86, adapted version).

In the STyrLogism project, we focus on new lexemes of the written standard local language ignoring misspellings/typos, named entities and inflected forms. Furthermore, our currently employed methodology initially focuses on the identification of neologism candidates, as only longitudinal studies with repeated observations of individual candidates can reveal true neologisms. And, as stated elsewhere, each neologism is originally an occasionalism stemming from an individual need for expression (cf. Kinne 1998: 77-78, referring to Coseriu 1958/74). As such, we are currently less interested in frequency distributions and focus on the collection of an initial data set that, consequently, may include hapax legomena as well.

The analyzed candidates are used in general language or common academic language (*alltägliche Wissenschaftssprache*, cf. Ehlich 1993). The reason to also consider commonly used technical terms in our study is due to the fact that radical social, political and economic changes activate the genesis of neologisms, necessitating designations for new circumstances, institutions etc. (cf. Kinne 1998: 87). South Tyrol was part of Austria annexed to Italy after the First World War, and thus it is obvious that a large number of neologisms are due to this radical historical change. With regard to our project, a further restriction has to be done in the sense that we are exclusively interested in STyrLogism candidates, which means in those candidates whose usage is limited to South Tyrol.

Thus, neologism candidates in the STyrLogism project can be briefly characterized as follows:

- new lexemes, not lexicalized
- used in general language or in common academic language
- consideration of the written standard language
- exclusion of misspellings/typos
- exclusion of named entities
- exclusion of inflected forms of lexicalized words
- no distinction from occasionalisms possible.

In particular, STyrLogism candidates exhibit the following features:

- neologism candidates
- usage limited to South Tyrol.

The restriction of the usage of the entities considered in the project means that they are not present in the reference corpus DECOW14 (Schäfer & Bildhauer 2012) nor in the German neologism platform *Wortwarte* (Lemnitzer 2000-2017).

### 3 Related Work

Up-to-date high-quality word lists and structured data is not only required for lexicography, but is also helpful for a wide range of human-language technologies (HLT), such as machine translation, named entity recognition, and spelling error detection. With the recent success of neural network methods in HLT and the related word embeddings, the need for large amounts of unlabeled data, i.e. corpora, has been emphasized, with word lists and structured data accessory parts of this. However, they are still used for supervised training to adapt to new genres, domains or languages, or for evaluation purposes. (For detailed insights into recent developments see, for example, Bethard et al. 2016; Ide, Herbelot, and Màrquez 2017; Calzolari et al. 2018). With more diachronic, genre- and domain-specific corpora becoming available, automatic neologism detection provides a head start to improve lexicographic resources and HLT tools and, as such, is becoming increasingly important.

Generally speaking, the approaches for neologism detection can be divided into two groups. One, usually applied to a single set of new data, uses language resources like word lists or linguistic patterns. The word lists are compiled from existing lexicographic resources, such as dictionaries or corpora, combined with filters for the elimination of non-words, typographical errors, named entities, and so on, and the linguistic patterns are, for example, markers of lexical novelty like punctuation marks that can signal new words, as shown in O'Donovan and O'Neill (2008) and Paryzek (2008). The other group, usually applied to multiple data sets, uses statistical measures or machine learning to calculate and assess the increase in usage or the change in meaning over time or in different registers. For examples, see Stenetorp (2010), Herman. and Kovár (2013) and Kilgarrieff et al. (2015). Finally, these two approaches can also be combined. The STyrLogism project is currently following the former approach.

*Wortwarte* (Lemnitzer 2000-2017) is the most relevant previous project with regard to our own, as it is an ongoing project with an online portal that has been regularly collecting and documenting new German words. The system is based on German online-newspaper texts: a web crawler regularly collects data from pre-defined sites, such as newspapers and magazines. After cleaning the HTML content, the plain text is used to build a new time slice of a corpus. The selection of appropriate neologism candidates is carried out on the basis of short-term evaluations, where the new corpus is compared with the continuously growing German reference corpus (Das Deutsche Referenzkorpus

– DeReKo. For an overview, cf. Kupietz & Lungen 2014) with approximately 42 billion word tokens (status: 03.02.2018). To avoid “random” errors (e.g., typing errors) and filter out misspellings, the selection of neologisms is done ‘manually’ after the comparison with the DeReKo. The results these analyses are put online at irregular intervals, but as a rule of thumb about once a week. The results usually include a few words with their exemplary use in a sentence, and the reference as to where it came from.

O’Donovan and O’Neill (2008) use a similar idea, but in lack of free access to a continuously growing reference corpus for English they use and update their own Chambers Harrap International Corpus (CHIC) web corpus. It consists of more than 500 million words of international English, and is in the tradition of the Bank of English<sup>1</sup> rather than a static, balanced resource such as the British National Corpus (BNC). They also make use of other resources, like lemmatization and morpho-syntactic information, such as a headword list augmented with inflected forms.

Kerremans, Stegmayr, and Schmid (2011) also crawl their own reference corpus and, additionally, use an explicit component for monitoring the change over time for selected terms: they use the commercial search engine Google and regularly crawl the content of search results returned for each ‘to-be-monitored’ neologism.

## 4 Methods and Data

We use a list of manually selected URLs from news, magazines and blog websites of South Tyrol, and regularly crawl their data with the Internet Archive’s open-source, extensible, web-scale, archival-quality web crawler Heritrix<sup>2</sup>. The whole content from the crawled web pages is saved in the Web ARChive (WARC) archive format (ISO 28500), a method for combining multiple pages into an archive file together with related meta information, like retrieval date, URL, IP address. We then use Schäfer and Bildhauer’s (2012) *texrex* toolkit for web corpus construction, which performs basic cleanups and boilerplate removal, simple connected text detection as well as shingling to remove duplicates from the corpora. The toolkit comes already set up to process WARC files, and directly works with the *heritrix* output. It removes HTML and scripts, and uses a simplistic heuristic to split paragraphs in the resulting text. So-called boilerplate, i.e. navigational elements and menus, date strings, copyright notices, among others, are then identified and quantified as an annotation on a paragraph level. Finally, a two-step duplicate detection is employed: the first step removes perfect duplicates, i.e. documents that are identical up to the last character; the second step removes near-duplicates by computing token-*n*-grams for each page and the corresponding fingerprint (w-shingle). This fingerprint has the property that similar pages end up with similar fingerprints, and thus the data can easily be de-duplicated by selecting a range of allowed similarity between the fingerprints.

The resulting data is converted into a list of word forms and a corpus for the NoSketchEngine (NoSkE) (Rychlý 2007). We then do case-insensitive comparisons of the list of word forms with a) the one from our reference corpora, b) the additional word lists, which is in practice a simple Named Entity Recognition, and c) with the combination of all formerly crawled data sets. Our reference corpora are DECOW14 (Schäfer & Bildhauer 2012) with around 60 million word forms, and the South Tyrolean Web Corpus (Schulz, Lyding, and Nicolas 2013) with around 2.4 million word forms; the additional word lists consist of named entities, terminological terms from the region, and specific terms of the German standard variety used in South Tyrol (altogether around 53,000 word forms). The cleaned data of the current crawl is then tokenized – but not lemmatized – and converted into a word list. This

1 <http://www.collins.co.uk/books.aspx?group=153>

2 <https://archive.org/projects/>



list of candidate words consists of those in the current crawl that appear less than a predefined number of times in all of the other data.

Finally, the candidates are manually checked in a specifically crafted streamlined interface. This interface shows a set number of neologism candidates on one page along with the first (and possibly only) results as a KWIC result. The user can then click the candidate to get the whole result page of this candidate's search query in the NoSkE, where all additional meta information for each search result is available. The user can also click a checkbox or enter a comment into a text field (which automatically triggers the checkbox) to make a note of this candidate for later curation. Finally, the user can go to the next page, which automatically discards all unmarked candidates from further processing.

In a second 'curation' step, a user can see all the previously marked candidates with single KWIC results of all occurrences of the candidate in different crawler runs. This stage gives an overview of the currently tracked neologism candidates with quick access to individual occurrences over time (cf. Figure 2).



Figure 2: Web interface being used within the STyrLogism project.

## 5 Preliminary Results

The manually selected candidates are checked on a regular basis in order to allow a long-term monitoring. At the same time, our up-dated corpus is regularly analyzed with respect to possible new neologism candidates. Here, we will be reporting on the two rounds of manual checks of data crawled approximately two years apart from each other.

### 5.1 First Round: Approach for a Classification

For this round we used our initial list of 43 manually selected URLs and let the crawler run for almost two days. The minimum occurrences for a wordform to be considered were: it needed to occur at least



once in our data and was filtered out if it occurred at least once in the reference material. The result was about 70GB of raw web content from roughly 250,000 web pages. After cleaning and deduplication roughly 40,000 web pages remained. After comparing the new word forms to all our reference material roughly 4,000 neologism candidates remained.

The manual evaluation of the first extracted word list showed that many of the rejected candidates were a) two or more words written as one, i.e. the words were missing a space; b) unrecognizable words – with both a) and b) being erroneous left-overs of the boilerplate cleaning – c) foreign (mostly Italian) words, d) misspelled words and e) common words or variants of common words that are rare but established. This led to the selection of 340 candidates for further analysis. So far, the analyses of the first round of manually checked data allowed is to elaborate a preliminary classification of STyr-Logism candidates, including different kinds of emerging new word forms.

Thus, we have (a) legal and administrative common terms, e. g. *Landeszusatzvertrag* (regional amendment of a national collective agreement). Furthermore, we find (b) compounds with components of lexicalized variants of the standard German in South Tyrol recorded in the *VWB*. An example is *Optantengesetz* (a particular law for those people from South Tyrol who in 1939 opted for German citizenship and, consequently, decided to emigrate). In this case *Optant* is a lemma in the *VWB* but neither *Optantengesetz* nor other compounds are recorded as their own lemmas or as corresponding word formation units as in other cases in the *VWB*. Striking are examples such as *Luxuspensionär* (a retired person receiving a very high pension). *Pensionär* is recorded as a lemma in the *VWB* but is typically used in Switzerland, whereas in Austria and South Tyrol *Pensionist* is the commonly used term to refer to a retired person.

There are also (c) common words used in the standard German in South Tyrol which are not yet lexicalized. For this we can mention *Wahlsektion* (a part of a municipality whose inhabitants go to the same voting center). Although not a term equally used in the whole German speaking area, it is not recorded in the *VWB*. In addition, the manual checking revealed a series of (d) common words with uncommon word formation features which are at the interface between lexicon and grammar. *Mittelstandperson* (middle class person) may serve for illustration: IN this case we would expect an “-s-” as a linking element. A long-term monitoring may show if it is only a lapsus or a trend. However, we noticed several word formations following the same pattern, e. g. *Namenregelung* (naming policy). Generally speaking, there seems to be a tendency to use a linking “-s-” in compounding in South Tyrol, also similarities to its use in Austria, above all after -g, -k, -ch (cf. Ammon, Bickel, and Lenz 2016: LXXVI), although the picture is anything but clear (cf. Abfalterer 2007: 191).

Finally, we distinguish a category that on an interim basis we call (e) “true” neologism candidates. An illustrative example is the term *Vollautonomist* referring to a person standing up for a “full” political autonomy for South Tyrol remaining, at the same time, part of the Italian state and being, in this specific meaning, a particularity of the South Tyrolean context. It can be put up for discussion if the term is rather a new meaning than a new lexeme. However, the lexica and word lists used for our analyses did not contain the word form. A further example shows the use of an Italian loan word which is, according to Abfalterer (2007: 167ff.), one of the three main features of primary South Tyrolisms (i.e. variants which are supposed to occur only in South Tyrol) next to loan translations and “others”. In the compound *Vollkornpizzetta* (small pizza made of whole grain) the Italian *pizzetta* with the diminutive suffix -etta is used. It is still debatable where to draw the line between (c) and (e), as the mentioned forms are commonly known.

Given that the category of “true” neologism candidates is particularly relevant within our study, an attempt to carry out a preliminary characterization was done. With regard to the goals of this category, different key aspects became apparent. Thus, the lexical items are used for (1) humorous, ironic

or sarcastic and, furthermore, for (2) polemic or malicious ways of expression. (3) Creative language usage and the play on words can be observed as well, and this may also appear together with (1) and (2). *Donnerwetterer* is a case in point designating in an original way someone railing against someone/something; *Donnerwetter* is a common German word, originally mainly used to refer to a thunderstorm, but nowadays referring to a loud confrontation or used as an interjection to express either anger or admiring astonishment; however, the unit, with the suffix *-er* that in German word formation is typically used to refer to persons, has not been lexicalized. Finally, the (4) designation of new circumstances, facts and objects can, of course, be found. For example, *Bausündennachlass* is used to indicate a legal measure in Italy for remitting a financial penalty for an eyesore.

The items have a number of typical features. With respect to word formation we notice a tendency towards (i) compounding (cf. also Abfalterer 2007: 189), and partly complex compounds are to be found. For instance, if we take *Regiokornbrot* then *Regiokorn* is used to designate regional grain, originally deriving from the name of a local project with the title *Regiokorn*; subsequently the term was being used more generally for bread made of local grain. Closely connected to this phenomenon is the (ii) strategy of turning phrasemes into single words. In the case of *Mundaufreißer* the idiomatic expression *das Maul* (also: *Mund*) *aufreißen* (to give oneself airs, literally to open the mouth wide (*Maul* in German regards to an animal, *Mund* to a person)) is used in an unusual way as a compound that is conflicting with the principle of fixedness of phraseologisms (cf. Burger 2007).

As expected, some candidates constitute (iii) loan words or loan translations from the Italian language. Here we might mention *Promotorenkomitee* (a committee of initiators, supporters of an action) which is a literal translation of the Italian *comitato promotore*, commonly used in South Tyrol, whereas in other German speaking areas words such as *Initiatoren* or *Befürworter* are used instead. We also have (iv) formal analogies to lexicalized variants. An interesting case is *Schwammlklauber* (a person picking mushrooms), a commonly used word form in South Tyrol but not recorded in the *VWB*. However, *Schwammerl* (mushroom) is a lemma in the *VWB* (used – with the diminutive suffix *-erl* – in southeast Germany and in Austria which, according to the rules applied for the *VWB*, means that the usage in South Tyrol is implied, cf. Ammon, Bickel, and Lenz 2016: LXXVI) but not the form *Schwamml*, which is the typical word form in the South Tyrolean context (the assumption of the use of the suffix *-erl* also in South Tyrol in the *VWB* is shown in other cases as well, e.g. concerning the lemma *Sackerl*, i.e. a carrying bag, which is not used in South Tyrol). The verb *klauben* (to harvest, to pick) is also lexicalized in the *VWB* (used in southeast Germany and in Austria) containing the diasystematic label “borderline case of the standard language”. On the other hand *Apfelklauber* is recorded as primary South Tyrolism in the *VWB*, and this without any diasystematic label. However, it has an own, limited meaning as it refers to a person helping to harvest apples being paid for this activity, whereas *Schwammlklauber* indicates someone doing the activity for leisure.

Among the affected domains, it is worth mentioning politics, environment, tourism, leisure and food.

## 5.2 Second Round: Some Remarks

For this round we used an updated list with 156 manually selected URLs and let the crawler run for three days. The minimum occurrences for a wordform to be considered were: it needed to occur once in our data and was filtered out if it occurred once in the reference material. The result was about 60GB of raw web content from roughly 500,000 web pages. After cleaning and deduplication roughly 50,000 web pages remained. After comparing the new word forms to all our reference material, roughly 7,000 neologism candidates remained. From the monitored candidates, only seven reappeared in the new data set.

Although the overlapping of the comparison was low, we might have a closer look at one of the word fields affected. It is notable that morphological variations of *autonomiefreundlich* (autonomy-friendly) and *autonomiefeindlich* (anti-autonomy) reappeared in the second round. These lexical units form a part of a word field of a constantly hot topic in the South Tyrolean context, being the political autonomy perceived as an important achievement for the German speaking population (cf. Autonome Provinz Bozen Südtirol: 2004). Thus, we also found *Vollautonomist* in the first round. Furthermore, a data checking in the DECOW14 corpus (Schäfer & Bildhauer 2012) confirmed the former usage of *Vollautonomie* (“full” autonomy) which, to the best of our knowledge, has never been in discussion for inclusion in the *VWB*. Furthermore, we can find the commonly used *Autonomie* (autonomy) exclusively in this narrow political sense, including patterns such as *dynamische Autonomie* (dynamic autonomy).

## 6 Conclusions and Outlook

In the paper we gave an overview on our work focusing on those neologism candidates with the potential to persist over time and to be lexicalized. The first findings of the initiative show that the approach is suitable to produce candidate lists for newly arisen words, or rather word forms not included in the corpora and word lists to date, even though a large amount of noise had to be eliminated manually. Within the time period taken into consideration and with the data basis used so far it was not possible to distill a larger amount of lexical units being characterized by persistence over time. However, the approach turned out to be a useful support for the overarching endeavor of language observation and documentation in South Tyrol.

We found that the online publishing attitude in South Tyrol makes our task more difficult in two ways: first, major newspaper and magazine publishers in the region often only put an excerpt or summary of an article online. This reduces the amount of actual text that can be used for our analyses, and also complicates the extraction of content from single web pages: extracting content from a web page is a balancing act between getting as much of the desired textual content as possible (recall), but at the same time *only* getting the desired content and not the superfluous boilerplate (precision). This task generally gets more difficult with lower amounts of available content, and produces more noise with a lower content-to-boilerplate-ratio (Schäfer & Bildhauer 2012). Second, articles only stay online for a short period of time. This period, depending on unknown factors, can be as short seven days.

Consequently, a methodological possible next step includes to shorten our crawl interval to be around the minimum content availability time in the region. This would mitigate the otherwise unavoidable loss of early onsets of new word forms and, additionally, would also enable precise time series analyses for word usage over time. To this end, we could use the SketchEngine’s “Trends: Neologisms and diachronic analysis of word usage” feature (cf. Herman & Kovár (2013) for the version currently implemented in the SketchEngine) as a start and see whether this yields promising results. Later, we could adapt the idea for our particular use-case.

Utilizing social media and thereby extending the basis for the data analyses could also prove helpful: users produce a tremendous amount of text each day on social media, much of which is readily available without the complications of boilerplate removal, as needed for web pages. This development has opened new possibilities for lexicographical analyses, such as, in “particular, corpus patterns that are very rare in conventional-size corpora turn out to have many occurrences in the very large corpora of social media” (Cook 2012).

A different direction could also be to detect novel senses, i.e. semantic changes in established word forms, based on distributional similarity between word models built from different corpora (cf.

Gulordava & Baroni (2011) for a successful application of vector space models in this context). Here, we could also employ word embeddings (Mikolov et al. 2013), a recently very successful language modeling technique with results, often on par or superior to the established vector space models.

## References

- Abel, A. (2018). Von Bars, Oberschulen und weißen Stimmzetteln: zum Wortschatz des Standarddeutschen in Südtirol. In S. Rabanus (ed.) *Deutsch als Minderheitensprache in Italien. Theorie und Empirie kontaktinduzierten Sprachwandels*. - Germanistische Linguistik: Themenheft, pp. 283-323
- Abfalterer, H. (2007). *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht*. Innsbruck: Innsbruck University Press.
- Ammon, U. (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. Berlin/New York: De Gruyter.
- Ammon, U., Bickel, H. & Lenz, A. N. (eds.) (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2nd ed. Berlin/Boston: De Gruyter Mouton.
- Autonome Provinz Bozen Südtirol (eds.) (2004). *Südtirol-Handbuch*. 23th ed. Bolzano/Bozen: Landespresseamt.
- Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P. & Zesch, T. (eds.) (2016). *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics.
- Burger, H. (2007). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.
- Ehlich, K. (1993). Deutsch als fremde Wissenschaftssprache. In A. Wierlacher et al. (eds.) *Jahrbuch Deutsch als Fremdsprache*, 19. München: iudicium, pp. 13-42.
- Kinne, M. (1998). Der lange Weg zum Neologismenwörterbuch. Neologismus und Neologismenlexikographie im Deutschen. Zur Forschungsgeschichte und zur Terminologie, über Vorbilder und Aufgaben. In W. Teubert (ed.) *Neologie und Korpus*. Tübingen: Gunter Narr Verlag, pp. 63-110.
- Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. & Tokunaga, T. (eds.) (2018). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).
- Cook, P. (2012). Using social media to find English lexical blends. In R. V. Fjeld, J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 846-854.
- Gulordava, K. & Baroni, M. (2011). A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 67-71.
- Herman, O. & Kovár, V. (2013). Methods for Detection of Word Usage over Time. In *Proceedings of the Seventh Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2013)*. Brno, Czech Republic: Tribun EU
- Ide, N., Herbelot, A. & Márquez, L. (eds.) (2017). *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (\*SEM 2017)*. Association for Computational Linguistics.
- International Organization for Standardization (2017). Information and documentation - WARC file format (ISO 28500).
- Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2011). The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. In K. Allan, J. A. Robinson (eds.) *Current Methods in Historical Semantics*. Berlin/Boston: De Gruyter, pp. 59-96. <http://doi.org/10.1515/9783110252903.59>
- Kilgariff, A., Ondřej, H., Bušta, J., Rychlý, P. & Jakubiček, M. (2015). DIACRAN: a framework for diachronic analysis. In *Corpus Linguistics (CL2015)*, United Kingdom.
- Kupietz, M. & Lungen, H. (2014). Recent Developments in DeReKo. In N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Lemnitzer, L. (2000-2017). Die Wortwarte. Accessed at: <http://wortwarte.de/> [April 28, 2017]
- Mikolov, T., Corrado, G., Chen, K. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1-12.



- O'Donovan, R. & O'Neill, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In J. D. E. Bernal (ed.) *Proceedings of the 13th EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 571–579.
- Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. In *Investigationes Linguisticae, XVI*; Adam Mickiewicz University: Poznań, Poland.
- Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*. Brno, Czech Republic: Masaryk University, pp. 65–70.
- Schäfer, R. & Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In N. Calzolari et al. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).
- Schulz, S., Lyding, V. & Nicolas, L. (2013). StirWaC: compiling a diverse corpus based on texts from the web for South Tyrolean German. In S. Evert, E. Stemle, P. Rayson (eds.) *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*. Lancaster, UK, pp. 35–45.
- Stenetorp, P. (2010). Automated extraction of swedish neologisms using a temporally annotated corpus. Stockholm, Sweden: Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.



# Neologisms in Online British-English versus American-English Dictionaries

**Sharon Creese**

Coventry University

E-mail: [creeses@uni.coventry.ac.uk](mailto:creeses@uni.coventry.ac.uk)

## Abstract

A common source of publicity for modern-day dictionary publishers is the regular (usually quarterly) release of lists of neologisms that have recently been added to their online dictionaries.<sup>1</sup> The publishing of updated versions of these sites every few months means it may no longer take years for new words to be included in a dictionary. However while different dictionaries may utilize neologisms in similar ways in order to improve brand awareness, the way in which these new words are presented and used in the dictionaries themselves can vary widely, including amongst those of differing varieties of English. This paper will describe differences in the approach and treatment of British-English neologisms in online editions of British-English dictionary *OED* (the *Oxford English Dictionary*) and American-English dictionary *Merriam-Webster*. In particular, the way in which each dictionary responds to potential new words will be discussed, as will the comprehensiveness of the resulting new entries and the differences found in the types of information each contains.

**Keywords:** neologism, lexicography, dictionaries, dictionary components, British-English, American-English, OED, Merriam-Webster

## 1 Introduction

Thousands of new words enter online dictionaries every year; according to their own publicity material, some 3,200 entered the *Oxford English Dictionary* (*OED, Third Edition*) during the year to January 2018,<sup>2</sup> while 1,100 entered *Merriam-Webster*<sup>3</sup> in the 12 months to March 2018.<sup>4</sup> Many of these novel words name or describe new items or concepts created through innovations in science and technology (Lehrer 2003: 371; Franc 2011: 417; Mitchell 2008: 33), and many others are coined by journalists or other professional writers, often in a bid to inject humor into a story, or simply as an expression of language play (Renouf 2007: 70). Neologisms can gain great popularity through repetitions of such use in the media, and indeed this is often where language users first come into contact with a new word.

As the disparity in the numbers of new words entering the two dictionaries above implies, there are significant differences in the approach to neologisms, leading to some dictionaries accepting more new words than others, and this is further reflected in the treatment of the new words which ultimately do appear. This paper discusses this issue in the context of the *OED* and *Merriam-Webster* online dictionaries, drawing upon the findings of a wider piece of research which investigated degrees of comprehensiveness between dictionary types and levels of responsiveness to new words (see Creese

1 See for example the September 2017 updates for the *OED* (<https://public.oed.com/the-oed-today/recent-updates-to-the-oed/september-2017-update/new-words-list-september-2017/>) and *Merriam Webster* (<https://www.merriam-webster.com/words-at-play/new-words-in-the-dictionary-sep-2017>).

2 See <https://public.oed.com/the-oed-today/recent-updates-to-the-oed/>.

3 Formerly based on the *Collegiate® Dictionary*, Eleventh Edition and now based on an amalgam of the company's dictionary products (Merriam-Webster 2018a).

4 See for example <https://www.merriam-webster.com/words-at-play/new-words-in-the-dictionary-march-2018>.

2017). Dictionary entries were analyzed based on the number and quality of industry-accepted “standardized” dictionary components they contained, plus a number of “non-standard” but increasingly common additional components, while responsiveness was assessed by how quickly new words (used in British-English national newspapers) appeared in dictionaries. However, the lack of accurate, or at times any, dating information made the latter extremely problematic for the *OED* and *Merriam-Webster*, particularly in the case of the latter, which has only recently begun adding “update” information, and still does not include any “inclusion date” details.

## 2 Materials and Methods

The wider study from which this paper is drawn involved examination, tracking and analysis of a sample of 34 neologisms, although only half of these featured in the *OED* and/or *Merriam-Webster*. As Table 1 demonstrates, 10 of these new words appeared in both dictionaries, with a further five appearing only in the *OED* and just two appearing only in *Merriam-Webster*.

Table 1: Neologisms appearing in *OED* and/or *Merriam-Webster* online dictionaries (source of definitions: *NeoCrawler* list, Ludwig-Maximilians Universität, n.d.).

Neologism	Part Of Speech	Meaning	<i>OED</i>	<i>Merriam-Webster</i>
acedia	noun	spiritual or mental sloth	Y	Y
bogof	noun	an advertising strategy that entices people to buy a product and get one for free	Y	
conurbation	noun	an extensive urban area resulting from the expansion of several cities or towns so that they coalesce but usually retain their separate identities	Y	Y
cyberbullying	noun	the use of internet and mobile phones to send embarrassing or hurting [sic] messages	Y	Y
cyberchondriac	noun	person who imagines they have a particular disease because their symptoms match those listed on an internet health site	Y	
earworm	noun	a piece of music that sticks in a person’s head		Y
e-tailer	noun	a company which uses the internet to sell its products	Y	Y
e-waste	noun	electronic products which have been discarded or have become useless		Y
frenemy	noun	a person you assume is a friend, although you don’t really like him/her	Y	Y
greenwashing	noun	the practice of making an unsubstantiated or misleading claim about the environmental benefits of a product, service, technology or company practice	Y	Y
hubristic	adjective	referring to someone or something behaving with hubris	Y	Y
promissory note	noun	a negotiable instrument, wherein one party (the maker or issuer) makes an unconditional promise in writing to pay a determinate sum of money to the other (the payee), either at a fixed or determinable future time or on demand of the payee, under spec	Y	Y
rewilding	noun	the process of returning species, habitats and landscapes to a natural state, as they would be without the intervention of humans	Y	
tenebrous	adjective	dark and gloomy	Y	Y

Neologism	Part Of Speech	Meaning	OED	Merriam-Webster
upskill	verb	to give employees extra training in order to improve their performance	Y	
warrantless	adjective	without a warrant; especially referring to governments' surveillance practices after 9/11	Y	
waterboarding	noun	a torture method of putting a cloth over the face and pouring water over it to make them believe they are drowning	Y	Y

## 2.1 Selection of Neologisms

The neologisms selected for this study were among those believed likely to be of interest to researchers in the fields of lexicography and neology, as well as linguists in general, based on the words' characteristics such as their development and behavior over time, and differences in the numbers and types of components appearing in their initial and developing dictionary entries.

Rather than effectively reinventing the wheel by identifying a set of neologisms from which to make this selection, it was decided to use an existing list of new words which had already been the subject of extensive analysis. This had been produced by the *NeoCrawler* program, designed and created by the EnerG team at Ludwig-Maximilians University, Munich.<sup>5</sup> The new words had been identified and tracked within the Google Blogs environment (Kerremans 2015: 80), and researching them in my own project, in the context of lexicography, would allow for an expansion of the body of knowledge available about them. A review of the literature at the time and currently reveals that very little work has been done on the relationship between neology and lexicography, making this study of particular interest. Whilst Moon (2008) explores lexical creativity and dictionaries, Fischer (1998), Kerremans (2015) and Renouf (2013) make only passing references to lexicography/dictionaries in their studies of new words. Weiner (2009) gives a brief overview of the history of neologism inclusion in the *OED*, from separate "Supplements" through the "NEWS" ("New English Words Series") to the "Additions" volumes incorporated into *OED2* (see 3.1) and the current system of quarterly updates (2009: 391, 401). Algeo, meanwhile, reports that more than half of the neologisms from a 30-year corpus had no dictionary presence just two decades later (1993: 281), raising questions about the long-term survival rates of new words.

The *NeoCrawler* database (the creation and use of which is explained in full in Kerremans 2015) contained many more neologisms than were necessary or practicable for use in this study. Thus it was necessary to select from the full list a manageable sample of new words (a maximum of 40). Prior to extracting a random sample from the list, it was necessary to exclude words such as:

- Trade names
- Words which were likely simple misspellings, but which had gained popularity (an example of what Neuman, Nave and Dolev define as "buzzwords", or "fashion words" that enter the language and rapidly acquire great popularity before fading into obscurity (2010: 58, 67).
- Non-British-English words. The purpose here was to see how British-English neologisms fared in both a domestic dictionary and an American-English one. Specifically, the objective was to see whether there was any delay in the acceptance of neologisms – or indeed any outright failure of take-up of these words – due to a lack of the frame of reference needed to understand them. Examples might include new cricketing terms, since cricket is not played in the US (Creese 2017: 68).

<sup>5</sup> See <http://www.neocrawler.anglistik.uni-muenchen.de/crawler/html/index.php?abfrage=about>.

Having removed these unsuitable terms, the remainder of the *NeoCrawler* list was checked against a number of UK national newspapers, in order to ascertain whether the words were in “real-world” usage. Words which had appeared in these newspapers were carried forward to the next stage of selection. Newspapers were chosen as an indicator of real-world usage because they are aimed at a broad-cross section of the population and are produced daily (giving large numbers of readers regular and frequent exposure to any neologisms used).

The remaining neologisms were then entered into an online “randomizing” program,<sup>6</sup> which was used to select the final list of words used in this study. It had been decided, in the interests of sample representativeness, that in terms of word class the make-up of this list of neologisms should reflect that of the original *NeoCrawler* list. As a result, a total of 34 neologisms were selected, 82% of which were nouns, with just four adjectives and two verbs. Of these 34, only 17 appeared in one or both of the dictionaries under discussion here, 76% of which were nouns, with three adjectives and a single verb (see Table 1 above).

Several of the selected neologisms were recognized as having actually been included in the *OED* for some time; long enough to not normally be considered “new”. These were “acedia”, “conurbation”, “hubristic”, “upskill” and “warrantless”. Three of them had also already appeared in *Merriam-Webster* (“acedia”, “conurbation”, “hubristic”), although as with all *Merriam-Webster* entries, there was no way to know *when* they had been accepted. All five, however had recently been accepted into other dictionaries covered by my wider research study, as well as obviously having been identified by the *NeoCrawler* system as “new” within the Google Blogs environment (Kerremans 2015). It seemed possible therefore that these terms might be experiencing some kind of revival following a period of possible dormancy, and it was decided to retain them as a kind of “neologic wildcard”, to see if anything could be learned from them. Although these “candidate neologisms” are really more “reincarnated” than they are “new”, for the purposes of this paper they are included in discussions of “neologisms”, unless stated otherwise.

## 2.2 Dictionary Inclusion Criteria

One of the key differences between the *OED* and *Merriam-Webster* is that the *OED* can be termed a “historical dictionary”, meaning that it aims “more than any other at comprehensiveness of inclusion rather than at a reportage of current use” (Algeo 1993: 283). Once a word enters the *OED*, it is never removed. It is “the most complete record of the English language ever assembled” (Oxford University Press 2015). *Merriam-Webster*, meanwhile, leans slightly more towards being a dictionary of current language use, although it rarely removes words, and when it does, it is only during major revisions conducted once every ten years (Mitchell 2008: 34).

Attestation, that is, proving that a word actually exists in the language, by showing it *in situ* (Atkins & Rundell 2008: 453 citing Simpson 2003: 268) is one of the “dictionary inclusion criteria” which are used to judge whether, for example, a new word or meaning is considered suitable for acceptance into that dictionary, either as a new entry, or as an additional sense for an existing word. While available space used to be another key factor in making these decisions in the days of the print dictionary, for today’s online dictionaries this is no longer the case. These days, the frequency and breadth of use of a term are key, as demonstrated by its citations (Atkins & Rundell 2008: 48). Both the *OED* and *Merriam-Webster* use citations as their main means of “attestation”, (although the *OED* is also now supported by corpus data, most notably the *Oxford English Corpus* (Oxford University Press 2018b) and the *Oxford New Words Corpus* (Oxford University Press 2018). (While *Merriam-Webster* calls its database of citations a corpus (Merriam-Webster 2018b), this, in my view, is a misnomer. A corpus is, according to Sinclair ‘a collection of pieces of language text in electronic form, selected according to external criteria to

6 Research Randomizer: <https://www.randomizer.org/>.

represent, as far as possible, a language or language variety as a source of data for linguistic research' (2004: 22). *Merriam-Webster's* "corpus" does not in my view fit this definition, and is instead simply a database of digitized information.) Collecting these citations is carried out through the continued use of one of the earliest methods of dictionary compilation: extensive reading programs which show that the new words or meanings in question have been in use for long enough, and in a wide enough variety of publications and sources, to make them suitable candidates for inclusion in the dictionary (Atkins & Rundell 2008: 51; Merriam-Webster 2018b; Oxford University Press 2018a). Ten years ago, the key publication types for this process were printed ones, however today it is increasingly online information that is used to prove a new word's credentials (Mitchell 2008: 33). In addition, the length of time that a word must have been in circulation, in order to be considered a viable candidate for dictionary inclusion, has significantly reduced, in response to the much faster pace at which words can develop, and dictionaries can respond, as a result of electronic technologies. Even as recently as 2015, the *OED's* publicity FAQ pages stated that citations had to show a potential new addition "in actual use over a period of at least ten years" (Oxford University Press 2015). Today, new words are simply required to have been "widely used in print or online" (Oxford University Press 2018d). *Merriam-Webster* has always been less specific, even four years ago giving no indication of how many citations a word should have, only that there should be enough "to show that it is widely used" and that they should "come from a wide range of publications over a considerable period of time" (Merriam-Webster 2015).

### 2.3 Dictionary Components

The components used as vehicles for comparison between the *OED* and *Merriam-Webster* online dictionaries were in the main those recognized as industry-standards, as demonstrated in Atkins and Rundell's guide to the practicalities of planning and building a dictionary (2008: 200-246; 385-462). These components should enable information on words to be presented uniformly and recognizably across all entries. In addition, several non-standard components were also used, and these are increasingly being found in online dictionaries, which do not suffer from constraints of space or typesetting. These include elements such as inclusion dates and audio files (Creese 2017: 83). The industry-standard dictionary components are shown in Table 2, while Table 3 presents the non-standard components. There are significant differences, however, in which of these components each dictionary includes for the same neologism, as will be shown below.

Table 2: Standardized dictionary components used in neologism entries in *OED* and *Merriam-Webster* (Atkins & Rundell 2008: 200-246; 385-462).

Component	Description
Headword (lemma)	Indication of how a word should be written, shown at the beginning of the entry
Lexical unit	Subdivisions of headwords (also known as senses)
Definition	Explanation of the meaning of a headword
Pronunciation	Information on how a word should be pronounced
Etymology	Origins of a word. I include here word formations and earliest known use of the word
Spelling variant	Permitted differences in spelling of the headword / senses
Word class	Parts of speech
Grammar label	Grammatical information on correct use of the headword
Examples/ quotations	Exemplars of how a word is used in real-life (including source information in the case of quotations)
Register/style/ attitude label	Indicators of the tone of the headword
Region label	Indication of where the headword is generally used
Cross-reference	Marker showing that more information is available on the headword elsewhere
Run-on	Indicator that the word derives from another headword



Table 3: Non-standard dictionary components used in neologism entries appearing in *OED* and *Merriam-Webster* (Creese 2017: 83).

Component	Description
Inclusion/update date	Date indicating when the word first entered the dictionary
Audio file	Sound file aiding correct pronunciation of unfamiliar words (usually in both British- and American-English)
Derivative	Marker showing that the neologism in question has derivatives
Related term	Indicator that another word is linked to the headword, without actually being a run-on. This information can alternatively appear as a standardized Usage Note

### 3 Findings

While each is widely viewed as being a leading dictionary (see for example Sullivan 2017 and Hanks 2013), the findings of this study show the *OED* to be significantly more open to neologisms than *Merriam-Webster*. It accepts more neologisms, includes more information in the new-word entries it carries, and the quality of the information included is higher. Unfortunately, it is not clear why, given such similar inclusion criteria, this disparity exists. There is no way to know why a lexicographer chooses one component and not another. One can only speculate that perhaps *Merriam-Webster* employs additional criteria which it does not make public. Or perhaps the type and scope of materials read as part of its attestation process is more limited than that of the *OED*. Or it may simply be that for some reason words which do meet the criteria are somehow simply not collected. Former *OED* Editor at Large Jesse Sheidlower is cited as saying that the *OED* is “less conservative than *Merriam-Webster* in how quickly it accepts new words” and maybe this is the reason for the difference (Mitchell 2008: 33).

In addition, it is often difficult, if not impossible, to tell how long a particular neologism or dictionary component has been present in one of these dictionaries – and therefore how responsive the dictionary is to new words’ existence – due to inconsistent and often absent dating information.

#### 3.1 Dictionary Entry Dating and Responsiveness to New Words

One of the mechanisms used in the wider study from which this paper is drawn to assess the responsiveness of dictionaries to the presence of new words was the date that a new word entered a particular dictionary. The sooner a dictionary accepted a new word, the more responsive it could be considered (notwithstanding adherence to the dictionary’s inclusion criteria). However in the case of *Merriam-Webster* there was no date information provided. In the *OED*, full entries included various date details – sometimes a date of entry, sometimes a date of update or that the entry was awaiting update. However, “run-ons” did not have any date information outside of that provided for the main entry; there was no indication of when the “run-on” was added, whether at the same time as the main entry, or later on. What information there was, was also often unreliable. As well as the main date information (appearing in the top right hand corner of the entry), many entries also carried a link to a “Publication History” (now titled “Entry History”) box. However, while these used to give information such as when an entry had been updated, they did not say what had been done to it: for example whether a meaning had been added or changed, or whether a new dictionary component had been included. If there was more than one sense to the word, there was no indication which one had been altered. Even more problematic was the fact that quite often these “Publication Histories” were simply wrong. In June 2016 the *OED* “Publication History” for “greenwashing” suggested that the

term had been added to the online dictionary in March 2016, yet my own research had shown it to be present as early as August 2014. Today, “Entry History” boxes include even less information. The “greenwashing” box now makes no reference to when the word entered the online dictionary, but includes the stock phrase:

oed.com is a living text, updated every three months. Updates may include:

- further revisions to definitions, pronunciation, etymology, variant spellings, quotations, dating or styling of citations;
- new senses or phrases (Oxford University Press 2018c)

Entries which have been included in the previous edition of the *OED* (known as *OED2*) show a “Previous Version” link, taking the user to the entry which appeared in the 1989 version of the dictionary.<sup>7</sup> All of the “reincarnated terms” except “upskill” (which entered after the *OED2* was published) include this link. “Acedia” is marked as having been updated in 2011, while the entries for “conurbation”, “hubristic”, “upskill” and “warrantless” all show that they are each still awaiting updates. This suggests that these terms may indeed be still developing and experiencing some form of revival, as suggested in 2.1.

Although *Merriam-Webster* carried no date information for any of its entries at the time of this study, it has now introduced some limited dating, although still not enough to be able to draw conclusions about its responsiveness to new words, since the most important date – when the word was first included – is still absent. However, some entries do now show when the entry was last updated, and some include a “first known use” component, giving an idea of the history of the word. It is not known when these new features were introduced, although the earliest “updated” date on any of these neologisms is November 2017. This was present on the entry for “acedia” in January 2018, however in March of the same year it had disappeared, as had its newly-added date of “first known use”. “Conurbation” has kept its new “first known use” and also gained an “updated” date (as have “promissory note” and “waterboarding”). “Hubristic” and “warrantless” have both remained dateless because they are “run-ons”, and “upskill” is not yet included in *Merriam-Webster*. Meanwhile “frenemy’s” new “update” information has been updated again, although it still has no other dates, and “earworm”, like “acedia” has lost its new-found “updated” information, although it has retained its new “first known use”. This latter date was 1802, yet there is no indication as to which of the two senses of “earworm” this applies to. It is only by referring to the dates of the attributed examples for the required sense, and comparing these with the dates of the attributed quotations in *OED*, that it becomes clear that this “first known use” applies to “corn earworm” (an agricultural pest) and not “a song or melody that keeps repeating in one’s mind”.<sup>8</sup>

It is not clear whether those *Merriam-Webster* entries which still do not have any dating features simply have not yet been updated, or whether they are missing in error. It would, however, seem from the many changes currently taking place on the website that the *Merriam-Webster* dictionary is perhaps undergoing another period of revision in March 2018. Perhaps when this is completed more entries will carry this important date information, and perhaps in time “inclusion dates” will also be added. Until that point, however, it is impossible to say with any accuracy which of these two dictionaries responds more quickly to new words.

### 3.2 Dictionary Components Findings

Of the 17 new words found to appear in one or both of the dictionaries under study here, the *OED* carried 15, while *Merriam-Webster* included only 12. Ten neologisms appeared in both dictionaries, although in all but two cases the *OED*’s entries were significantly more comprehensive, as Figure 1 shows.

<sup>7</sup> See for example <http://www.oed.com/view/Entry/199107?redirectedFrom=tenebrous#eid>.

<sup>8</sup> <https://www.merriam-webster.com/dictionary/earworm>

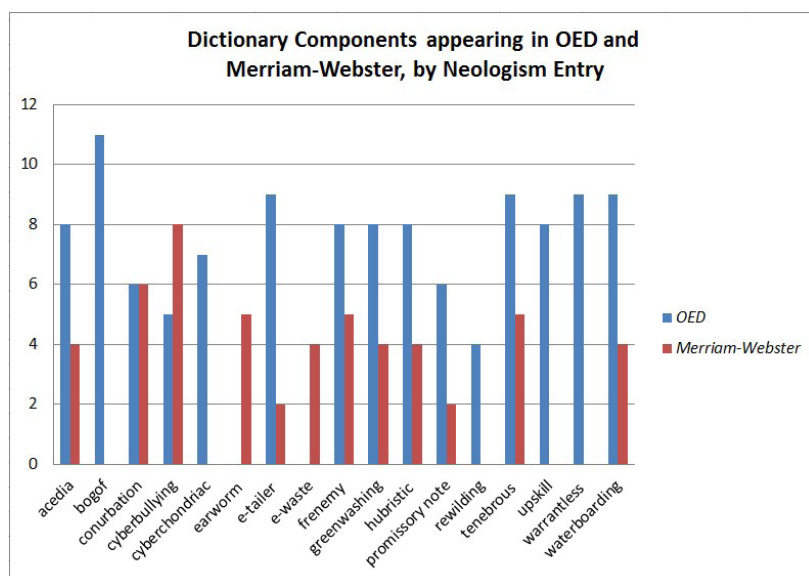


Figure 1: Dictionary components appearing in neologism entries in the *OED* and/or *Merriam-Webster*.

The *OED* provides these new words with more detailed entries, with ten of its 15 neologisms featuring between eight and ten dictionary components, while nine of the US dictionary's 12 entries feature between just five and seven. In only one instance does *Merriam-Webster* include more information than the *OED*, that being “cyberbullying”, and the reason is that in the *OED* this appears only as a “run-on” of “cyber”, while *Merriam-Webster* includes a full entry. Interestingly, although it includes “cyberbullying” it does not include “cyberchondriac”, whereas *OED* also includes this, again as a “run-on” of “cyber” (although for reasons unknown, the former is hyphenated, and the latter is not).<sup>9</sup> “Run-ons” do not include definitions in *Merriam-Webster*, presumably because it is assumed that the user will understand their meaning based on the definition for the word from which they are derived.<sup>10</sup> The *OED*'s “run-ons” are more comprehensive because they *do* generally include definitions, as “cyberbullying” and “cyberchondriac” showed, but for some (unknown) reason “rewilding” (the process of returning land to a more natural state) is not defined (though its meaning is easily understood from the definition of the “headword”). In all cases however, “run-ons” lack “headwords”, since they are simply a subdivision of a larger entry.

In addition to its entries being less comprehensive in terms of numbers of dictionary components, definitions in *Merriam-Webster* have frequently been confusing. Prior to recent changes (see below) they regularly featured more than one definition for the same word (as distinct from multiple senses of the word), but with no indication as to why, or which definition should take precedence. In rare instances there would be a label marking one of the definitions as, for example, “business”, suggesting that it came from a business dictionary. However in most cases there were simply two definitions with the same meaning, the second of which would be labelled “Full Definition”. In some instances this was the more detailed of the two, and in others it was the simplest. For example, the entry for “promissory note” featured:

(1) initial definition: “business: a written promise to pay an amount of money before a particular date” followed by “Full Definition of PROMISSORY NOTE : a written promise to pay at a fixed or determinable future time a sum of money to a specified individual or bearer” (Merriam-Webster 2014).

The current entry is much clearer: the “Full Definition” is the main one and the other in fact comes from the LearnersDictionary.com,<sup>11</sup> with “business” as a “domain label” indicating the subject area

<sup>9</sup> See <http://www.oed.com/view/Entry/250879?rskey=xTMDWS&result=2&isAdvanced=false#eid>.

<sup>10</sup> See for example <https://www.merriam-webster.com/dictionary/e-tail>.

<sup>11</sup> [Http://www.learnersdictionary.com/](http://www.learnersdictionary.com/).

to which the entry belongs (Atkins & Rundell 2008: 227). These changes may well have occurred during the course of a major overhaul of the *Merriam-Webster* online dictionary which is believed to have taken place at the end of 2017, (and is believed to have been the point at which the site ceased to be based solely on the *Collegiate® Dictionary* and was instead widened to overtly incorporate input from a broad range of the publisher's products<sup>12</sup> (Merriam-Webster 2018a)). This new site is somewhat clearer, inasmuch as all definitions are now properly labelled and users need only visit one page to find any number of definitions of the word from different subject areas and contexts. However, drawing on a broader range of the publisher's products means that some entries are now extremely long, because the definitions from three, four or more dictionaries are included in one online entry. Different word classes are also all in that same single entry, instead of a menu system appearing in response to a search query, as is the case in the *OED*, allowing the user to select the version of the word they require. *Merriam-Webster*'s entry for "warrant", for example carries a detailed entry, including definitions, multiple senses and examples, for both noun and verb in the generalized dictionary, as well as the "financial", "learners", "kids", and "law" dictionaries. It also includes the "run-on" "warrantless", the only extra word from this study to be added to *Merriam-Webster* in the past four years (and even this is actually a "reincarnated" term).<sup>13</sup> This welter of information, coupled with the inevitable advertisements found across the Web, makes for a very daunting and confusing entry. It is also in stark contrast to the earlier format, which was extremely sparse.

In monolingual dictionaries like these we would expect to always see a definition given for a full entry, yet in rare instances this is not the case. For example "acedia" in the *OED*, rather than having a definition, simply has a cross-reference to an even earlier entry, "accidie", a term borrowed from French, meaning "physical or mental slothfulness".<sup>14</sup> This kind of cross-referencing appears most often in the *OED*, with only one instance in *Merriam-Webster*, that of "conurbation". Figure 2 demonstrates this disparity, along with the many components which are used only by the *OED*, and which therefore serve to make the British-English dictionary significantly more comprehensive than its American-English counterpart.

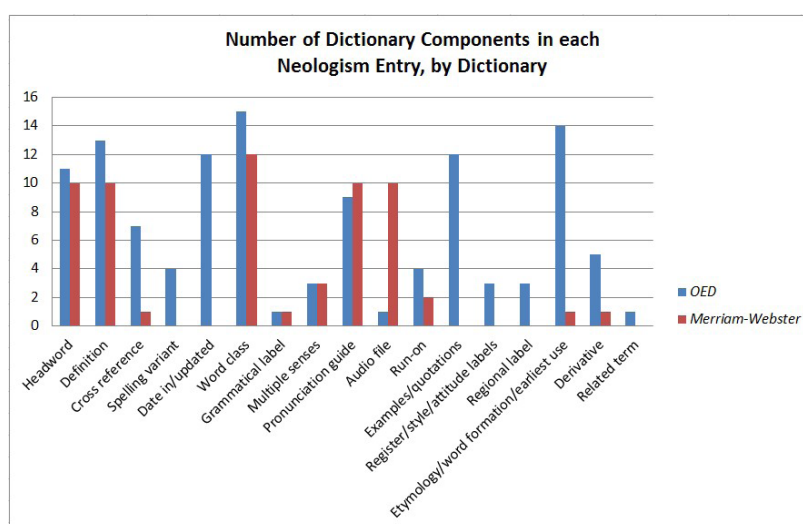


Figure 2: Total number of neologisms featuring each dictionary component, broken down by dictionary.

<sup>12</sup> This means that the links from this paper to example entries are able to show only how the entry appears now, not the layout and degree of information that was present at the time of data collection. However in each case, the component or aspect being exemplified demonstrates the same characteristics, albeit in a different style, as it did previously.

<sup>13</sup> See <https://www.merriam-webster.com/dictionary/warrantless>.

<sup>14</sup> See <http://www.oed.com/view/Entry/1068#eid39596420>.

The *Merriam-Webster* cross-reference from “conurbation” was to other dictionaries in the range, adding to the confusion surrounding this entry since one of those was the language learners’ dictionary, whose definition was already included without a label. As with all of the other neologism entries in the US dictionary, this has changed following the revisions to the site as a whole. In the *OED*, meanwhile, in most cases cross-references linked from an entry to the dictionary’s thesaurus function, as well as to alternate spellings, derivatives, related terms and grammatical information, thus making it easy to find a broad range of information, as Table 5 demonstrates.

Table 5: Cross-referencing in the *OED*.

Neologisms Containing Cross-Reference Links	Alternate Spelling	Thesaurus	Derivative	Related Word	Grammatical Information
acedia			Y		
bogof	Y				
hubristic		Y			
promissory note		Y			
tenebrous		Y			
upskill		Y		Y	Y
warrantless		Y	Y		Y

The *OED* has now also added frequency indicators to all of its neologism entries (except “run-ons”), to show how widely each word is used. This is another “non-standard” dictionary component, and one which *Merriam-Webster* does not use. Like the “audio file” change below, this is a dictionary-wide amendment, but it reinforces the pattern that the *OED* uses more non-standard components than *Merriam-Webster* does, and is further evidence of the *OED*’s more comprehensive approach to neologisms. Indeed, of the four non-standard components originally included in one or both of these dictionaries (“inclusion date”, “audio file”, and having “derivatives” or “related terms”), the *OED* used all except one (“audio files”) more than its American counterpart (see Figure 2). In the last few years the *OED* has added audio files to most of its neologism entries, while *Merriam-Webster* has made no change to its use of non-standard dictionary components, instead continuing to focus largely on the most basic of the standardized ones (see Figure 2).

One key addition it has made, however, is the introduction of examples, which it had previously not included in its neologism entries. These are now present for seven of its 12 entries: “conurbation”, “earworm”, “frenemy”, “greenwashing”, “promissory note”, “tenebrous” and “waterboarding”. The *OED*, meanwhile, carries “quotations” rather than “examples”. These quotations comprise attributed samples of the use of the word, and are generally drawn from Citation Banks made up of the kind of citations used in the attestation process described in section 2.2. They are also increasingly being complemented by examples from corpora (Atkins & Rundell 2008: 455). These quotations are present in all 15 of the *OED*’s neologism entries. *Merriam-Webster*’s new examples are generally, but not always, attributed, although in line with its position as a slightly more “current” dictionary it chooses to use the more common heading “examples”, and those included are usually quite recent and to be found on the internet.<sup>15</sup> However the presence of examples is still not consistent across *Merriam-Webster*, and this may be because their inclusion is so new. As mentioned above, it is believed that this change happened only very recently, and that prior to November 2017 these entries may have still been lacking examples of any kind. It may be that their addition is an ongoing process; certainly just two years ago none of the neologisms examined here were exemplified in *Merriam-Webster* online.

<sup>15</sup> See for example <https://www.merriam-webster.com/dictionary/frenemy>.



In terms of the quality of the components appearing in these two dictionaries, it is interesting to note that while most of the neologisms include pronunciation guidance, in the *OED* this takes the form of the widely-recognized IPA or International Phonetic Alphabet system, while *Merriam-Webster* appears to use its own bespoke guidance system. The symbols used are different both from IPA and from the usual alternative, SAMPA (Speech Assessment Methods Phonetic Alphabet). This makes the *Merriam-Webster* dictionary harder to use, since no key is given to the pronunciation symbols it uses. The *OED*, meanwhile provides links from the IPA guidance in each entry to a detailed IPA chart (although these are not hyperlinks (underlined and a different color), meaning it may be only by accident that users discover their presence). The *OED* is not without its problems with regard to pronunciation, however. Where a word can be pronounced in several different ways (sometimes giving multiple IPA versions in the two varieties of English), it occasionally gets them the wrong way around. For example, it is my view that the US- and the British-English IPA has been transposed in the entry for “frenemy”.

One of the dictionary components that we might expect to be of particular use in the current discussion would be that of regional labels. We might expect words which are considered especially “British” or “American” to carry the relevant geographical label. Indeed in the *OED*, both “warrantless” and “waterboarding” are labelled as being mainly used in the United States, yet *Merriam-Webster* does not carry similar labels, making its entries again slightly less comprehensive. Indeed none of its neologism entries uses regional labels. “Bogof” is labelled in the *OED* as being mainly used in Britain, and this term does not appear at all in *Merriam-Webster*. This lack of acceptance into the American-English dictionary may perhaps indicate that a different term is used in the US to promote “buy one get one free” deals in shops, or indeed that a whole raft of US-specific marketing and promotional vocabulary exists, making “bogof” irrelevant for a US audience. Perhaps similar reasons may explain why “rewilding” appears in the *OED* but not in *Merriam-Webster*. It may be that the idea of reclaiming land and returning it to its natural habitat has yet to gain currency in the US, or perhaps that a different term is used, one which has yet to enter British-English. Without further specific research on this topic it is impossible to know; however, if this is the case it offers some resistance to the widely-held belief that language flows from the United States to the United Kingdom, not least because of the prevalence of American television programming and music on British shores (see, for example, Anderson 2017). While these new words have not gained a place in this US dictionary, they have held their own and become sufficiently well established to gain acceptance into the *OED*.

While geographical information may not be as extensive as one would hope, *Merriam-Webster* has increased the amount of “etymological/word formation” information included in its neologism entries. New information has been added to eight neologisms which previously had none: “conurbation”, “earworm”, “frenemy”, “greenwashing”, “promissory note”, “tenebrous” and “waterboarding”. Finally, the *OED* has added several new words and senses to its webpages. A new sense for “earworm” has been included (a tune that you cannot get out of your head),<sup>16</sup> and entirely new entries have been added for “bankster” (a dishonest banker)<sup>17</sup> and “hyperlocal” (something extremely local).<sup>18</sup> These entered in March, December and June of 2015, respectively. Thus, while both dictionaries are clearly continuing to develop their neologism entries, it appears that *Merriam-Webster* is still (intentionally or simply due to a lack of data) limiting its new-word entries to only the more basic of components. The *OED*, meanwhile, is continuing to expand its “neologic” offering.

16 See <http://www.oed.com/view/Entry/318883?rskey=sCYLzS&result=1&isAdvanced=false#eid>.

17 See <http://www.oed.com/view/Entry/425803?rskey=SDWaju&result=2&isAdvanced=false#eid>.

18 See <http://www.oed.com/view/Entry/34824186?redirectedFrom=hyperlocal#eid>.

## 4 Conclusion

As Former Chief Editor of *OED* John Simpson points out, “neologisms are a window both on language change and continuity” (2007: 147). From the discussion above it is clear that, despite the availability of “standardized” dictionary components which should enable information on new words to be presented uniformly and recognizably across all entries, there are significant differences in how the *OED* and *Merriam-Webster* dictionaries approach and treat new words. *Merriam-Webster* appears to take a very cautious approach to these terms, limiting its entries to basic components such as definitions, senses, word classes and pronunciation guidance. The *OED*, meanwhile, presents a broader and more comprehensive range of information in its entries, including geographical details, date information and quotations. Both dictionaries have experienced site-wide changes in the past few years, and for *Merriam-Webster* these mean that it is beginning to catch up with the *OED*, particularly in the areas of dating and examples. However it still has some way to go, and for the time being, it is the “historical” dictionary, and not the more “current” one which is leading the way with a more comprehensive approach and treatment of new words.

## References

- Algeo, J. (1993) Desuetude among New English Words. In *International Journal of Lexicography* 6(4), pp. 281-293
- Anderson, H. (2017) *How Americanisms are Killing the English Language*. In BBC – Culture. Accessed at <http://www.bbc.com/culture/story/20170904-how-americanisms-are-killing-the-english-language> [05/03/18]
- Atkins, BTS. & Rundell, M. (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press
- Creese, S. (2017) Lexicographical Explorations of Neologisms in the Digital Age. Tracking New Words Online and Comparing Wiktionary Entries with ‘Traditional’ Dictionary Representations. PhD thesis. Coventry University, Coventry, UK
- Fischer, R. (1998) *Lexical Change in Present-Day English. A Corpus-Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Gunter Narr Verlag
- Francel, M. (2011) Neolexia. In *Nature Chemistry*, 3, pp. 417-418
- Hanks, P. (2013) Patrick Hanks. In *OED Symposium 2013 Newsletter, Issue 7. OED Symposium*, 1 August 2013. Oxford: University Press, Oxford, UK
- Kerremans, D. (2015) *A Web of Words. A Corpus-Based Study of the Conventionalization Process of English Neologisms*. Frankfurt: Peter Lang GmbH
- Lehrer, A. (2003) Understanding Trendy Neologisms. In *Rivista di Linguistica*, 15(2), pp. 371-384
- Ludwig-Maximilians Universität München (n.d.) *NeoCrawler List*. Accessed at <http://www.neocrawler.de/crawler/html/> [1 February 2014]
- Merriam-Webster Dictionary (2018a) *About Us*. Accessed at <https://www.merriam-webster.com/about-us/faq> [25/03/18]
- Merriam-Webster Dictionary (2018b) *How does a Word get into a Merriam-Webster Dictionary?* Accessed at <https://www.merriam-webster.com/help/faq-words-into-dictionary> [25/03/18]
- Merriam-Webster Dictionary (2015). *How does a Word get into a Merriam-Webster Dictionary?* Accessed at [https://www.merriam-webster.com/help/FAQ/words\\_in.html](https://www.merriam-webster.com/help/FAQ/words_in.html) [19/10/15] (NB this webpage has since been changed)
- Merriam-Webster Dictionary (2014) *Promissory note* Accessed at <http://www.merriam-webster.com/dictionary/promissory-note> [31/08/14] (NB this webpage has since been changed)
- Mitchell, RL. (2008) My Word: Why Google is in the Dictionary but AJAX isn’t. In *ComputerWorld*, 27 October 2008, pp. 32-34
- Moon, R. (2008) Lexicography and Lexical Creativity. In *Lexikos* 18 (AFRILEX-reeks/series 18), 131-153
- Neuman, Y., Nave, O. & Dolev, E. (2010) Buzzwords on Their Way to a Tipping Point: a View from the Blogosphere. In *Complexity*, 16(4), 58-68
- Oxford University Press (2018) *The Oxford New Words Corpus*. Accessed at <https://en.oxforddictionaries.com/explore/oxford-new-words-corpus> [15/03/18]

- Oxford University Press (2018a) *Reading Programme*. Accessed at <https://public.oed.com/history-of-the-oed/reading-programme/> [24/03/18]
- Oxford University Press (2018b) *What is a Corpus?* Accessed at <https://en.oxforddictionaries.com/explore/what-is-a-corpus> [22/03/18]
- Oxford University Press (2018c) *Greenwashing. Entry History*. Accessed at <http://www.oed.com/view/Entry/249122?rskey=aA0io4&result=2&isAdvanced=false#eid> [28/03/18]
- Oxford University Press (2018d) *How Do We Decide Which Words are Included in English Dictionaries*. Accessed at <https://www.oxforddictionaries.com/our-story/creating-dictionaries> [29/03/18]
- Oxford University Press (2015) *Not like Other Dictionaries: A Brief Introduction to the OED* Accessed at <http://public.oed.com/about/frequently-asked-questions/#newword> [19/10/15] (NB this webpage has since been changed)
- Renouf, A. (2007) Tracing Lexical Productivity and Creativity in the British Media: “The Chavs and The Chav-Nots”. In J. Munat (ed) *Lexical Creativity. Texts and Contexts*. Amsterdam: John Benjamins Publishing, pp. 61-89
- Renouf, A. (2013) A Finer Definition of Neology in English. The Life-Cycle of a Word. In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling (eds.) *Corpus Perspectives on Patterns of Lexis*. Amsterdam: John Benjamins Publishing, pp. 177-208
- Simpson, J. (2007) Neologism: The Long View. In *Dictionary. Journal of the Dictionary Society of North America*, pp. 146-148
- Simpson, J. (2003) The Production and Use of Occurrence Examples. In P. van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Amsterdam: J. Benjamins, pp 260-272
- Sinclair, J. (2004) Corpus and Text – Basic Principles’. in *Developing Linguistic Corpora: A Guide to Good Practice*. M. Wynne (ed.) [online] Oxford: Oxbow Books. Accessed at: [www.ahds.ac.uk/linguistic-corpora/](http://www.ahds.ac.uk/linguistic-corpora/) [21/02/13], pp. 5-24
- Sullivan, J. (2017) *How the Dusty Merriam-Webster Dictionary Reinvented itself. Bigly*. In *Globe Magazine*. Accessed at <https://www.bostonglobe.com/magazine/2017/02/14/how-dusty-merriam-webster-dictionary-reinvented-itself-bigly/Knytvd8ZHBjHismFi4fK9N/story.html> [08/03/18]
- Weiner, E. (2009) The Electronic OED: The Computerization of a Historical Dictionary. In AP. Cowie (ed.) *The Oxford History of English Lexicography, Volume 1 General-Purpose Dictionaries*. Oxford: Oxford University Press 378-409

## Acknowledgements

With thanks to Daphne Kerremans and the team at Ludwig-Maximilians University, Munich, for use of the *NeoCrawler* and its neologism list.



# New German Words: Detection and Description

**Annette Klosa, Harald Lüngen**

*Institut für Deutsche Sprache, Mannheim*

*E-mail: klosa@ids-mannheim.de, luengen@ids-mannheim.de*

## Abstract

In this paper, we discuss an efficient method of (semi-automatic) neologism detection for German and its application for the production of a dictionary of neologisms, focusing on the lexicographic process. By monitoring the language via editorial (print and online) media evaluation and interpreting the findings on the basis of lexicographic competence, many, but not all neologisms can be identified which qualify for inclusion in the *Neologismenwörterbuch* (2006-today) at the Institute for the German Language in Mannheim (IDS). In addition, an automated corpus linguistic method offers neologism candidates based on a systematic analysis of large amounts of text to lexicographers. We explain the principles of the corpus linguistic compilation of a list of candidates and show how lexicographers work with the results, combining them with their own findings in order to continuously enlarge this specialized online dictionary of new words in German.

**Keywords:** detection of neologisms, description of neologisms, corpus linguistics, lexicography

## 1 Introduction

Entries on new words in general dictionaries of German or in specialized dictionaries of neologisms in German attract many users, as new words pose problems regarding their meaning and usage, their grammatical features, their orthography, and their pronunciation. Neologisms are also often subject of language criticism in German: while some speakers consider them as unnecessary additions to German mainly taken from other languages (with English often as the most common donor language), others realize their importance in filling lexical gaps or providing means of enriching the language with a multitude of possibilities of expression. Generally, language change, which we can see most prominently at the lexical level, is a topic of high interest for all speakers (cf. Lemnitzer 2010: 65).

Therefore, a central issue in lexicography (for German as well as other languages) is to find new lexemes and to identify new meanings for existing lexemes. By monitoring the language via editorial media evaluation and interpreting the findings on the basis of lexicographic competence, many but not all neologisms can be identified. Only automated methods of corpus linguistics can provide a systematic analysis of large amounts of text, offering neologism candidates to lexicographers. In this paper, we discuss our method of neologism detection for German and its application for the production of a dictionary of neologisms in this language.

In the following (cf. Section 2), we present the *Neologismenwörterbuch* (2006-today; cf. Steffens 2017) at the Institute for the German Language in Mannheim (IDS). In Section 3, we discuss related work (other German dictionaries of new words and other methods of detection of neologisms). We describe our semi-automatic method of detection of neologism candidates in Section 4, which we evaluate in Section 5. In this section, we also focus on the impact of this method for both our dictionary as well as the corpus tool. In a short outlook (cf. Section 6), we then discuss the possibilities of a more extensive use of corpus linguistic findings in our online dictionary of neologisms in the future.



## 2 The Neologismenwörterbuch

Dictionaries of new words are specialized dictionaries describing the meaning and usage of lexemes in a specific language which became part of the vocabulary at a certain time (for more details cf. Barnhart & Barnhart 1990; Lemnitzer 2010; Wiegand 1990). The *Neologismenwörterbuch* published online by IDS Mannheim is a typical example of this type of dictionary. It covers new words and new meanings established in the past thirty years. The online publication in the dictionary portal OWID (Online-Wortschatz-Informationssystem Deutsch) at IDS Mannheim allows for the continuous addition of new entries. For the decades of 1991-2000 and 2001-2010, print dictionaries are also available (Herberg et al. 2004; Steffens & al-Wadi 2015). The lexicographic concept for these dictionaries goes back to the late 1980s (cf. Heller et al. 1988; Kinne 1989) and 1990s (cf. Herberg 1997 and 1998), when German lexicography had not yet embraced the potential of corpus linguistics. Only in the first decade of the 21st century was a corpus linguistic method of neologism detection developed to supply lexicographers working on the *Neologismenwörterbuch* with candidates for inclusion in the dictionary in combination with candidates taken from the project's own editorial media evaluation (cf. Section 4.1).

The *Neologismenwörterbuch* comprises entries on single words (e.g., *Avatar*), multi-word expressions (e.g., *in der Pipeline*), and new elements of word formation (e.g., [...]*holic*). Not only new words, but also new meanings for existing words in German are described (e.g., *texten* 'send a (short) text message in electronic media'). Proper names are basically excluded from the lemma list in the *Neologismenwörterbuch*; only derivatives with a proper name as their base are included in the dictionary, for example *twittern* ('to send a Twitter message'), but not *Twitter*. As many compounds and derivatives in German are semantically transparent, i.e., can be fully interpreted based on the knowledge of the meaning of their components, they are not lemmatized in the *Neologismenwörterbuch*, either, for example *Eurokrise*, 'crisis because of the weak Euro'.

Lexicographic information comprises etymology, orthography, pronunciation, meaning, usage, grammar, word formation, encyclopedic information, illustrations, and frequency in the corpus. The dictionary aims at covering all neologisms established throughout the last two as well as the current decade, describing each neologism as new for each decade, respectively. Table 1 gives information on the number of entries in the *Neologismenwörterbuch* in March 2018. All entries in the dictionary meet the following definition of "neologism" in our project: A neologism is a lexical unit or a meaning which emerges in a communication community in a specific period of time of language development, which diffuses, is generally accepted as language norm, and which the majority of speakers perceives as new for some time. To sum up: neologisms in our project are not nonce words, but are defined as fully lexicalized lexemes. Thus, only in retrospect it is possible to decide which words are neologisms and which are not.

Table 1: Numbers of entries in the *Neologismenwörterbuch* in March 2018

All entries	more than 1.800
Neologisms from 1991-2000	over 1.000
Neologisms from 2001-2010	almost 700
Neologisms since 2011	almost 150
New lexemes	almost 1.550
New elements of word formation	almost 20
New meanings	over 160
New multi-word units	almost 120
Other new lexemes (synonyms, other sense-related words, derivations, compounds, etc.) contained in entries and accessible via list	almost 5.000

Our definition contains several criteria which cannot be easily operationalized: How do we measure whether a new word is generally accepted or that a majority of speakers perceives it as new? For each neologism candidate, a decision on its possible inclusion in the dictionary has to be based on an individual analysis of the data available. Not only do we look at the number of years and/or months since the lexeme has shown up in the German corpora and the development of its frequency, but also at the way in which it is being used. There are several textual indicators for words which are not yet fully lexicalized (cf. Lemnitzer 2010: 69): quite often, they are used in quotation marks or are followed by short definitions. In particular words borrowed from other languages initially do not exhibit a full declination paradigm in German; nouns often show different genders, before they settle for one grammatical gender. Pronunciation as well as orthography show a lot of variation in the beginning as well. Moreover, only fully lexicalized words in German can enter into word formation products in combination with Germanic as well as loan morphemes. Candidates for inclusion in the dictionary, whether from our editorial reading or detected automatically, are evaluated according to these criteria.

### 3 Related Work

#### 3.1 German Dictionaries of New Words

German dictionaries of new words are a fairly recent development: the first German print dictionary of neologisms was published at the beginning of the new century at IDS Mannheim (Herberg et al. 2004), shortly followed by the *Neologismenwörterbuch* online (in 2006; cf. section 2). In 2007, a first, strictly corpus-driven print dictionary (Quasthoff 2007) was published. It contains almost 2,300 entries, giving short definitions, corpus examples, information on the subject area, and frequency diagrams for each head word. Neologisms in this dictionary are words whose frequency has increased significantly between 2000 and 2006. However, not all of these are proper new words according to our definition, as quite a number of the entries comprise nonce words

Since 2000, the website Wortwarte (2000–today) has collected and published German neologisms automatically extracted from newspaper web pages and other online sources. Here, neologisms are defined as new, but not yet fully lexicalized lexemes. The Wortwarte-dictionary aims at recording new words “*in statu nascendi*” (Lemnitzer 2010: 67). In 2017, for example, approximately 2,900 new words were registered with short grammatical information and one corpus example.

Of course, general dictionaries of German also update their lists of headwords, adding new entries for new editions. The latest edition of *Duden – Die deutsche Rechtschreibung* (2017) contains more than 5,000 new entries (as part of 145,000 entries in total) according to Duden publishing house. Many of these are not neologisms in our definition, but, for example, transparent compounds or proper names.

#### 3.2 (Semi-)Automatic Detection of New Words

Most approaches to a corpus-based detection of neologisms use press corpora and to some extent also web-specific corpora (Lemnitzer 2010; Quasthoff 2007). Wortwarte (2000–today) automatically compares a current word list (i.e., a word list of the day derived from press and web texts) with a reference word list built from older corpora and previous word lists, thus obtaining a list of new words of the day. Subsequently, a lexicographer decides which words from the generated candidate list is a proper neologism according to the used definition. Following Falk et al. (2014), this approach may be classified as based on exclusion lists.

Another type of approach uses corpora which are partitioned into sub-corpora by year, and then tries to identify typical frequency timelines of neologisms, with a rise of frequency in the present and with minimum frequency conditions to model its establishment in the language and to exclude nonce words. Such an approach has been used by Quasthoff (2007), and our approach falls into this category, too, cf. Section 4.2. An alternative approach is exemplified by Falk et al. (2014), who combine exclusion lists with a supervised machine-learning approach in which typical features of the linguistic context of neologisms are extracted from (French) press corpora. The authors note that the advantage of their approach is that it does not heavily rely on large, diachronic, time-annotated corpora.

All of these approaches are semi-automatic in that candidate lists are generated which must be post-processed manually in order to select the proper neologisms according to the respective definitions.

## 4 Finding Neologism Candidates

### 4.1 Editorial Evaluation of Print and Online Media

Since starting the *Neologismenwörterbuch*, lexicographers working on the dictionary have collected candidates for new words through intensive daily browsing through a number of print and (later also) online media. Candidates are entered into a database with information on the date of the first sighting (or hearing), the source, and a first frequency result of a search in our corpus or on the web. A tentative short definition and links to other sources are added, if possible. All candidates are classified according to the decade they entered the German language. Some candidates remain in the database for further monitoring for several years, other candidates directly become entries in the dictionary.

In addition to the dictionary staff, a small number of users of the *Neologismenwörterbuch* send in suggestions for new entries. In the future, we plan to systematize collaboration with users in the editorial evaluation of print and online media by introducing an appeal to readers to send in their findings via a form.

### 4.2 Quantitative Method

Our method for the quantitative detection of neologism candidates was originally developed by Holger Keibel and described in detail in the technical report Keibel et al. (2010). According to the setting of the *Neologismenwörterbuch*, it is designed to identify neologisms which are associated with one specific decade, and which are already advanced in their lexicalization process, i.e., excluding ad-hoc formations and nonce words. To achieve this, frequency timelines of all words are compared in corpus data from two adjacent time periods A and B. Those words in the more recent period B which exhibit a typical timeline are subject to further filtering processes aimed at reducing the remaining non-neologisms such as names and regionalisms; the resulting list is considered the “neologism candidate list” that should be more closely inspected by the lexicographer.

In the recent application of the method that we evaluated for this paper, we have compared the corpus data from the period 2010-2015, representing the current decade, with data from the previous decade 2000-2009 in order to discover neologism candidates of the current decade. (The delimitation of decades in this application is the one originally applied by Keibel et al. (2011) and deviates slightly from the one used in the *Neologismenwörterbuch* (2001-2010 and 2011-2020). This was adjusted in later applications.)

#### 4.2.1 *Corpus and Frequency List*

As corpus data, we use a virtual corpus with over three billion tokens of press text which in the application described in Section 4.2.3 and evaluated in Section 5, spanning the years 2000-2015, and which is derived from the German Reference Corpus DeReKo. DeReKo is hosted by IDS and is the largest linguistic text archive for the German language, currently with 42 billion tokens (Institut für Deutsche Sprache 2017). The bulk of DeReKo is made up of press corpora, but many other genres such as fiction, science, specialized texts, debate protocols, and computer-mediated communication, are represented, too. The newspaper titles represented in the current project corpus were selected because: a.) most of them are available in DeReKo for most of the years in the period under scrutiny, and b.) they are well distributed across the four language regions North, East, South, and Southwest of Germany.

Initially, a huge frequency list of all word forms occurring in the project corpus representing both periods A and B is built. The quantitative method strictly investigates word forms only, i.e., unclassified corpus tokens. Base forms or lemmata are not considered, primarily because automatic lemmatization tools will frequently fail to lemmatize new and unknown word forms, but also because it is interesting to keep track of possibly differing frequency properties of the different inflectional forms or spelling variants.

#### 4.2.2 *Filtering of the Frequency List*

From the full frequency list, obvious non-lexical items such as numbers or URLs are identified based on graphematic criteria and removed in a first filtering step. The second step consists of filtering out all words with an overall absolute frequency below a certain minimum (currently 10). In a third step, those forms are filtered out that do not conform to a set of quantitative criteria which, amongst other things, define a maximum frequency in period A, a minimum frequency in period B, and minimum frequencies for the years after the year of appearance and the peak year. These criteria are to characterize the typical timeline of a neologism of period B and to filter out other words, including short-lived, ad-hoc, non-lexicalized formations and usages.

The resulting list still contains many names and regionalisms that conform to the quantitative criteria; hence in the fourth filtering step we try to filter out regionalisms by removing word forms that show a bias of occurrence for one of the four sub-corpora representing German language regions according to the DP (deviation of proportions) dispersion measure by Gries (2008). In a fifth filtering step, names are filtered by sending KWIC result lines with candidates derived from the corpus to the Stanford Named Entity Recognizer (Finkel et al. 2005). Filtering is applied in a conservative fashion, because in the lexicographic context a maximization of recall is considered more important than a maximization of precision. The latter would bear a higher risk of losing relevant candidates in the filtering processes.

The resulting list is our “neologism candidate list”, which still contains many obvious non-neologisms, mostly names that had not been correctly identified by the NER tool. Keibel et al. (2010) point out that these could be filtered efficiently manually even by someone who is not an expert before the final candidate list is analyzed by a lexicographer.

#### 4.2.3 *Recent Application and Candidate List*

In one of our recent applications of the method to detect neologisms of the current decade, the corpus spanned 2000-2009 (period A) and 2010-2015 (period B) and together contained over three billion word form tokens yielding around 10 million different word form types. The resulting CANDIDATE-LIST\_2016 contained 5,483 word form types (neologism candidates).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Pegida	14845 0%		0.47625	DE-S	28523828	23955.54	2015	12948	2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1897	12948
2	dapd	52487 1.96%		0.54312	DE-S	100781217	21172.48	2010	24053	2012	0	0	0	0	0	0	0	0	0	3	6458	20799	24053	1105	51	18
3	Fracking	7313 55.45%		0.36442	DE-S	14051574	11801.89	2011	2421	2013	0	0	0	0	0	0	0	0	0	0	3	173	1237	2421	2024	1455
4	iPad	10736 6.67%		0.41906	DE-S	20607144	8661.82	2011	2614	2012	0	0	0	0	0	0	1	0	0	0	2160	2362	2614	1410	1181	1008
5	Varoufakis	3890 100.00%		0.46889	DE-S	7474.46	6278.53	2013	3838	2015	0	0	0	0	0	0	0	0	0	0	0	0	48	2	2	3838
6	apn	6776 1.01%		0.54944	DE-S	12999.16	5466.89	2011	6764	2010	0	0	1	0	0	0	0	0	0	0	6764	9	1	0	1	0
7	Instagram	3139 50.98%		0.41574	DE-S	6031.45	5066.71	2012	1531	2015	0	0	0	0	0	0	0	0	0	0	1	2	434	428	743	1531
8	Mietpreisbremse	2995 19.61%		0.5192	DE-S	5754761	4834.35	2014	1483	2015	0	0	0	0	0	0	0	0	0	0	0	0	0	560	952	1483
9	Eurokrise	5219 37.62%		0.39701	DE-S	10007.99	4210.7	2011	1969	2012	1	0	0	0	0	0	0	0	0	0	290	1067	1969	1096	394	402
10	Selfie	2491 6.25%		0.40586	DE-SW	4786348	4021.1	2014	1567	2015	0	0	0	0	0	0	0	0	0	0	0	0	0	78	846	1567
11	Bundesfreiwilligendienst	2424 0.00%		0.50295	DE-SW	4676925	2020.12	2011	750	2011	0	0	0	0	0	0	0	0	0	0	60	750	617	242	242	406

Figure 1: Top of CANDIDATE-LIST\_2016, including frequencies by year, LLF, M-score (cf. Keibel et al. 2010), peak year, probability of being a name or a regionalism, and ranked according to the M-score.

The initial word lists derived from our corpora and the full resulting CANDIDATE\_LIST\_2016 will shortly be published on the research data server of our institution<sup>1</sup> so that others can try out alternative approaches on the same data.

## 5 Evaluation of the Quantitative Method

### 5.1 Linguistic Evaluation of the Candidate List and Comparison with Editorial Media Evaluation

Lexicographers working on the *Neologismenwörterbuch* annotated the first 500 false positives (FPs) in CANDIDATE-LIST-2016 with the six categories *Proper Name*, *Semantically Transparent*, *Occult Neologism*, *Inflectional Form*, *Spelling Variant*, and *Other*. Table 2 shows an extract of annotated datasets 1-15 from the CANDIDATE-LIST\_2016 with explanatory comments.

Among the top 500, 219 of the false positives were annotated as proper names, 179 as semantically transparent, 79 as “other”, 13 as inflectional forms, and 10 as occult neologisms. CANDIDATE LIST\_2016 was annotated while large parts of the entries for the first decade and the first half of the second decade of the 21st century had already been published online. Thus, quite a number of candidates were already contained in the *Neologismenwörterbuch* either as a full lemma or as a sub-lemma (e.g., as a word formation product, or as a less frequent synonym).

The majority of candidates which had not already been found through editorial media evaluation and accordingly included in the dictionary are either proper names (e.g., *Instagram*) or semantically transparent lexemes (e.g., *Eurokrise*). At the moment, proper names are basically excluded from the lemma list in the *Neologismenwörterbuch*; only derivatives with a proper name as their base are included in the dictionary, for example *twittern* (‘to send a Twitter message’), but not *Twitter*, *Youtuber* (‘somebody watching or somebody producing YouTube videos’), but not *Youtube*. As many compounds and derivatives in German are semantically transparent, i.e., can be fully interpreted based on the knowledge of the meaning of their components, they are not lemmatized in the *Neologismenwörterbuch*, either, for example *Eurokrise* ‘crises because of the weak Euro’. Only the top 500 words were annotated in CANDIDATE-LIST\_2016, because further down the list the number of proper names and semantically transparent lexemes increases considerably, while the frequency of each candidate in our corpora decreases.

<sup>1</sup> <https://repos.ids-mannheim.de/>



Table 2: Extract from annotated list of candidates for dictionary of neologisms; abbreviations: pn = proper name, st = semantically transparent; oth = other (e.g., abbreviations)

Candidate	Lemma/Sub-lemma in Dictionary	Category	Comment
<i>Pegida</i>	no	pn	name of a political group
<i>dapd</i>	no	oth	abbreviation
<i>Fracking</i>	yes		'hydraulic fracturing'
<i>iPad</i>	no	pn	product name
<i>Varoufakis</i>	no	pn	family name
<i>apn</i>	no	oth	abbreviation
<i>Instagram</i>	no	pn	app name
<i>Mietpreisbremse</i>	no	st	'political measure for slowing down the increase of rents'
<i>Eurokrise</i>	no	st	'crisis because of the weak Euro'
<i>Selfie</i>	yes		engl.: <i>selfie</i>
<i>Bundesfreiwilligendienst</i>	no	pn	name of German nationwide voluntary service
<i>Grexit</i>	yes		engl.: <i>grexit</i>
<i>iOS</i>	no	oth	abbreviation
<i>Fiskalpakt</i>	yes		'fiscal pact'
<i>Kobane</i>	no	pn	geographical name

Therefore, these candidates are presumably less relevant for inclusion in our dictionary. As of 2017, a candidate list has been compiled and evaluated annually so that the number of words which are already part of the dictionary should decrease in the future. The editorial media evaluation is still needed to supplement this method, because we also include multi-word expressions, new elements of word formation, and new meanings in the dictionary which currently cannot be found automatically. Also, many new lexemes in our dictionary had not been discovered by the corpus-linguistic method explained above (false negatives). The following list gives an overview of entries from the first half of the second decade of the 21st century in the *Neologismenwörterbuch* showing which ones were detected in our editorial media evaluation (grey) and which ones were also contained among the top 500 of the automatically compiled CANDIDATE-LIST\_2016 (black):

- New lexemes: *3-D-Drucker, Antänzer, Arabellion, Bestellbutton, BFD, Biodeutscher, Blitzmarathon, Blockupy, Bodycam, Boxspringbett, Brexit, BRICS, Bubble-Tea, Bufdi, Buttonlösung, Cakepop, Chia, Chiasamen, Clickworker, Craftbier, Cross-fit, Crowdfunding, Crowdworker, Crowdworking, Cybergrooming, Darknet, Doodle, Doodleliste, Emoji, Entscheidungslösung, ESM, Facebookparty, Fairteiler, Fakeshop, Faszientraining, Femenaktivistin, Fingerwisch, Fiskalpakt, Fitnessarmband, Flexiquote, Flexirente, Flexitarier, Foodtruck, Fotobombe, fracken, Fracking, Freistoßspray, Frutarier, Fukushima-Effekt, Garagengold, Gettofaust, Glamping, Googlebrille, Grexit, GroKo, Guerillastricken, Hashtag, Helikoptereltern, Hipsterbart, Homestaging, Hugo, Hygieneampel, Inklusionsklasse, Jahnbehörde, Kampfhradler, Keniakoalition, Kinesiotape, Kryptohandy, Kryptoparty, leaken, Leo, Like, liken, Loopschal, Memoriamgarten, Mikrojob, Mingle, Netzpartei, Occupybewegung, Pflege-Bahr, Phablet, Pinkifizierung, Pop-up, Pop-up-Restaurant, QR-Code, Reichweitenangst, Repaircafé, Retweet, retweeten, Scamming, Selfie, Selfiestick, Seniorazubi, Sexting, Shapewear, Shitstorm, Smart-TV, Smartwatch, Spot- ted-Seite Stadtgärtnern, Streetfood, Strickgraffito, Strickguerilla, Superfood, tinder, Tofutier, Upcycling, Vatileaks, veggie, Veggieday, Veggietag, Vöner, Webinar, WhatsApp ('short mes- sage'), whatsappen, Willkommensklasse*

- New meanings: *aufpoppen*, *Computeruhr*, *dampfen*, *Dampfer*, *Energiearmut*, *Flugmodus*, *Loop*, *stromern*, *Tunnel*, *wischen*
- New multi-word expressions: “*Gefällt mir*”, “*Gefällt mir*”-Button, *arabischer Frühling*, *falsche Neun*, *falscher Neuner*, *grüner Smoothie*, “*hätte, hätte, Fahrradkette*”, *Natural Running*, *Second Screen*, *vertrauliche Geburt*, *ziemlich beste [X]*

These findings by our lexicographic team shall lead to finding better parameter settings to use with the quantitative method for the detection of neologism candidates. The majority of the false positives clearly arise because the respective form is not identified as a proper name by the NER tool. We shall try to find a better setting of the tool in new experiments, but at the same time we expect this problem to be hard to eliminate. Note that semantically transparent words would count as proper neologisms in other approaches to automatic detection (cf. Section 3.2). Note also that inflectional forms of head words should rather count as true positives from the perspective of the quantitative method.

## 5.2 Evaluation of the Filter Criteria

We also extracted a reference list from the underlying database of the *Neologismenwörterbuch* containing all simplex words associated with the entries of the current decade 2011–2016, which as explained above had originally been compiled solely through the editorial media evaluation. This reference list contained 845 word forms which were mapped on 127 base forms in the database. Eight-one of the word forms were true positives, i.e., also contained in our CANDIDATE-LIST\_2016, and these in turn were associated with 51 different base forms in the database. Thus our CANDIDATE-LIST\_2016 yielded a recall of  $51/127 = 40\%$  in terms of simplex base forms. In view of the number of 5,483 items on our Result List 2016, the precision is of course far lower.

We removed those items from the reference list that, according to the database, represented sense relations of head words (e.g., *Radrowdy* ‘bike rowdy’ as a synonym of the headword *Kampfradler* ‘bike rowdy’), and those that represented word formations derived from or composed with head words (e.g., *Kampfradlerin* as a derivative of *Kampfradler*), and obtained a smaller reference list with 390 word forms strictly representing only either the base form, an inflectional form, or a spelling variant of a headword. This reduced reference list still contained 130 true positives, i.e., for the reduced reference list the recall remained the same by a minor difference. From this calculation we concluded that those words that had been classified by the lexicographers as morphological variants of head words or sense relations of headwords only, but not deserving headword status themselves, had received a secondary status by our quantitative method, too, which was a nice confirmation.

Next we wanted to know whether we would still obtain a reasonable recall if we cut off our ranked candidate list at some point. It turned out that if we cut off the list at 4,000 entries it would still contain 78 of the 81 true positives. Cutting of the CANDIDATE-LIST\_2016 at a higher point would deteriorate the recall, e.g., 69 true positives remaining when cutting off at rank 2500. We thus concluded that  $\frac{4}{5}$  or more of the CANDIDATE-LIST\_2016 would have to be taken into consideration in order to find nearly all neologisms contained in it. These figures were confirmed in later applications.

Another form of evaluation was a closer inspection of the False Negatives (FNs) to find out when and why they were falsely filtered out, with the ultimate goal of improving the performance of the filters. As pointed out above, FNs (i.e., type II errors) imply a lower recall and are considered more severe errors in the lexicographic context than false positives. Having identified 51 true positives in a reference list of 127 base forms, the number of false negatives (FN) was 76. It turned out that eight of the FNs had not occurred in the corpus in the first place, one had been filtered out by the graphemic cleaning procedure, eleven had been filtered because their absolute corpus frequency was under the required minimum of ten corpus hits, 44 were filtered out by the complex quantitative criterion, two

were filtered out by the regional dispersion filter, and finally, ten were removed by the proper name filter, cf. Table 3 for details.

The distribution illustrated in Table 3 shows that most (44) of the FNS are falsely filtered out by the complex quantitative criterion. Step 0 and Step 2 falsely filtered out 19 FPs – for these, we assume that the editorial media evaluation program had been ahead of the corpus-based method, i.e., the lexicographers had already discovered them before they showed up in our press corpora at a reasonable number of occurrences. We would assume that they will be included in a future candidate list when the corpora are extended by data from beyond 2015. Ten FNs were filtered out because they were falsely analyzed to be proper names by the named entity recognizer.

Table 3: When were the false negatives, i.e., neologisms that were in the corpus, but did not make it in the final CANDIDATE-LIST\_2016, filtered out?

Filtering Step	#	FNs removed (as base forms)
Step 0 Not in corpus	8	<i>3-D-Drucker, Doodleliste, Fukushima-Effekt, Jahnbehörde, Pflege-Bahr, Pop-up-Restaurant, QR-Code, Seniorazubi</i>
Step 1 Graphemic cleaning	1	<i>ESM</i>
Step 2 Absolute frequency < 10	11	<i>Chiasamen, Fakeshop, Garagengold, Gettofaust, Gruselclown, Guerillastricken, Mikrojob, Selfiestick, Strickgraffito, tindern, whatsappen</i>
Step 3 Quantitative filter	44	<i>aufpoppen, Bestellbutton, Biodeutscher, Bubble-Tea, Buttonlösung, Chia, Clickworker, Craftbier, Cybergrooming, dampfen, Dampfer, Doodle, Energiearmut, Facebookparty, Flexirente, Fotobombe, Googlebrille, Hipsterbart, Hugo, Keniakoalition, Kinesiotape, Kryptohandy, Leo, Like, Loop, Mingle, Netzpartei, Pinkifizierung, Pop-up, Reichweitenangst, Sexting, Shapewear, Spotted-Seite, Stadtgärtnern, Streetfood, Strickguerilla, stromern, Tofutier, Tunnel, Upcycling, veggie, Vöner, Webinar</i>
Step 4 Regional dispersion	2	<i>Antänzer, Memoriamgarten</i>
Step 5 Proper name	10	<i>Arabellion, BFD, Blockupy, BRICS, Bufdi, Darknet, Occupybewegung, Vatileaks, Veggieday, Whatsapp</i>
	$\Sigma = 76$	

## 6 Conclusion and Outlook

With respect to our quantitative method for detecting new words, the majority of false positives clearly occur because the respective form is not identified as a proper name by the NER tool. On the other hand, several false negatives were not found because they are falsely filtered out as proper names by the NER tool. We shall try to find a better setting of the NER tool in further experiments, but at the same time we expect this problem to be hard to eliminate using current NER technology.

The majority of false negatives are filtered out by the complex quantitative criterion describing the typical timeline of a lexicalized new word (cf. Table 3). This is contrasted by the already fairly low precision (more than 5,000 candidates) our application yielded. Nevertheless, we shall try to improve the quantitative filter settings through more experiments and by carefully adjusting them for each new application, i.e., new corpora and corpus partitions.

As shown in Table 2, many neologism candidates are semantically transparent compounds or derivatives. Instead of excluding these completely from our dictionary, we consider describing them in

short entries in the future with the following (reduced) lexicographic information: orthography, pronunciation, word formation, grammar, definition, a small number of corpus examples, encyclopedic information and illustration where necessary or possible. Short entries are also considered for other candidates that could enhance the *Neologismenwörterbuch* in the future:

- synonyms for already existing entries with a significantly lower frequency, e.g., *Generation Y* (headword) – *Y-Generation* and *Yps-Generation* (synonyms), *Millenial* (headword) – *Millenial-generation* and *Milleniumsgeneration* (synonyms)
- extended usage of lexemes, e.g., *teilen*: ‘to share’ – extended usage ‘to share in social media’
- phrases from other languages, e.g., *never ever*, *powered by*, *all you can eat*, *sharing is caring*
- proper names which are the base for new lexemes, e.g. *YouTube*, *WhatsApp*, *Twitter*, *Tinder*
- terminological lexemes entering the general language, e.g., *Koenzym*, *Coretraining*, *SWIFT-Code*, *Prokrastinationxxx*

Many of these possible new short entries for the *Neologismenwörterbuch* belong to the following discourses: sports, society, political measures, economy, media, transport, crime, and terrorism. As the *Neologismenwörterbuch* online already offers access to all entries via subject areas (cf. Figure 2), the expansion of the dictionary in this direction seems worth paying attention to.

	90er	Nuller	Zehner
Arbeitswelt/Bildung	Account Adresse		
Gesellschaft	at Attachment	App	
Soziales	Banner Bannerwerbung		
Demografischer Wandel	Barcode Beamer		
Politik	Bildschirmschoner		Bestellbutton Bezahlshranke
Wirtschaft/Handel	Bluetooth	Blog	
Banken/Finanzwesen	Brenner Browser Button	Blu Ray Blu-Ray-Disc	
Umweltschutz/Energie	CD-Brenner		
Computer/Internet/Technologie	Chat Chatraum Chatroom		Buttonlösung
Tätigkeiten mit Bezug auf Computer/Internet			
(Computer-)Kriminalität			
Telekommunikation			
Medien			

Figure 2: Access to entries in the *Neologismenwörterbuch* via subject area and listing of entries in their decade of emergence (<http://www.owid.de/docs/neo/gruppen.jsp>)

## References

- Barnhart, R., Barnhart, C. (1990). The Dictionary of Neologisms. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (eds.) *Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin/New York: de Gruyter, pp. 1159-1166.
- Duden – Die deutsche Rechtschreibung (2017). 27th edition. Ed. Dudenredaktion. Berlin: Dudenverlag.
- Falk, I., Dernhard, D. & Gérard, C. (2014). From Non Word to New Word: Automatically identifying Neologisms in French Newspapers. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC, The 9th edition of the Language Resources and Evaluation Conference, May 2014, Reykjavik, Iceland*.
- Finkel, J. R., Grenager, T. & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, USA, 2005*.

- Gries, St. Th. (2008). Dispersions and adjusted frequencies in corpora. In *International Journal of Corpus Linguistics*, 13(4), pp. 403-437.
- Heller, K., Herberg, D., Lange, C., Schnerrer, R. & Steffens, D. (1988). *Theoretische und praktische Probleme der Neologismenlexikographie. Überlegungen und Materialien zu einem Wörterbuch der in der Allgemeinsprache der DDR gebräuchlichen Neologismen*. Berlin: Zentralinstitut für Sprachwissenschaft.
- Herberg, D. (1997). Neologismen im allgemeinen Wörterbuch oder Neologismenwörterbuch? Zur Lexikographie von Neologismen. In K.-P. Konerding, A. Lehr (eds.) *Linguistische Theorie und lexikographische Praxis. Symposiumsvorträge*. Heidelberg 1996. Tübingen: Niemeyer, pp. 61-68.
- Herberg, D. (1998). Auf dem Weg zum deutschen Neologismenwörterbuch. In A. Zettersten, V. H. Pedersen & J. E. Mogensen (eds.) *Symposium on Lexicography VIII. Proceedings of the Eighth International Symposium on Lexicography May 2-4, 1996, at the University of Copenhagen*. Tübingen: Niemeyer, pp. 187-192.
- Herberg, D., Kinne, M. & Steffens, D. (2004). *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. In collaboration with E. Tellenbach and D. al-Wadi. Berlin/New York: de Gruyter.
- Institut für Deutsche Sprache (2017). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-II* (Release: 01.10.2017). Mannheim: Institut für Deutsche Sprache. Accessed at <http://www.ids-mannheim.de/DeReKo> [20/03/2018].
- Keibel, H., Hennig, S. & Perkuhn, R. (2010). *Effiziente halbautomatische Detektion von Neologismuskandidaten*. Technical Report IDS-KL-2010-01. Mannheim: Institut für Deutsche Sprache.
- Kinne, M. (1989). Endlich: Ein deutsches Neologismenwörterbuch. In *Der Sprachdienst*, 4, pp.115-116.
- Lemnitzer, L. (2010). Neologismenlexikographie und das Internet. In: *Lexicographica*, 26, pp. 65-78.
- Neologismenwörterbuch* (2006-today), in: OWID – Online Wortschatz-Informationssystem Deutsch. Mannheim: Institut für Deutsche Sprache. Accessed at <http://www.owid.de/wb/neo/start.thml> [20/03/2018].
- Quasthoff, U. (2007). *Deutsches Neologismenwörterbuch. Neue Wörter und Wortbedeutungen in der Gegenwartssprache*. de Gruyter: Berlin.
- Steffens, D. (2017): Vom Print- zum Onlinewörterbuch – Zur Erfassung, Beschreibung und Präsentation von Neologismen am IDS. In J. Dąbrowska-Burkhardt, L. M. Eichinger & U. Itakura (eds.) *Deutsch: lokal – regional – global*. Tübingen: Narr, pp. 281-294.
- Steffens, D., al-Wadi, D. (2015). *Neuer Wortschatz. Neologismen im Deutschen 2001-2010*. Mannheim: Institut für Deutsche Sprache.
- Wiegand, H. E. (1990). Neologismenwörterbücher. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (eds.) *Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin/New York: de Gruyter, pp. 2185-2187.
- Wortwarte (2000-today). *Die Wortwarte. Wörter von heute und morgen. Eine Sammlung von Neologismen*. Ed. by Lothar Lemnitzer. Accessed at <http://www.wortwarte.de> [20/03/2018].





# “Brexit means Brexit”: A Corpus Analysis of Irish-language BREXIT Neologisms in The Corpus of Contemporary Irish

**Katie Ní Loingsigh**

*Fiontar & Scoil na Gaeilge, Dublin City University*

*E-mail: katie.niloingsigh@dcu.ie*

## Abstract

The primary objective of this paper is to map the introduction and adoption of various Irish-language BREXIT neologisms in The Corpus of Contemporary Irish (CCI)<sup>1</sup>. This study follows a corpus approach and aims to record the development of Irish-language BREXIT neologisms in a subcorpus of media texts compiled from CCI. Firstly, a general overview is given of the Irish language along with a background to the development of various Irish-language BREXIT neologisms. The corpus is utilised to examine these terms and to identify any lexicogrammatical patterns of use. The emerging linguistic patterns are surveyed and discussed in the results section. This is a narrow study and is limited to print media, as there is currently no comprehensive corpus of spoken data available for Irish. While BREXIT is still a relatively new term, it is highly topical and in widespread use at the time of writing. This analysis aims to provide an insight into the development of unique Irish-language neologisms along with providing a base for future comparisons of similar neologisms in other minority languages

**Keywords:** Irish language, corpus linguistics, lexicography, neologisms

## 1 Introduction and background

The term BREXIT was added to the Oxford English Dictionary in December 2016 and has become a global word since its initial coinage in 2012. BREXIT is a blend of Britain or British with exit, which captures the meaning “Britain exiting from the EU” or “Britain’s exit from the EU” or “British exit from the EU” (Fontaine 2017: 2). Its rapid rise in use reflects the importance of the phenomenon it describes, along with the need to fill a lexical gap in language. It is currently in widespread use and is universally accepted and understood in various languages around the world. There are two primary Irish-language equivalent terms in use – *Brexit* and *Breatimeacht*. While the term *Brexit* is commonly used in both spoken and written Irish, the term *Breatimeacht* has gained increased currency, and both terms are now frequently used. *Breatimeacht* is a blend of *Breatain* (Britain) with *imeacht* (to leave), which captures a similar meaning to its English-language equivalent.

The Irish language is an official language in Ireland and the European Union. Although it is recognized as the first official language in Ireland, it holds minority status. New Irish-language terms continually enter the lexicon in various domains, but it is difficult to predict which terms will be adopted and which will fall into decline.

Table 1: A sample of newly-coined Irish-language terms published in the National Terminology Database for Irish in 2018.

English term	Irish term	Domain
hackspace	<i>ceárta chomhspéise</i>	Computers, Computer Science
horizon scanning	<i>faire na fáistine</i>	Government › Administration
follow churning	<i>suaithheadh leanúna</i>	Computers, Computer Science › Information Technology › Internet › Social Media

<sup>1</sup> <https://www.gaois.ie/g3m/ga/>

While Crystal (1995: 132) claims that there is never any way of telling which neologisms will stay and which will go, Metcalf (2002: 152) suggests that the “success or failure of new words is not entirely random”. The continual need for new words reflects the importance of language as a communication device (Kerremans 2013: 18) and the term BREXIT, in both English and Irish, is exceptional in the sense that it is a highly topical term and is much more frequently used than other newly-coined words.

BREXIT has not simply come into use in an ad hoc and temporary way (i.e. not fully adopted into the language). It has on the contrary, gained media currency and is, at the time of writing, a term that all UK residents know. (Fontaine 2017: 13)

Kerremans (2013: 18) suggests that neologisms are often coined due to a combination of a social need which is frequently intertwined with a semantic need in the language, i.e. to fill a lexical gap. The creation and adoption of an individual Irish-language term for BREXIT to fill this empty space in Irish is unique inasmuch as the English-language term BREXIT has gained currency in German-, Spanish- and French-language print media (Fontaine 2017: 5), and equivalent translated terms have not been commonly used or created in other languages.

## 2 Formation and development

The first recorded use of the term BREXIT in an Irish-language context stems from social media in May 2015 (confirmed by analysis undertaken by Kevin Scannell on Irish-language terms for BREXIT in online texts) (K Scannell 2018, personal communication, 22 March).

- (1) Cúis faoiseamh dom nach iad an DUP atá ina “déantóir Rí”. Cúis imní féidearthacht Brexit. Cúis amhrais easaontú na Ríochta. [8 May, 2015, Twitter]<sup>2</sup>
- (1) It is a cause for relief for me that the DUP are not “king-makers”. The possibility of Brexit is a cause for anxiety. Dissolution of the Empire is a cause for doubt.<sup>3</sup>

Similarly, the first recorded use of the Irish-language term *Breatimeacht* can be found on social media in November 2015. The initial Irish-language references focus primarily on the selection of an appropriate Irish-language term for BREXIT, with the term *Breatimeacht* prevailing as suitable term in subsequent discussions.

- (2) Cé acu seo is fearr mar Ghaeilge ar Brexit? Breatamach, Breatimeacht, Breatscor nó Breatéalú? [5 November 2015, Facebook]<sup>4</sup>
- (2) Which of these is best as an Irish term for Brexit? *Breatamach*, *Breatimeacht*, *Breatscor* nó *Breatéalú*?
- (3) ... nárbh fhearr Breatimeacht? (ainneoin pbhreith RTE/BBC, Breatimeacht thar am, dar le go leor de mhuintir na 6 chontae 😊) [5 November, 2015, Twitter]<sup>5</sup>
- (3) ... would *Breatimeacht* not be better? (despite RTE/BBC poll, about time for *Breatimeacht*, according to a lot of people in the 6 counties 😊)

Following initial references online, both *Breatimeacht* and *Brexit* were added to *The National Database for Terminology* ([téarma.ie](http://www.tearma.ie)<sup>6</sup>) and to *The New English-Irish Dictionary* ([foclóir.ie](http://www.foclóir.ie)<sup>7</sup>), two of the

<sup>2</sup> <https://twitter.com/aonghusoha/status/596693772156706816>

<sup>3</sup> All translations are the author's own unless otherwise specified.

<sup>4</sup> <https://www.facebook.com/groups/166677873392308/permalink/973884952671592/>

<sup>5</sup> <https://twitter.com/Cormacag5/status/662276243002404864>

<sup>6</sup> <http://www.tearma.ie/Home.aspx>

<sup>7</sup> <https://www.foclóir.ie/en/>

primary online Irish-language terminological and lexicographical resources currently available, in September 2016 and January 2017, respectively. *An Coiste Téarmaíochta* (The Terminology Committee) is responsible for the creation and development of authoritative standard Irish-language terms which are published in the *tearma.ie* database. Precedent is given to the term *Brexit* in the database, as the term *Breataimeacht* is marked with the explanatory note “in use”. The term *Breataimeacht* is recorded in the database as an additional term in recognition of its frequent use in Irish (J Ní Mhaolain, personal communication, 22 March). However, the semantic meaning of the term *Breataimeacht* created difficulties for *An Coiste Téarmaíochta* prior to its inclusion in the *tearma.ie* database. As the term stems from the noun Britain (*Breatain*), it does not include Northern Ireland which is part of the United Kingdom but not part of Britain. Similar issues arose in relation to other proposed Irish-language terms which were in circulation during the decision-making process, e.g. *Sasamach* (a blend of *Sasana* (England) and *amach* (to go out)), *Bréalú* (a blend of *Breatain* (Britain) and *éalú* (to escape) along with other similar variations (ibid.)

Ní Ghallochobhair (2014: 226) sets out the guidelines followed by *An Coiste Téarmaíochta* in relation to Irish-language term creation, and both the terms *Brexit* and *Breataimeacht* fall within these.

- *Roghnú ó théarmaí dúchasacha a bailíodh* (to select from previously collected native terms);
- *Síneadh brí a chur le focail/míreanna dúchasacha* (to add an extended meaning to native words/parts);
- *Diorthú ó fhocal dúchasach* (e.g. *ainmfhocal* as *aidiacht*) (to derive from a native word (e.g. a noun from an adjective));
- *Iasachtaí a thraslitriú* (transliteration of loanwords);
- *Iasachtaí a fhágáil sa bhunteanga* (to leave loanwords in the source language);
- *Meafar Gaeilge in ionad meafair iasachta* (to choose an Irish-language metaphor over a loan metaphor);
- *Tearma tuairisciúil in ionad meafair iasachta* (to choose a descriptive term over a loan metaphor);
- *Lomaistriúchán* (calque, loan translation).

Additionally, both *Brexit* and *Breataimeacht* are listed as Irish-language words in *The New English-Irish Dictionary* (NEID), developed by *Foras na Gaeilge*. NEID is the most comprehensive English-Irish online dictionary currently available, and precedence is given to the word *Breataimeacht* over *Brexit*. *Breataimeacht* is listed as the initial entry and *Brexit* is marked with the explanatory note “foreign”. Furthermore, only *Breataimeacht* is used in the Irish-language examples for “hard Brexit” (*Breataimeacht cruá*) and “soft Brexit” (*Breataimeacht bog*).

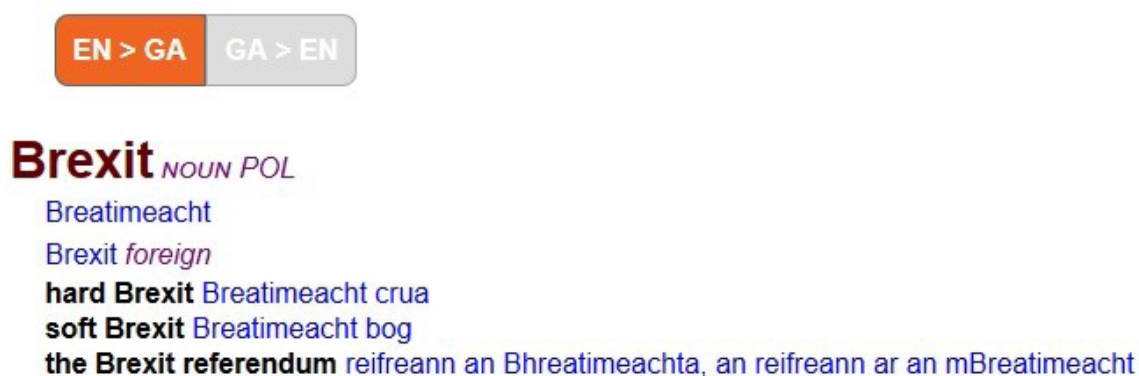


Figure 1: *New English-Irish Dictionary*, s.v. ‘Brexit’<sup>8</sup>

8 <https://www.focloir.ie/en/dictionary/ei/Brexit>

Kerremans (2013: 20) suggests that a positive influence in the adoption of neologisms “is to be expected when the neologism is used in more formal types of sources like newspapers and large Internet portals, because they guarantee a large readership that in turn can diffuse the neologism further in ever expanding circles.” It could be suggested that the inclusion of the term *Breatimeacht* in both foclóir.ie and téarma.ie has assisted in its adoption over other Irish-language variants. Figure 2 maps various Irish-language BREXIT neologisms as used online. A significant increase in use is noted in June 2016, which reflects the term’s topicality during the referendum regarding the United Kingdom’s membership of the European Union. However, the increased usage of the term *Breatimeacht* over other Irish-language neologisms occurs from September 2016 onwards, which coincides with its inclusion in the téarma.ie database and subsequent addition to NEID.

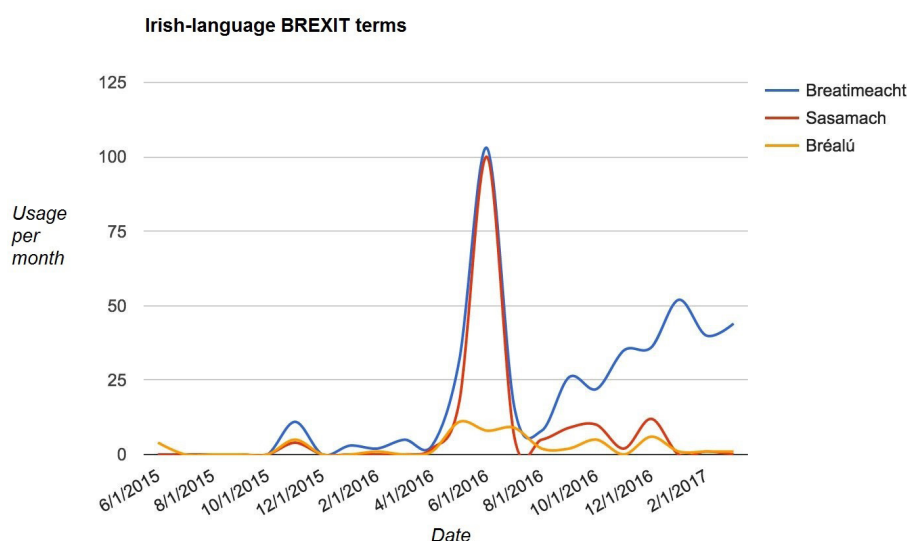


Figure 2: Use of Irish-language BREXIT neologisms online<sup>9</sup>.

Ní Laoire (2008: 191 as cited in O’Connell 2013: 199–200) suggests that even though there is limited evidence “of a direct casual, measurable link between media and language styles in a speech community”, it is accepted that all aspects of the media “form part of the linguistic mix that is a speech community and can be accepted as reflecting current language use to some degree”. Moreover, she argues that “the media may influence the speed and spread of changes in language use through their role in picking up and reflecting changes in progress” (ibid.) Although the sociolinguistic factors influencing the adoption of the term *Breatimeacht* over other Irish-language equivalent terms is not the central focus of this paper, it is submitted that the selection and conscious use of the term *Breatimeacht* on the primary Irish-language current affairs radio show *Cormac ag a Cúig* (F Ó Drisceoil, personal communication, 11 May) has helped its dissemination and adoption in Irish-speaking communities. However, further analysis is needed to comprehensively assess why the term *Breatimeacht* has gained precedence over other Irish-language equivalent terms.

### 3 Methodology

The Corpus of Contemporary Irish, the principal Irish-language corpus of published texts in the 21st century, is utilised in this study to examine the incremental rise in use of the term *Breatimeacht* over other Irish-language neologisms coined to describe BREXIT. A subcorpus of print media texts was

<sup>9</sup> <https://twitter.com/kscanne/status/844958295047766017> (With permission from Kevin Scannell)



created to facilitate this study, ranging from the term's first recorded use in November 2015 to December 2017. The corpus was compiled using the AntConc corpus analysis toolkit (Anthony 2018), and contains 8.2 million words. This is an unannotated corpus, and its contents are set out in the appendix of this paper. This study follows a corpus-based approach, as previously identified Irish-language BREXIT neologisms are analysed in the corpus search. However, it is also corpus-driven as it does not limit the search to these terms alone, but seeks to uncover other Irish-language terms and phrases used to describe BREXIT. Since the corpus is unannotated, each of the variant morphological forms for both terms *Brexit* and *Breataimeacht* were queried to ensure a complete reflection of frequency and usage was reported.

Initially, a word list was generated, and any potential Irish-language equivalent terms for BREXIT were selected and analysed. These terms were chosen by examining entries on the word list beginning with the letter 'b' (*Breatain* / Britain), 's' (*Sasana* / England) and 'r' (*Ríocht Aontaithe* / United Kingdom). The resultant terms along with their frequency in the corpus are listed in Figure 3.

Figure 3: Frequency of BREXIT terms in CCI

Term	Frequency
<i>Breataimeacht</i>	925
<i>Brexit</i>	314
<i>Sasamach</i>	17
<i>Bréalú</i>	14
<i>Breataimeacht</i>	2
<i>Breatamach</i>	1
<i>Breatéalú</i>	1
<i>Brimeacht</i>	1

The following section discusses the primary usage patterns for the terms *Brexit*, *Breataimeacht*, *Sasamach* and *Bréalú*. Results with a frequency lower than five are not included in this analysis. As this is an unannotated corpus, a collocational analysis of the various terms could not be undertaken due to numerous morphological variations which were presented in the results. However, a future analysis of this would provide an interesting insight into the various collocational and frequency patterns of these terms.

## 4 Results

### 4.1 Markers

Fontaine (2017: 4) states that as new words are coined they are generally marked in some way as they enter the language.

When a word first appears in a language, whether as a loan or calque, or as a nonce formation, it appears that speakers are aware of its newness, that is they are aware that they are exploiting the productivity of the language system. Thus, in modern journalistic language the word is often put in inverted commas, a phrase is added such as “what has been called”, “as it is termed” and so on, or a complete gloss is provide. (Bauer (1983: 42) as cited in Fontaine (2017: 4))

There is strong evidence in the corpus of each of these factors in the Irish-language forms presented in the results.

*Bréalú*

- (4) Leis an reifreann maidir le “Bréalú” nó “Brexit” ag tarraingt orainn i mí Meithimh, tá an díospóireacht faoi lán seoil, ach cad é an tionchar a bheas aige orainne sna Sé Chontae? (2016-03-22, *Meon Eile*)
- (4) With the “Bréalú” or “Brexit” referendum approaching us in June, the debate is in full swing, but what impact will it have on us in the Six Counties?
- (5) Léiríonn taighde an ESRI gur gnónna i dTuaisceart Éireann agus i gcontaethe ar an dteorainn is measa a bhuaifí dá dtarlódh ‘Bréalú’, sé sin dá bhfágadh an Bhreatain an t-Aontas Eorpach. (2015-11-05, *Nuacht RTÉ*)
- (5) ESRI research reflects that businesses in Northern Ireland and in border counties would be worst affected if ‘Bréalú’ takes place, that is if Britain leaves the European Union.

*Sasamach*

- (6) Ach go luath tar éis an vóta ar son ‘Sasamach’ thug sé tuairim uaidh don *Belfast Telegraph* ag rá go mbeadh daoine ó thuaidh anois ag meá cén todhchaí ab fhearr leo – fánacht sa Ríocht Aontaithe ach lasmuigh den Aontas Eorpach nó bheith mar bhaill d’Éirinn Aontaithe a bheadh mar chuid den Aontas Eorpach. (2016-07-18, *Meon Eile*)
- (6) But shortly after the vote for ‘Sasamach’ he gave his opinion to the *Belfast Telegraph*, stating that people in the north were now weighing up which future they would prefer – to remain in the United Kingdom but outside the European Union or as members of a United Ireland which would be part of the European Union.
- (7) Ach, tá tírdhreach na hAlban athraithe ó bhonn anois agus is é Páirtí Náisiúnta na hAlban atá anois i mbun cúrsaí agus iad ag ullmhú leo le haghaidh athreifrinn i ndiaidh an vóta ar son Breatimeachta (nó ‘Sasamach’ mar is gnáth le hEoghan Ó Néill a thabhairt air ar Raidió Fáilte). (June 2016, *An tUltach*)
- (7) But the landscape in Scotland is now utterly changed and it is the Scottish Nationalist Party that is now in charge and they are preparing for a re-referendum after the vote for *Breatimeachta* (or ‘Sasamach’ as Eoghan Ó Néill is inclined to refer to it on Raidió Fáilte).

*Breatimeacht*

- (8) Agus é ag trácht ar cheist na ‘Breatimeachta’ (Brexit) an tseachtain seo caite dúirt ár dTaoiseach Ionúin linn go gcrochfaí claí na teorann arist ar an bpointe boise dá mba rud é go bhfágadh an Ríocht Aontaithe an tAontas Eorpach. (2016-01-26, *Tuairisc.ie*)
- (8) While referring to the ‘Breatimeachta’ (Brexit) question last week, our Beloved Taoiseach informed us that the border fence would be re-established immediately if the United Kingdom left the European Union.
- (9) Má tá an Bhreatain idir dhá cheann na meá maidir le fanacht nó imeacht, ní bheadh a lán daoine in Éirinn a bheadh in amhras ná gur chóir dúinne cloí go daingean lenár mballraíocht, pé bóthar a thógadh an Bhreatain nó pé fadhbanna a chruthódh ‘Breatimeacht’ dúinn. (April 2016, *Feasta*)
- (9) If Britain is hanging in the balance regarding staying or leaving, not many people in Ireland will be in doubt that we should adherently stick to our membership, despite the road Britain will take or whatever problems ‘Breatimeacht’ will create for us.

*Brexit*

- (10) Ag caitheamh leide a bhí sé ag Méara Londain Boris Johnson, atá ar son “Brexit”, is é sin an Bhreatain ag fágáil an AE. (2016-02-22, *Nuacht RTÉ*)
- (10) He was giving a hint to the Mayor of London, Boris Johnson, who is for “Brexit”, that is Britain leaving the EU.

- (11) Cé go mbíodh Sinn Féin ar an dul Eoraisceipteach céanna leis na haontachtaithe tá siad iom-paithe i bhfabhar agus ag tabhairt foláirimh faoin gcontúirt d'Éirinn, thuaidh agus theas, dá nglacfaí le 'Breatamach' – Brexit mar a thugtar air. (2016-02-23, *Tuairisc.ie*)
- (11) Even though Sinn Féin have been on the same Eurosceptic side as the unionists, they have turned in favour and are warning about the dangers that exist for Ireland, both north and south, if 'Breatamach' – Brexit as it is termed, is accepted.

As seen in the above examples, the newness of each term is primarily marked in results relating to late 2015 and 2016. The use of markers is not as frequent in the corpus results from 2017 onwards, which suggests the terms have been conventionalized to some degree. Additionally, an explanatory gloss follows the newly-coined terms in examples (5) and (11). Both *Sasamach* and *Bréalú* mainly occur in results from 2016, and both terms are chiefly found in publications based in Northern Ireland, e.g. *An tUltach* and *Meon Eile*. The contextual use of these terms reflects the potential political and economic impact the United Kingdom leaving the European Union could have on Northern Ireland. Furthermore, example (7) suggests the term *Sasamach* is predominantly used by Eoghan Ó Néill, a radio presenter and journalist based in Belfast, Northern Ireland. Both *Bréalú* and *Sasamach* have not been widely adopted, as evidenced by the limited number of examples produced in the corpus results. The following section focuses on the development of the term *Breatimeacht* along with its primary patterns of use.

## 4.2 Gender

While the term *Breatimeacht* is classified as a masculine noun, it is used both as a masculine and feminine noun in the corpus results. In Irish, a masculine noun following a definite article is not typically lenited, e.g. *an fear* (the man). However, a feminine noun following a definite article is typically lenited, e.g. *an bhean* (the woman). Examples (12) and (13) show *Breatimeacht* used as both a masculine and feminine noun.

- (12) Agus an Breatimeacht ag druidim linn, tá inní ar an gCoiste go ndéanfar maolú ar an aitheantas a tugadh don Ghaeilge i dTuaisceart na hÉireann le blianta beaga anuas. (August 2017, *Feasta*)
- (12) With the *Breatimeacht* getting ever closer, the Committee is worried that the recognition given to the Irish language in Northern Ireland in recent years will be reduced.
- (13) Is í an Bhreatimeacht an t-athrú is mó ar shaol na ndaoine le fada. (2016-07-27, *Tuairisc.ie*)
- (13) The *Breatimeacht* is the biggest change affecting people's lives in a long time.

Another example of variation in gender can be seen in the use of *Breatimeacht* in conjunction with an adjective. In Irish, an adjective following a masculine noun is not typically lenited, but an adjective following a feminine noun is. Examples (14) and (15) show the Irish-language term for "hard Brexit", with example (14) reflecting its use as a masculine noun, e.g. *Breatimeacht crua* and example (15) reflecting its use as a feminine noun, e.g. *Breatimeacht chrua*.

- (14) Níl aon mhíniú tairgthe aici ach oiread ar a ráiteas "no deal is better than a bad deal", ná ar an bhfáth go mb'fhearr Breatimeacht crua ná socrú éigin a chuideodh le comhlachtaí na Breataine earraí a dhíol san AE. (2017-06-01, *Tuairisc.ie*)
- (14) She has offered no explanation either on her statement "no deal is better than a bad deal", nor a reason as to why a hard *Breatimeacht* would be preferable nor some type of an arrangement which would help British businesses sell products in the EU.
- (15) Ba chosúil nár mhian léi eolas ar bith a scaoileadh cé is moite de leide a thabhairt gur Breatimeacht chrua a bheadh ann. (November 2016, *Comhar*)
- (15) It appears that she would not like to give anything away except to hint that a hard *Breatimeacht* will happen.

There are a few possible reasons for the variance in gender exemplified in the corpus results. As previously stated, *Breatimeacht* is a blend of the terms *Breatain* and *imeacht*. The term *imeacht* is classified both as a verbal noun and masculine noun in Irish. However, many nouns ending in -(e)acht are classified as feminine in Irish, and this could lead to confusion in relation to the gender of a newly-coined term such as *Breatimeacht*. The primary Irish-English dictionary, *Foclóir Gaeilge-Béarla* (Ó Dónaill 1977), classifies *imeacht* as a masculine noun. However, an additional variant form is included in the dictionary, which suggests that *imeacht* is also used as a feminine noun (ibid. s.v. 'imeacht') which reflects specific dialectal usage. Furthermore, the term *Breatain* (Britain) is a feminine noun which might lead to the presumption that *Breatimeacht* should be classified as a feminine noun. A future quantitative comparative analysis of both *Breatimeacht* and *imeacht* in relation to gender variance could provide insight into whether *Breatimeacht* as a feminine noun is more common than *imeacht* as a feminine noun. This analysis would also assist in explaining whether the variance in gender of *Breatimeacht* is influenced by the term *Breatain*, classified as a feminine noun, or whether the variance of gender in the corpus examples is a result of the term's relevant newness and recent coinage.

### 4.3 Definite article

In addition to variance in gender, the corpus results reflect the intermittent use of a definite article in conjunction with the term for BREXIT, e.g. *an Breatimeacht*. The inclusion and exclusion of a definite article is shown in a sample of results below.

- (16) Luadh an Breatimeacht agus na deacrachtaí leis an Fheidhmeannas i rún na hArd-Chraoibhe i dtaca le hAcht Gaeilge de. (March 2017, *An tUltach*)
- (16) The *Breatimeacht* and the issues with the Executive were mentioned in a resolution by the *Ard-Chraobh* in relation to an Irish Language Act.
- (17) Cás ar leith é Tuaisceart Éireann agus dá bhrí sin caithfear é a chosaint ó Bhreatimeacht. (2016-08-14, *Tuairisc.ie*)
- (17) Northern Ireland is an exceptional case and therefore needs to be protected from *Bhreatimeacht*.
- (18) B'fhéidir gur dóigh leat gur fada é seo ón mBreatimeacht. (2016-03-14, *The Irish Times*)
- (18) Maybe you think that this is a long way from the *mBreatimeacht*.
- (19) Sholáthair Nicola Sturgeon plécháipéis polasaí faoi stádas na hAlban san Aontas Eorpach agus cé gur vótáil an Bhreatain Bheag i bhfabhar Breatimeachta, chuir an Chéad-Aire Carwyn Jones polasaí chun tosaigh i bhfabhar ceangal leis an margadh Eorpach. (2017-01-31, *Tuairisc.ie*)
- (19) Nicola Sturgeon provided a discussion policy document in relation to Scotland's status in the European Union and even though Wales voted for *Breatimeachta*, First Minister Carwyn Jones put forward a policy in favour of linking with the European market.
- (20) Reifreann sa Ríocht Aontaithe inar vótáladh i bhfabhar na Breatimeachta, nach léir a toradh fós. (December 2016, *Comhar*)
- (20) A referendum in the United Kingdom in which the *Breatimeachta* was voted for, its results are still not clear.
- (21) Ar *Monocycle*, ní luaitear toghchán Mheiriceá, ní luaitear Breatimeacht, ní luaitear ráta malairte. (2016-07-08, *NÓS*)
- (21) On *Monocycle*, the American election is not mentioned, *Breatimeacht* is not mentioned, exchange rates are not mentioned.

Examples (16), (18) and (20) show the use of the term *Breatimeacht* in conjunction with a definite article. However, examples (17), (19) and (21) show the use of the term *Breatimeacht* without a definite article. *An Caighdeán Oifigiúil* (2016: 4), the official standard for Irish, states a definite article should be used when a noun has an abstract or conceptual meaning associated with it. This suggests that a

definite article should be used in conjunction with the term *Breatimeacht*, but the implementation of this rule varies greatly in the corpus results, as evidenced in example (16) to (21).

#### 4.4 Periphrastic sentences

Zabaleta et al. (2008: 207) suggests that one of the primary difficulties facing journalists working in minority languages is “the translation and/or formation of technical words that, as a rule, refer to specialized topics and come from wire services or sources in the majority language.” The following points are listed as strategies which are utilised by communities, media organizations and journalists to combat these difficulties:

- (1) looking up in dictionaries; (2) writing periphrastic sentences that would circumvent the need for a specific term; (3) newsroom discussions and agreement on new terms; (4) language training courses organized by the organizations; (5) elaboration of internal stylebooks; and (6) hiring of linguists to correct and standardize the writing of journalists, a capacity only in the hands of large media outlets (ibid.)

The corpus results show some evidence of Zabaleta et al.’s second point, i.e. writing periphrastic sentences that would circumvent the need for a specific term. The primary phrase employed in reference to BREXIT in the corpus is *imeacht na Breataine as an Aontas Eorpach* or its variant form *imeacht na Breataine ón Aontas Eorpach* which translates as “Britain leaving the European Union”.

- (22) Chaith Seansailéir an Stáitchiste, George Osborne, dhá lá abhus i mbun feachtais ag tathant ar dhaoine vótáil in aghaidh imeacht na Breataine as an Aontas Eorpach sa reifreann ar an 23 Meitheamh. (2016-06-07, *Tuairisc.ie*)
- (22) The Chancellor of the Exchequer, George Osborne, spent two days there campaigning and urging people to vote against Britain leaving the European Union in the referendum on the 23 June.
- (23) ...agus go gcinnteoidh an Rialtas ó dheas roimh imeacht na Breataine as an Aontas Eorpach nach ndéanfar aon laghdú ar stádas oifigiúil na Gaeilge mar theanga oibre san Aontas Eorpach, agus go leanfar ar aghaidh leis an aitheantas don Ghaeilge mar phríomhtheanga na hÉireann sa bhaile agus thar lear. (August 2017, *Feasta*)
- (23) ... and that the government in the south would confirm before Britain leaves the European Union that the status of the Irish language as an official language of the European Union will not be reduced, and the recognition of the Irish language as the primary language of Ireland at home and abroad will continue.
- (24) Bhuail an Taoiseach Enda Kenny le príomhaire na Breataine David Cameron an tseachtain seo caite agus bhí Brexit – imeacht na Breataine ón Aontas Eorpach – ar cheann de na hábhair a phléigh siad. (2016-02-03, *The Irish Times*)
- (24) The Taoiseach Enda Kenny met with the British Prime Minister, David Cameron last week and Brexit – Britain leaving the European Union – was one of the topics under discussion.

However, it is accepted that the phrase “Britain leaving the European Union” and its variant forms is also used in other languages, and a comprehensive comparison of use would need to be undertaken to confirm whether the Irish-language phrase is more prevalent than its equivalent form in majority languages. Ní Ghallchobhair (2014:3) suggests that journalists in spoken media focus on concise and clear language and are inclined to simplify, or completely avoid, technical terms. It is suggested that print journalists are not as restricted regarding space and have an opportunity to review a new term or phrase (ibid.), which could explain the use of such periphrastic sentences in this corpus of print media texts. However, such phrases are also in use in other non-minority languages, and as there is no spoken corpus of Irish language media currently available it is difficult to confirm whether periphrastic sentences are more commonly used to describe BREXIT over the equivalent terms *Brexit* or *Breatimeacht* in Irish.



## 5 Conclusion

This paper aimed to map the adoption and development of Irish-language BREXIT neologisms through undertaking a corpus analysis of use in print media. The corpus results show that the newly-coined term, *Breatimeacht*, has gained precedence in Irish-language print media over other Irish-language equivalent terms. These results contrast to analysis undertaken on the use of both *Brexit* and *Breatimeacht* in Irish-language content online (K Scannell 2018, personal communication, 22 March). At the time of writing, the term *Brexit* (1998 references) has gained precedence over *Breatimeacht* (1813 references) in online references. While the research presented in this paper provides an insight into initial use of Irish-language BREXIT neologisms in print media, further work is needed to provide a more comprehensive overview of the terms' adoption and patterns of use, especially in spoken media. It is yet to be seen whether *Breatimeacht* will become a fully conventionalized term in Irish or whether it is just a transitory vogue word such as “millennium bug or Y2K, which were in vogue towards the end of 1999... and do not (or hardly) occur in current language” (Kerremans 2013: 38). The linguistic analysis and contextual examples suggest that *Breatimeacht* has gained widespread dissemination in print media and this has assisted in its use and adoption in other domains. It has gained the upper hand over other Irish-language BREXIT neologisms in print media, but the contextual examples reflect that there is still some uncertainty in relation to its uniformity of use, e.g. gender. However, its topicality can only further increase its awareness and dissemination, and thus it has strong potential of becoming fully conventionalized in the Irish-language.

## References

- Anthony, L. (2018). AntConc (Version 3.5.2) [Computer software]. Tokyo, Japan: Waseda University. Accessed at: <http://www.laurenceanthony.net/software> [15/02/2018].
- An Caighdeán Oifigiúil (2016). *Gramadach na Gaeilge: An Caighdeán Oifigiúil*. Tithe an Oireachtais: Baile Átha Cliath.
- Corpus of Contemporary Irish*. Accessed at: <https://www.gaois.ie/g3m/en/> [13/02/2018].
- Crystal, D. (1995). *The Cambridge Encyclopaedia of the English Language*. Cambridge: Cambridge University Press.
- Fontaine, L. (2017). The early semantics of the neologism BREXIT: a lexicogrammatical approach. In *Functional Linguist*, 4:6. Accessed at: <https://doi.org/10.1186/s40554-017-0040-x> [12/12 2017].
- Kerremans, D. (2013). *Web of Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms*. Frankfurt am Main: Peter Lang.
- Metcalf, A. (2002). *Predicting New Words*. Boston: Houghton Mifflin.
- Ní Ghallchobhair, F. (2014). *Ar dTearmaí Féin*. Cois Life: Baile Átha Cliath.
- O'Connell, E. (2013). Towards a Template for a Linguistic Policy for Minority Language Broadcasters. In E. H. G. Jones, E. Uribe-Jongbloed (eds.) *Social media and minority languages: Convergence and the creative industries*, pp. 187-201. Bristol, United Kingdom: Multilingual Matters.
- Ó Dónaill, N. (1977). *Foclóir Gaeilge-Béarla*. An Gúm: Baile Átha Cliath.
- The National Database for Terminology*. Accessed at: <http://www.tearma.ie/Home.aspx> [09/02/2018].
- The New English-Irish Dictionary*. Accessed at: <https://www.focloir.ie/> [10/02/2018].
- Zabaleta, I., Xamardo, N., Gutierrez, A., Urrutia, S. and Fernandez, I. (2008). Language Development, Knowledge and use among Journalists of European Minority Language Media. In *Journalism Studies*, 9:2, pp. 195-211, Accessed at: DOI: <https://doi.org/10.1080/14616700701848238> [28/02/2018].

## Acknowledgements

This research was undertaken with support from Dublin City University and Fiontar & Scoil na Gaeilge, DCU. Valuable feedback and assistance was received from *An Coiste Téarmaíochta*, *The New English-Irish Dictionary* research team, Kevin Scannell, Brian Ó Raghallaigh and Gearóid Ó Cleircín. I would also like to thank the useful comments from the two anonymous reviewers.

## Appendix

Corpus Source	Date	Available
<i>An tUltach</i>	2015-2017	<a href="https://antultach.wordpress.com/">https://antultach.wordpress.com/</a>
<i>Comhar</i>	2015-2017	<a href="https://comhar.ie/">https://comhar.ie/</a>
<i>Feasta</i>	2015-2017	<a href="http://www.feasta.ie/">http://www.feasta.ie/</a>
<i>NÓS</i>	2015-2017	<a href="https://nos.ie/">https://nos.ie/</a>
<i>Nuacht RTÉ</i>	2015-2017	<a href="https://www.rte.ie/news/nuacht/">https://www.rte.ie/news/nuacht/</a>
<i>The Irish Times</i> (Irish-language articles)	2015-2017	<a href="https://www.irishtimes.com/culture/treibh">https://www.irishtimes.com/culture/treibh</a>
<i>Tuairisc.ie</i>	2015-2017	<a href="https://tuairisc.ie/">https://tuairisc.ie/</a>



# **Lexicography of Lesser Used Languages**





# Synonymy in Modern Tatar reflected by the Tatar-Russian Socio-Political Thesaurus

**Alfia Galieva**

*Tatarstan Academy of Sciences*

*E-mail: amgalieva@gmail.com*

## Abstract

This paper discusses some aspects of lexical synonymy in the modern Tatar language that have been revealed in the course of compiling the bilingual *Russian-Tatar Socio-Political Thesaurus*. Building the thesaurus is aimed at fixing all Tatar single words and multiword items related to the socio-political sphere with their Russian equivalents. A distinguishing feature of the contemporary Tatar lexicon is a great deal of absolute synonyms which emerged due to a combination of intralinguistic and extralinguistic factors. We disclose social and linguistic causes of the emergence of synonyms, describe the main structural types of synonymous items, and present corpus data on their frequency. Corpus data prove that synonymy in socio-political terminology is rather an artificial and superficial phenomenon. Currently most Tatar socio-political terms are coined by calquing the corresponding Russian terms, and lexical preferences of translators and terminology developers may differ, which leads to a large number of competing items of different origin and structure. On the level of multiword items, lexical variation is complicated by the factor of syntactic variation, which in its turn multiplies the number of synonymous compounds. Parallel denominations are used for a wide range of phenomena, including official names of state structures and social institutions.

**Keywords:** lexical synonymy, absolute synonyms, socio-political vocabulary, bilingual thesaurus, the Tatar language

## 1 Introduction

Collecting lexical data, systemizing it and mapping semantic relations between the lexical items are the important stages of lexicographic work, ones that gives material for comprehending current processes in languages. In this paper, we discuss some facts of lexical synonymy in the modern Tatar language that have been encountered in the process of compiling the bilingual *Russian-Tatar Socio-Political Thesaurus*. The Tatar language of recent decades has experienced a stage of renewal which caused significant changes in vocabulary, including the emergence of a large number of synonymous items. In present work, the main attention is paid to absolute synonymy.

The conventional viewpoint is that true synonyms are rarely found in a language. However, the situation with Tatar socio-political vocabulary in many respects contradicts this thesis about the rareness of absolute synonymy; this phenomenon is observed due to a set of extralinguistic and intralinguistic causes, which require explanation.

The body of the paper is organized as follows. Section 2 gives a brief description of the current language situation in Tatarstan, and puts forwards an idea about the social motivation of changes in Tatar vocabulary. Section 3 presents the main goals and methodology of compiling the *Russian-Tatar Socio-Political Thesaurus*. Section 4 outlines the basic theoretical background of the study. Section 5 describes the most important aspects of lexical synonymy in the modern Tatar language that have been registered in the course of compiling the bilingual thesaurus; synonymy at the level of single

words and multiword items is discussed; corpus evidence about the distribution of synonyms is given. Section 6 lists the conclusions and outlines the prospects for future work.

## 2 Social Motivation of Changes in Tatar Vocabulary

An important feature of the contemporary Tatar language is the emergence of a large number of synonymous items, this phenomenon being caused by a set of intralinguistic and social factors, which requires an explanation.

Tatarstan is a republic of the Russian Federation which is located in the Volga region of the European part of Russia, with the capital city of Kazan. According to the 2010 census, the population of Tatarstan consists of 53% of ethnic Tatars, 40% of ethnic Russians, 7% of people of other ethnic origin. At present practically all Tatars speak Russian (but not vice versa).

By the late 1980s, the minority languages of the USSR had been largely supplanted by Russian in the urban areas of most Soviet republics; ethnic groups used Russian in the public sphere, while state and public administration and office work were also conducted in Russian. The danger of losing the national language was one of the main driving forces of the ethnic renaissance that developed in Tatarstan in the late 1980s, as well as in other ethnic regions of the USSR.

The legal basis for promoting the use of the Tatar language in the public sphere emerged with the adoption of the Language Law in 1992, which triggered the implementation of the government-sponsored Tatar language revival program and the establishment by state and municipal bodies of special committees for Tatar language use. Active translation work was thus launched in state structures and institutions, and an urgent need in Tatar socio-political terminology arose.

Currently in accordance with the Constitution of the Republic of Tatarstan and the Language Law, the two state languages of the republic are Tatar and Russian, which formally have equal legal status. In particular, the texts of laws of the Republic of Tatarstan and other legal acts are to be published in both Tatar and Russian, while state authorities and institutions use both state languages of the Republic of Tatarstan.

So the late 1980s had an active social impact on the language situation in Tatarstan, and the Language Law significantly raised the status of the Tatar language. At the same time a religious revival began in Tatarstan; Tatars recognized themselves as a part of the vast Arab-Muslim cultural area, and direct cultural connections with the Arabic countries and Turkey were initiated.

The late 1980s and 1990s thus became a period of intensive linguistic innovation in Tatarstan. A new regard for Tatar culture and the increased social role of the Tatar language generated a movement for purifying it from numerous Russian and Western international words. As a result, the available vocabulary underwent revision; numerous loanwords of Russian and European origin were offered to be substituted with their Tatar, Arabic or Persian equivalents; many Arabic and Persian words that in the Soviet period were regarded as obsolescent, returned to the active lexicon again. Consequently a large number of new words appeared to designate existing and newly created realities, and Tatar vocabulary was enriched by numerous synonymous items.

So the vast changes in Tatar vocabulary of the last decades were socially motivated. The location of Tatar culture at the intersection of Occidental and Oriental civilizations leads to active language contacts both with the Arab-Muslim and the European cultural areas. At the same time, the dominant role of Russian as the state language of the Russian Federation remains the main cause of a huge number of words and constructions calqued (translated component-by-component) from Russian.

Currently most Tatar socio-political terms are coined by calquing the corresponding Russian terms. Diverse groups of specialists with dissimilar world-views and ideological guidelines develop terminology in Tatar; in terminology design they may principally be oriented towards different cultural spaces with different languages spoken:

- Tatar and Turkic vocabulary;
- Arabic and Persian vocabulary;
- international Greek, Latin or English vocabulary;
- Russian vocabulary, etc.

The heterogeneous preferences of terminology developers as well as of other language users lead to proposing and choosing different designations for the same entity. In these circumstances there is a need for new lexicographic resources that would embrace synonymous items and cover all existing variants of terms to provide users with more effective access to information.

### 3 *Tatar-Russian Socio-Political Thesaurus: Main Goals and Methodology of Compiling the Work*

For low-resource languages like Tatar, compiling special domain vocabulary is a challenge in many respects:

- insufficient degree of terminology development that leads to the absence of relevant designations for special language concepts
- a lack of special sources for terminology compiling;
- a lack of terminology management, etc.

The project of developing the *Russian-Tatar Socio-Political Thesaurus* (<http://tattez.antat.ru/>) is aimed at compiling the whole body of modern Tatar vocabulary related to the following basic domains: state government, economy, social life, justice, warfare, culture, religion. The thesaurus also comprises some general lexicon branches representing lexical items which can be found in various domain specific texts. All items have Russian equivalents (correspondences) and are represented at concept and lexical entries levels which are arranged hierarchically.

This new lexical resource is being developed on the basis of the Russian RuThes ([http://www.labinform.ru/pub/ruthes/index\\_eng.htm](http://www.labinform.ru/pub/ruthes/index_eng.htm)) thesaurus. Both thesauri are implemented as a hierarchy of concepts viewed as units of thinking. Each concept is linked with a set of language expressions (single words and multiword expressions) that refer to it in texts (lexical entries). Each RuThes concept is a set of synonyms or near-synonyms (plesionyms). RuThes developers, Loukachevich and Dobrov, use a weaker term, *ontological synonyms*, to designate words belonging to different parts of speech (like *stabilisation*, *to stabilise*) and related to different styles and genres; idioms and even free multiword expressions which are potentially synonymous to single words are also included (Loukachevich & Dobrov 2014).

The conceptual structure of RuThes determines the direction of concept development in the Tatar part, and the basic structure of the conceptual relations of RuThes is preserved, i.e. the Tatar component is based upon the list of RuThes concepts.

The general methodology of creating the Tatar part of the thesaurus includes the following steps.

1. Search for equivalents (corresponding words) which are actually used in Tatar as translations of Russian words.

2. Adding new concepts representing topics which are important for the socio-cultural life of the Tatar society and are not presented in the original RuThes (the list of required vocabulary is compiled, the concept names and lexical entries are distinguished and arranged according to RuThes structure).
3. Revising relations between the concepts considering the place of each new concept in the hierarchy of the existing ones and, if necessary, adding the new concepts of the intermediate level (Galieva et al. 2017).

The Thesaurus is mainly being compiled by manual translation of terms from the Russian RuThes into Tatar. Tatar language specific concepts and their lexical entries are also added, so each part of the Thesaurus – the Russian and Tatar ones – represents a unique language-internal system of lexicalizations. At the same time, the languages are interconnected so that it is possible to go from the concepts and words in one language to the corresponding items in the other.

The main challenge of working on this project is concerned with acquiring lexical data and representing Tatar socio-political vocabulary as fully as possible, including a large number of synonymous items in actual use. Searching for translation equivalents and adequate correspondences that are practically used in texts becomes, in many cases, a laborious and time-consuming task. The available bilingual Russian-Tatar dictionaries of general lexicon and special dictionaries are outdated, and even new lexicons do not contain the required items or include obsolete lexical material (See (Galieva et al. 2017)).

In the process of compiling the thesaurus, data from the following available Tatar corpora is used:

1. Tatar National Corpus (<http://tugantel.tatar/?lang=en>);
2. Corpus of Written Tatar (<http://www.corpus.tatar/en>).

These corpora include texts of various genres, from official documents and scientific publications to media texts, fiction, and textbooks. They are being permanently replenished, which provides a constant inflow of fresh linguistic material. The corpora have comparable volumes, each containing more than 100 million tokens, and are supplied with a system of morphological annotation (Suleymanov et al. 2013; Nevzorova et al. 2015). The data provided by these corpora allow us to acquire reliable information on meanings, typical contextual relations and frequency of use of Tatar words, which is a necessary stage in compiling a thesaurus. Currently the *Russian-Tatar Socio-Political Thesaurus* contains 8,000 concepts provided with lexical entries.

#### **4 Theoretical and Lexicographic Aspects of Synonymy: Brief Outline**

Synonymy is one of the fundamental concepts in linguistics; it manifests itself at different levels of the language. A prevailing point of view in linguistics is that lexical synonymy shows a high degree of development of a language, enabling us to encode the most subtle differences between entities or ideas and different perspectives of viewing them.

The notion of synonymy has been subject to a large number of studies as to what it constitutes, where its borders lie and what its scale is, and there are a number of controversial interpretations. Researchers distinguish between cognitive synonyms (Cruse 1986) and near-synonyms, or plesionyms (Hirst 1995; Edmonds & Hirst 2002; Divijac 2006; 2010, Desagulier 2014).

Mapping synonyms in special lexicographic resources demands some kind of technical approach, because the developers cannot carry out a meticulous study on semantic and contextual differences of thousands of words represented in a resource. A synonym, in *Webster's New Dictionary of Synonyms*,

means “one of two or more words in the English language which have the same or very nearly the same essential meaning” (Egan 1984: 24 a).

Synonymy is the basis for organizing lexical items in Princeton WordNet (Miller et al. 1990, Miller 1995) and other WordNet-like resources (Vossen 2002). Miller and Fellbaum suggest a vaguer term for synonymy, namely “semantic similarity”, which is defined as follows: “two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value” (Miller et al. 1990).

In *EuroWordNet General Document* words are considered to be synonyms if they “denote the same range of entities, irrespective of the morpho-syntactic differences, differences in register, style or dialect or differences in pragmatic use of the words” (Vossen 2002: 18). Moreover, developers rely upon another, more practical, criterion which follows from the homogeneity principle that synonyms cannot have any other semantic relation (Vossen 2002: 18).

The notion of absolute synonymy also leaves room for different interpretations. Cruse maintains that two lexical units would be absolute synonyms if and only if all their contextual relations were identical (Cruse 1986: 168).

According to Lyons, two (or more) synonymous expressions are absolute synonyms if they satisfy the following three conditions:

- 1) all their meanings are identical;
- 2) they are synonymous in all contexts;
- 3) they are semantically equivalent (i.e. their meaning or meanings are identical) in all the dimensions of meaning, both the descriptive and the non-descriptive ones (Lyons 1995: 61).

The attitudes in related works range from total rejection of absolute synonymy in a language to claiming their extreme uncommonness (Ulmann 1962; Lyons 1995; Cruse 1986 and others). Cruse emphasizes that “natural languages abhor absolute synonyms just as nature abhors a vacuum” (Cruse 1986: 270), and maintains that it would be reasonable to assume “that if the relationship were to occur, it would be unstable”, because one of the items would fall into obsolescence, or some differences between the items would develop (Cruse 1986: 270). Synonymous items are usually distinguished from each other by a set of individual senses, added implications and connotations, or idiomatic use.

## 5 Synonymy in Current Tatar Socio-Political Vocabulary

### 5.1 Synonymy of single words

In the Soviet era many words denoting socio-political realities were borrowed from Russian and European languages or were coined by component-by-component translation of Russian items. A movement for the Tatar language revival led to an active language renewal, which enriched the Tatar vocabulary with a great number of synonymous items of Tatar and Oriental (Arabic and Persian) origin. Many Arabic and Persian words and their derivatives considered as obsolescent in the Soviet period (and correspondingly marked in dictionaries of the time) were thus entered into the active vocabulary fund. So in modern Tatar, words of different roots (Turkic, Russian, Arabic, Persian, Greek, Latin, and English) denoting the same referent, are competing. Selection of homosemous words is not completed yet, and the concurrence brought dissimilar results for different lexemes. Corpus data allows us to observe and fix the process of coining new terms, although this process is laborious and contradictory in many aspects. Table 1 represents one-word synonyms of European and Russian origin which are currently being displaced with equivalents of Oriental (Arabic and Persian) and Tatar



(Turkic) origin. It should be noted that words of Oriental origin in Table 1 abound in present-day official and media texts which underwent preliminary editorial revision. However, in actual oral and especially colloquial speech, words of Russian origin and international words and their derivatives are actively used.

Table 1: Examples of words of European and Russian origin which are presently being displaced with Oriental and Turkic equivalents.

Words	Word origin	English translation	Frequency in Corpus of Written Tatar / Tatar National Corpus	Type of texts of words using
экономик	Greek	'economic'	621 / 676	Soviet official and media texts
икътисади	Arabic	'economic'	27,351 / 22,274	Present-day official and media texts
политик	Greek	'political'	1,005 / 1,536	Soviet official and media texts
сәяси	Arabic	'political'	25,011 / 24,489	Present-day official and media texts
страхование	Russian	'insurance'	239 / 109	Soviet official and media texts
страховкалау	Tatar	'insurance'	814 / 1,331	Soviet official and media texts
иминиятләштерү	Tatar	'insurance'	2,222 / 1,221	Present-day official and media texts
иминият	Arabic	'insurance'	3,894 / 1,667	Present-day official and media texts
больница	Russian	'hospital'	2,832 / 6,383	Soviet official and media texts
хастаханә	Persian	'hospital'	19,675 / 13,397	Present-day official and media texts
сырхауханә	Persian	'hospital'	3,492 / 1,235	Present-day official and media texts
налог	Russian	'tax'	547 / 1,423	Soviet official and media texts
салым	Tatar	'tax'	23,223 / 17,322	Present-day official and media texts

Loanwords produced stems for sets of derivatives of different structure, which compete in their turn (like the Tatar words *страховкалау* and *иминиятләштерү* derived from corresponding Russian and Arabic stems, represented in Table 1).

Words of Oriental and Turkic origin do not always dominate; in many cases international and Russian words demonstrate high frequency in written texts. Table 2 presents lexical items which retained their positions despite the emergence of Oriental competitor words.

Table 2: Examples of international and Russian loanwords maintaining high frequency

Words	Word origin	English translation	Frequency in Corpus of Written Tatar / Tatar National Corpus
республика	Latin	'republic'	316,667 / 258,433
жәмһүрият	Arabic		2,479 / 1,631
суд	Russian	'court of law'	15,725 / 13,873
мәхкәмә	Arabic		3,203 / 4,957
компьютер	English	'computer'	13,469 / 10,032
санак	Turkic		631 / 67

Words of Tatar (Turkic) origin that are formed of different stems or of the same stem may also compete. Notable examples are Tatar words referring to *businessman* and *business* (entrepreneurship), from the same Turkic stem *эш* 'affair, business'. Table 3 presents the quantitative distribution of synonymous words denoting *businessman*, in corpus collections, and the number of their most frequent derivatives.

Table 3: Native Tatar words with a business-related meaning and their most frequently used derivatives

Lexeme	English translation	Frequency in Corpus of Written Tatar	Frequency in Tatar National Corpus
<i>эшмәкәр</i>	'businessman'	22,671	13,839
<i>эшкуар</i>	'businessman'	8,606	11,289
<i>эшмәкәрлек</i>	'business, entrepreneurship'	6,454	4,209
<i>эшкуарлык</i>	'business, entrepreneurship'	2,641	3,859

To assess the semantic similarity of words we may consider their distributional similarity using corpus data. It is easy to see that words *эшмәкәр* and *эшкуар* are actively used in present-day socio-political texts; they have the same meaning and are characterized by essentially identical contextual environment (see Table 4). So we may conclude that words *эшмәкәр* and *эшкуар* are absolute synonyms. The same can be said about other examples in the tables above.

Table 4: The most frequent collocations of words *эшкуарлык* and *эшмәкәрлек*

Typical collocations of word <i>эшкуарлык</i>	Frequency in Corpus of Written Tatar	Typical collocations of word <i>эшмәкәрлек</i>	Frequency in Corpus of Written Tatar
<i>эшкуарлык комитеты</i> 'business committee'	341	<i>эшмәкәрлек субъекты</i> 'business entity'	447
<i>эшкуарлык субъекты</i> 'business entity'	186	<i>эшмәкәрлек комитеты</i> 'business committee'	183
<i>эшкуарлык эшчәнлеге</i> 'business activity'	94	<i>эшмәкәрлек үсеше</i> 'development of business'	269
<i>эшкуарлык үсеше</i> 'development of business'	65	<i>эшмәкәрлек эшчәнлеге</i> 'business activity'	257

In many cases absolute synonyms form nests of elements of different frequency. In Table 5 one can find the distribution of words referring to *patriotism*.

Table 5: Synonyms of different origin and structure denoting *patriotism*.

Words	Word's origin	Frequency in Corpus of Written Tatar	Frequency in Tatar National Corpus
<i>патриотизм</i>	Greek	763	703
<i>патриотлык</i>	Tatar derivative from the Greek stem	446	105
<i>ватанпәрвәрлек</i>	Tatar derivative from the Arabic stem	853	797
<i>ватандарлык</i>	Tatar derivative from the Arabic stem	80	83
<i>ватанчылык</i>	Tatar derivative from the Arabic stem	12	30

To declare that the words above are absolute synonyms, we proceed from the assumption (which can be debated) that we do not observe a real word borrowing here (new or unknown referents with their unique links do not appear), rather one verbal label is changed by another. For example, international *economics* was arbitrarily replaced by Arabic *икътисад*, consistently and in all collocations, and ordinary Tatar speakers may not know anything about the real Arabic word denoting economics, its senses, collocations and connotations in the Arabic language. So we observe a rather superficial change of lexical items with one another.

## 5.2 Synonymy of Multiword Expressions and Compound Terms

On the level of multiword terms and phrases synonymy is complicated by grammatical factors. An interesting issue concerns the absence of primordial relative adjectives in Turkic languages. A great number of Russian terminological word combinations contain relative adjectives. Because of a comparatively small number of relative adjectives in Tatar (all of them borrowed from Oriental and European languages or from Russian), when translating such multiword items the so called *ezāfe* constructions are usually formed by two nouns. *Ezāfe* constructions in Turkic languages are used to express relative characteristics abstracted from the meaning of the attributing nominal component; this tool significantly reduces the need for adjective-noun phrases. In cases of available relative adjectives in Tatar (all of them borrowed, as we mentioned above), we see the following very frequently observed correspondence of grammatical patterns of noun phrases (abbreviations: N – noun, ADJ – adjective, POSS\_3 – possessive affix, 3d person):

*ADJ + N* and *N + N, POSS\_3*:

*икътисади кризис* (ADJ + N), *икътисад кризисы* (N + N, POSS\_3)  
‘economic crisis’.

Other regular correspondences follow the models *N, NMLZ + N, POSS\_3* and *N, PL + N, POSS\_3* (abbreviations: NMLZ - nominalization, PL - plural):

*акционерлык жәмгыяте* (N, NMLZ + N, POSS\_3), *акционерлар жәмгыяте* (N, PL + N, POSS\_3)  
‘joint-stock company’.

Such regular correspondences multiply the number of grammatical variants of multiword terms. In cases where structural (syntactic) variation of synonymous multiword items is complicated by synonymy of their components, we find ramified sets of synonymous items and their derivatives (see Table 6).

Table 6. Synonymous items designating *Constitutional court* and their frequency of use in corpus collections

Compound term	Structure of compound	Frequency in Corpus of Written Tatar	Frequency in Tatar National Corpus
Конституция суды	N + N, POSS_3	905	119
Конституцион суд	ADV + N	169	66
Конституция мәхкәмәсе	N + N, POSS_3	261	126
Конституцион мәхкәмә	ADV + N	48	19

Parallel denominations can be used in a wide range of cases, including official names of state structures and social institutions. So in thesaurus compilation a separate task is to build the most complete list of synonymous denominations related to the same concept to fix them as lexical entries for the thesaurus, because the lists of synonyms in available dictionaries are incomplete (in many cases only one translation variant of a term is represented).

## 6 Conclusion

Collecting Tatar synonymous items for the Russian-Tatar bilingual thesaurus and analyzing lexical data allows us to conclude that the abundance of absolute synonyms in Tatar is a peculiarity of the Tatar language in its current stage. In the late 1980s a more active social impact on the language situation in Tatarstan began; the Language Law significantly raised the status of the Tatar language, the social movement for purifying Tatar from redundant Russian and Western elements took place, and nowadays the language situation in Tatarstan remains unstable.

The emergence of a great number of exact synonyms in Tatar is caused by a combination of intralinguistic and extralinguistic conditions. The location of Tatar culture at the intersection of Occidental and Oriental civilizations leads to active lexical borrowing both from Arab-Muslim cultural area and European cultural area; borrowing vocabulary from European languages is carried out through mediation of the Russian language, and, certainly, a huge number of words and constructions are taken from Russian. Besides, a significant number of synonyms are built from Turkic and Tatar lexical material. Therefore, words of different origin (Turkic and Tatar, Arabic and Persian, Greek, Latin, English and Russian), coexisting in Tatar, are related to the same referent, which causes a great number of synonyms at the single word level.

Some peculiarities of lexical, derivational and grammatical systems of the Tatar language also lead to the originating of a great number of synonyms. On the level of multiword terms and phrases, lexical synonymy is complicated by the factor of using different grammatical structures, the most characteristic of these being the parallelism of *ADJ + N* and *N + N*, *POSS\_3* constructions. As a result, parallel denominations are used in official and media texts for a wide range of phenomena.

Coexistence of different centers of building Tatar terms and the lack of coordination among them leads to instability and imbalance of terminology. We can conclude that a certain degree of Tatar terminology management is thus required, and this may be descriptive (describing how terms are used in documents and media texts) and prescriptive (or even normative) for document compilers (prescribing what terms must be used in standard work and official documentation and how they must be used).

Corpus data analysis proves that synonymy both in socio-political terminology and in general lexicon are characterized by dissimilar features. Absolute synonyms are mainly related to socio-political and scientific terminology (they are characterized by unambiguousness and have a content which in many cases may be well defined) and have direct correspondences in Russian (a Russian term may have a set of Tatar equivalents). Synonymy in general lexicon has a more intricate structure, and the number of absolute synonyms is very limited. A comparative study of synonymy in socio-political domain and in general lexicon is intended as the next step of this research.

## References

- Corpus of Written Tatar*. Accessed at: <http://www.corpus.tatar/en> [28/03/2018].
- Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Desagulier, G. (2014). Visualizing Distances in a Set of Near-Synonyms. In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, 43, pp. 145-178.
- DiMarco, C., Hirst, G., & Stede, M. (1993, March). The Semantic and Stylistic Differentiation of Synonyms and Near-Synonyms. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*. Stanford, CA, pp. 114-121.
- Divjak, D. (2010). *Structuring the Lexicon: A Clustered Model for Near-Synonymy*. Walter de Gruyter.
- Divjak, D. (2006). Ways of Intending: Delineating and Structuring Near-Synonyms. In St. Th. Gries, A. Stefanowitsch (eds.). *Corpora in Cognitive Linguistics*. Berlin: Mouton, pp. 19-56.

- Edmonds, P., Hirst, G. (2002). Near-Synonymy and Lexical Choice. In *Computational Linguistics*. 28(2), pp. 105-144.
- Egan, R.F. (1984). Synonym: Analysis and Definition. In Ph. B. Gove, (ed.) *Webster's New Dictionary of Synonyms*. Springfield, Massachusetts: Merriam-Webster, pp. 23a-25a.
- Galieva, A., Nevzorova, O. & Yakubova, D. (2017). Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, the Status of the Project. In G. Angelova et al. (eds). *International Conference Recent Advances in Natural Language Processing*, 2 - 8 September, 2017. Varn: INCOMA Ltd., pp. 245-252.
- Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.
- Inkpen, D., Graeme H. (2006). Building and Using a Lexical Knowledge-Base of Near-Synonym Differences. In *Computational Linguistics*, 32(2), pp. 223-262.
- Lyons, J. (1995). *Linguistic Semantics*. Cambridge: Cambridge University Press.
- Loukachevitch, N., Dobrov, B. (2014). RuThes Linguistic Ontology vs. Russian Wordnets. In *Proceedings of the Seventh Global Wordnet Conference*. Tartu: University of Tartu Press, pp. 154-162.
- Miller, G.A., Charles, W.G. (1991). Contextual Correlates of Semantic Similarity. In *Language and Cognitive Processes*, 6(1), pp. 1-28.
- Miller, G.A. (1995). WordNet: A Lexical Database for English. In *Communications of the ACM*, 38(1)1, pp. 39-41.
- Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K.J. (1990). Introduction to WordNet: An On-line Lexical Database. In *International Journal of Lexicography*, 3(4), 235-244
- Murphy, M.L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*. Cambridge: Cambridge University Press.
- Nevzorova, O., Mukhamedshin, D. & Bilalov R. (2015). Search Engine for the 'Tugan Tel' Tatar National Corpus: Main Decisions. In *Proceedings of the International Conference "Turkic Languages Processing TurkLang-2015"*. Kazan: Academy of Sciences of Tatarstan Republic Press, p. 236-244.
- Niina N., Nina P. (2010). Synonymy in Specialised Communication – a Terminological Approach. In *Re-thinking synonymy: semantic sameness and similarity in languages and their description*, book of abstracts, pp. 66-67.
- Russian-Tatar Socio-Political Thesaurus*. Accessed at: <http://tattez.antat.hru> [28/03/2018].
- RuThes Linguistic Ontology*. Accessed at: [http://www.labinform.ru/pub/ruthes/index\\_eng.htm](http://www.labinform.ru/pub/ruthes/index_eng.htm) [28/03/2018].
- Suleymanov, D., Nevzorova, O., Gatiatullin A., Gilmullin, R. & Khakimov, B. (2013). National Corpus of the Tatar Language "Tugan Tel": Grammatical Annotation and Implementation. In *Procedia - Social and Behavioral Sciences*, 2013, 95, pp. 68-74.
- Taylor, J. R. (2002): *Cognitive Grammar*. Oxford: Oxford University Press.
- Tatar National Corpus. Accessed at: (<http://tugantel.tatar/?lang=en>) [28/03/2018].
- Ullmann, S. (1962). *Semantics, An Introduction to the Science of Meaning*. Oxford: Basil Blackwell.
- Vossen, P. J. T. M. (ed.) (2002). *EuroWordNet: General Document*. Amsterdam: University of Amsterdam.

## Acknowledgements

The reported study was funded by Russian Science Foundation, research project No. 16-18-02074.



# Revision and Extension of the OIM Database – The Italianisms in German

**Anne-Kathrin Gärtig**

Paris Lodron Universität Salzburg

E-mail: anne-kathrin.gaertig@sbg.ac.at

## Abstract

The paper presents the *Osservatorio degli italianismi nel Mondo* (OIM), an online database and a homonymous research project on Italianisms in various European target languages, and the revision of the existing data and their structure which has been underway since 2017.

The OIM is the digitized version of the *Dizionario di italianismi in francese, inglese, tedesco* (Stammerjohann et al. 2008). The paper focusses on the Italian loanwords in German registered in two opera, and outlines how further loans in this target language are being systematically integrated during the revision process, and how gaps and weaknesses in their lexicographical description are being filled.

**Keywords:** multilingual lexicography, specialized dictionaries, online lexicography, language contact, history of the Italian/German lexicon

## 1 Introduction: The OIM Database

With the turn of the millennium, we can see a turn in the lexicography of language contact. Besides classical dictionaries of foreign words, including foreign material in a recipient language, there have been various projects which, starting from the source language, have retraced the paths loan words have taken into one or more target languages, such as Görlach's *Dictionary of European Anglicisms* (2001), van der Sijs's *Nederlandse woorden wereldwijd* (2010) and the *Lehnwortportal Deutsch* (Institut für Deutsche Sprache), which offers access to lexicographical registration of Germanisms in dictionaries of Polish, Slovenian and Hebrew. Joining this group of dictionaries classified as “aktives polylaterales Sprachkontaktwörterbuch” (‘active, polylateral dictionary of language contact’) by Wiegand (2001: 125), Italian has been represented since 2008 by the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT) by Stammerjohann et al., registering Italianisms in the three European languages of French, English and German.

Concerning the number of speakers, Italian is clearly not comparable to the other source languages mentioned above. However, its status as a language of culture, which has been learned all over Europe beginning in the 14<sup>th</sup> and 15<sup>th</sup> centuries (cf. Christmann 1992), and affected domains of use such as trade and finance, art and architecture, music, gastronomy and so on like no other language, has made it an important source for borrowings in a multitude of languages all over the world. The DIFIT is the first work to register them systematically.

As the following figure, the entry for *cappuccino* shows, the DIFIT lemmatizes Italian source words, followed by information on their semantic fields or domains of use and one or more meaning descriptions. It then lists the form in which they were borrowed in the individual target languages within the microstructure. Following Gusmani (1986), a “borrowing (It. *prestito*, G. *Entlehnung*) is seen here as the result of the imitation of a foreign linguistic pattern” (Heinz & Gärtig 2014: 1100) and includes single words like G. *Pizza*, multiword expressions like G. *Frutti di mare*, formatives as the suffix

-issimo, but also loan translations as G. *Großherzog* (< It. *granduca*) and pseudo-loans (e.g. G. *pico-bello*), which are formatives that seem to be borrowed, but do not exist in the assumed source langue, or exist with another meaning (cf. Winter-Froemel 2011: 44-45).

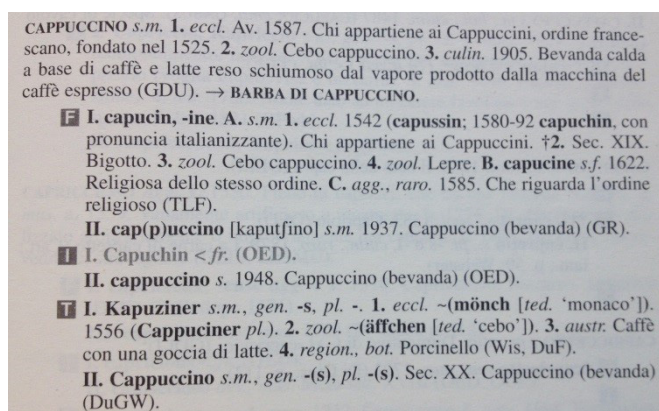


Figure 1: DIFIT, s.v. *cappuccino*.

For each borrowing, the entry contains information on the grammatical category, a definition of its meaning(s) in the target language, the domain where it is used/the semantic field it belongs to, the first attestation and information on the type of borrowing (direct borrowing, calque, pseudo-loan, etc.).

The dictionary is based on lexicographical sources, i.e. mainly on the large reference dictionaries of the target languages, from which the borrowings from Italian were extracted: the OED for English, the *Trésor de la langue française* (TLF), the *Grand Robert* (GR) and the *Dictionnaire étymologique de la langue française* by Bloch and Wartburg (BW) for French. For the Italianisms in German, the authors had to face the problem that:

il famoso *Deutsches Wörterbuch* dei fratelli Grimm contiene solo forestierismi molto antichi, da lungo integrati, mentre i *Fremdwörterbücher* [...] o sono selettivi, come il *Deutsches Fremdwörterbuch* (DFwb), o non danno datazioni, come *Das große Fremdwörterbuch* (DuF), e sono selettivi anche i dizionari etimologici, come *l'Etymologisches Wörterbuch der deutschen Sprache* (Kluge 1995) e il *Herkunftswörterbuch der deutschen Sprache* (DuE), o storici, come il *Deutsches Wörterbuch* di Paul (Paul 2002) [...]. (Stammerjohann & Seymer 2007: 42)<sup>1</sup>

This divergent lexicographical tradition leads to a lack of datings and attestations for a multitude of Italianisms in German as their target language, especially for newer borrowings arriving from the 20<sup>th</sup> century on.

Since 2014, the DIFIT data has been available online at [www.italianismi.org](http://www.italianismi.org) (cf. Figure 2 for the entry s.v. *cappuccino*, cited in the printed version in Figure 1). The international project working on its transfer and on its extension is called *Osservatorio degli italianismi nel mondo* (OIM) and is hosted by the Accademia della Crusca. Its platform offers a web interface with various filters (cf. Figure 3), allowing for varied search options, e.g. for limiting the search to features such as Italianisms of certain grammatical categories, in only one or two recipient languages, of a certain domain of use or a limited period of borrowing, and so on. This creates a useful instrument that can be used for researching the phenomenon of borrowing and language contact (cf. the exemplary studies in Stammerjohann

<sup>1</sup> 'The Grimms' famous *Deutsches Wörterbuch* contains only very old borrowings, which have long been integrated, while the *Fremdwörterbücher* [...] or are selective, as is the *Deutsches Fremdwörterbuch* (DFwb), or do not give datings, like *Das große Fremdwörterbuch* (DuF). The etymological dictionaries are also selective, e.g. the *Etymologisches Wörterbuch der deutschen Sprache* (Kluge 1995) and the *Herkunftswörterbuch der deutschen Sprache* (DuE), or they are historical, such as Paul's *Deutsches Wörterbuch* (Paul 2002).' On the process of the compilation of the DIFIT and the lexicographical sources, also see Heinz 2008.

& Seymer 2007 and in Heinz & Gärtig 2014). At present, the database contains 8,951 Italianisms derived from 4,660 Italian etyma, 3,145 of those in English, 2,525 in French and 3,281 in German.



Figure 2: OIM, s.v. cappuccino. Accessed at: <http://www.italianismi.org/scheda.aspx?id=3963> [19/03/2018].

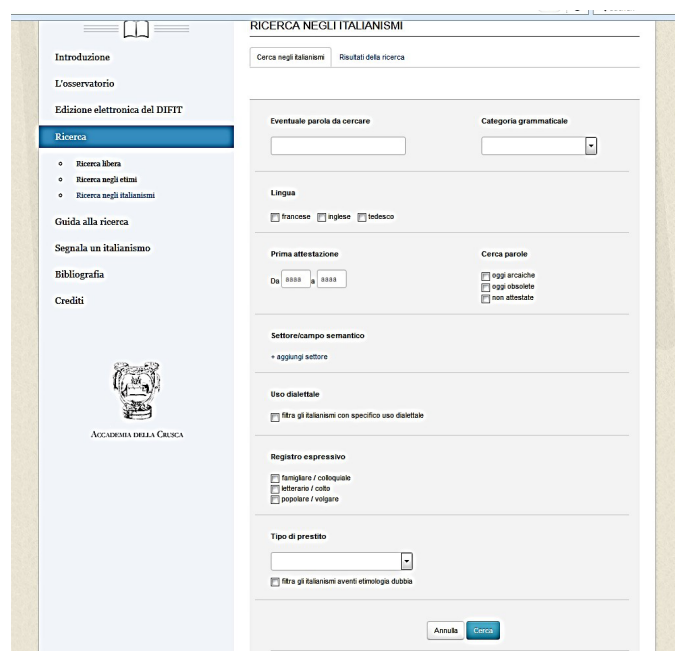


Figure 3: OIM, search mask. Accessed at: [www.italianismi.org/ricerca-italianismi](http://www.italianismi.org/ricerca-italianismi) [19/03/2018].

In 2014 a new project phase was given the green light, coordinated by Matthias Heinz, one of the authors of the DIFIT, and by Luca Serianni, who in the beginning of the new millennium had run a similar project on Italianisms in a much larger number of languages, which was, however, shelved in 2008 (cf. Serianni 2017; for more details on OIM and its new targets see Heinz 2017 & Pizzoli 2017). The new project, on the one hand, investigates Italianisms in further major European target languages. Since 2017, three working groups around Gloria Claveria Nadal, Yorick Gomez Gane, Gianluca Miraglia and Paolo Silvestri, Elżbieta Jamrozik and Zsuzsanna Fábíán have been working on the integration of borrowings in Spanish, Portuguese, Catalan, Polish and Hungarian. On the other hand, in 2017 the project group also began to review the existing data and rethink some aspects of the database structure and its sources. This process is coordinated by Matthias Heinz and Lucilla Pizzoli and is being carried out mainly in Florence and Salzburg.

This paper focusses on the data surrounding Italianisms in German,<sup>2</sup> which the author is working on within the current project phase, shows problems and restrictions in the current form of the database, and illustrates how they are dealt with in the revision process.

## 2 The Representation of Italianisms in German

As mentioned above, OIM at present contains 3,281 Italianisms in the target language, German. Searching them with the database's filters, one can gain a clear picture of how those borrowings are structured within German and can compare them to those in English and French. Figure 4 shows the division according to domains of use.

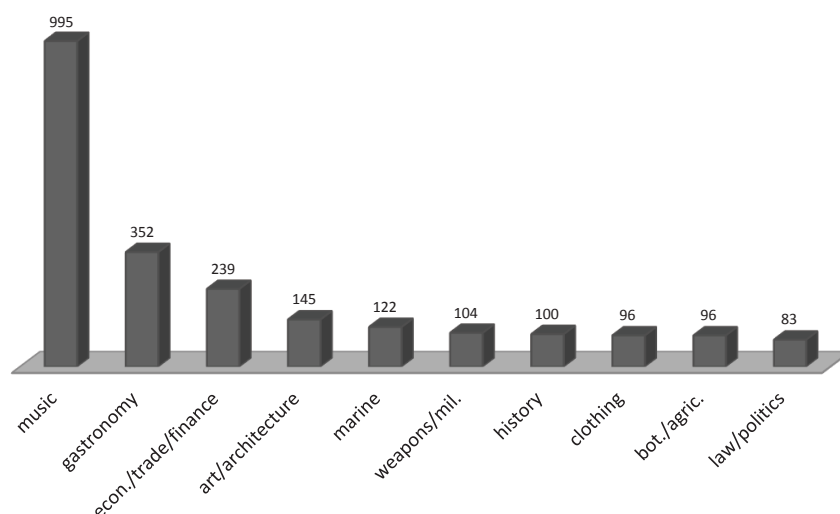


Figure 4: Italianisms in German, frequency of semantic fields/domains of use in OIM (cf. Gärtig 2017: 368).

Borrowings from the field of music (e.g. *accelerando*; *Viola da gamba*) are by far the most frequent ones, followed by Italianisms in the area of gastronomy (e.g. *Brokkoli*; *Ravioli*), economy and finance (e.g. *Konto*; *netto*) and art and architecture (e.g. *Aquarell*; *Fresko*). The predominance of musical terminology is of course explainable by the important influence the Italian language has exerted in this field, but it is also due to the sources used by the DIFIT: In order to compensate for shortcomings in the general lexicography of German, lexicographers and studies on music were systematically taken into consideration, whilst works on other semantic fields were used far less extensively.<sup>3</sup>

The filter for the years of first attestation allows users to create a timeline of the periods with the highest influence of the Italian language on the German lexicon. As Figure 5 shows, peaks were reached in the 15<sup>th</sup> and 16<sup>th</sup> centuries, when Renaissance Italy became the major center of cultural expansion, and again and even stronger in the 19<sup>th</sup> century, in which most of the musical terms deriving from Italy were registered in German dictionaries for the first time. The broken line for the 20<sup>th</sup> century reveals a decreasing number of Italianisms, but also reveals the lack of precise, reliable data of first attestation for this period, which makes it impossible to specify a number: “va ricordato che le fonti non sempre forniscono datazioni, specie per il tedesco [...]. Oltre a qualche datazione sommaria come ‘medio alto tedesco’, invece di un secolo preciso, la collocazione delle datazioni mancanti del tedesco è da

<sup>2</sup> For a detailed bibliography of the research on Italianisms in German, see Gärtig 2017.

<sup>3</sup> The DIFIT's bibliography lists 18 sources on music, eight of them regarding German, cf. DIFIT: XXV-XXXIX; for further explanations cf. Heinz 2008: 167-170.

cercare soprattutto nel Novecento [...]” (Stammerjohann & Seymer 2007: 47)<sup>4</sup> But as mentioned above, not only the newer Italianisms in German are registered without first attestation: The graph includes a total number of 2,311 borrowings, which means that almost one third of all the Italianisms in this recipient language in OIM do not have a dating, and the present system, in a chronological query, does not label them explicitly.

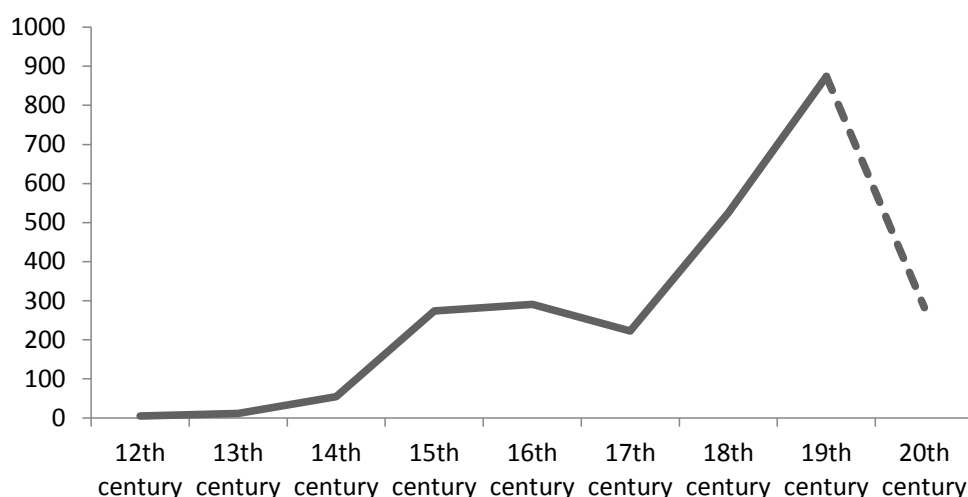


Figure 5: Italianisms in German, chronology of first attestations in OIM (cf. Gärtig 2017: 374, based on Stammerjohann & Seymer 2007: 47-48).

A very useful search option for a target language such as German, which presents a differentiated internal diatopic variation and is also a polycentric language, is the filter for Italianisms that have entered only one of its varieties. Aside from the unfortunate labelling as filter for special “uso dialettale”, ‘dialect use’, which doesn’t represent the linguistic reality, it shows that at present about 6% of the registered Italianisms belong to only one variety, most of them (119 or 64%) to Austrian German, 37 or 20% to Southern German varieties, 17 or 9% to Suisse German and 7 or 4% to the German of South Tyrol (cf. Gärtig 2017: 369-370). It is obvious that the South Tyrolian area is especially underrepresented, and the new project phase has to work on a balanced integration of its specific Italianisms.

An overall criticism of today’s version of the OIM, already remarked in a review on the DIFIT (Marazzini & Marelllo 2011: 164-165), is that it could represent a kind of museum-like lexicon, without information on the effective use of the individual Italianisms as found in Görlach (2001) which includes speaker’s surveys. Ten years after the publication of the DIFIT, another potential problem is related to the immediacy of its data, which has been the basis for the OIM until now. One just has to go to a modern café in a German-speaking area to see a *Barista* at work, ask for a *Latte macchiato* – or do you prefer *Chai Latte*? – and observe if he or she is able to produce the perfect *Crema* to find quite a number of Italianisms not registered in the actual database.

The review and extension process, thus, has to include the search for attestations and their datings, for information on the effective use of Italianisms in German. It should also create a balanced basis of sources that respects the current lexicon as well as the actual usage and frequency in certain geographical and semantic areas.

4 ‘It should be remembered that sources do not always provide dating, especially for German [...]. In addition to some summary dating as Middle High German, instead of a precise century, missing German dating is to be found especially in the twentieth century [...].’



### 3 Revision and Extension of the Database Structure and its Data

The fundamental question that has to be answered at the beginning of a revision regards the basic structure of the DIFIT, which uses mainly lexicographic works for sourcing Italianisms to be integrated. Should this decision be maintained, or should text corpora, readily available today and also relatively easily to incorporate into an online dictionary, be taken into consideration? One should bear in mind that the OIM structure should remain stable as it will be possible to add other source languages step by step, perhaps even languages of which the accessibility of corpora and lexicographic description greatly varies. It has been a strength of the DIFIT that its rather rigid structure and having been based on dictionaries guarantee a high level of comparability and reliability.

For this reason, the decision was reached to keep this basic structure, at least for now. New data will be mainly extracted from lexicographical sources, but corpora will be checked to gain attestations and information on the real use. The entries, however, shall be completed with fields on the “grado di stabilizzazione dei termini e la loro diffusione tra le varietà della lingua” (‘degree of stabilization of the terms and their diffusion among the varieties of the language’, Pizzoli 2017: 177) and with references to more detailed studies and/or lexicographic description in other (electronic) dictionaries.

Since the DIFIT went to press, there has been a notable increase in innovation within German lexicography. A substantial number of publications that can be used as sources for the OIM’s extension have emerged. To mention only a few, from 2011 to the present, there is the revised edition of the *Deutsches Fremdwörterbuch* (DFWB), published until the letter *i* and accessible across the portal OWID, hosted by the IDS, where you can also find the electronic German dictionary *ellexiko* and the online version of the *Neologismenwörterbuch* on the new and latest entries in the German lexicon. For a systematic investigation into Italianisms in common language, the new editions of *Wahrig* and of the various DUDEN dictionaries, in print and online, are very useful, as is the *Digitales Wörterbuch der deutschen Sprache* (DWDS), a digital information system on the German vocabulary with direct access to a vast part of its base corpora and to detailed word profiles and diachronic curves, elaborated since 2007 at the *Berlin-Brandenburgischen Akademie der Wissenschaften*. For the integration of Italianisms in singular varieties of German, the project can make use of the new edition of the *Variantenwörterbuchs des Deutschen* (2016). For those in South Tyrol, the linked dissertation of Abfalterer (2007) is of interest. For the verification of its entries, the *Korpus Südtirol*, developed at the EURAC in Bolzano, is used, whilst for the other entries the corpus platform of the DWDS is consulted.

#### 3.1 Completion of the German Data

The work flow is divided into the revision, correction and completion of the data on Italianisms already present in OIM, and into the addition of new borrowings, extracted from the works mentioned above. For the revision process, an important aim is to guarantee correct and meaningful answers to research questions on the chronology of borrowing. For this purpose, all the entries at present listed without an attestation are extracted and looked up in newly available dictionaries. That has led for instance to attestations of *Bruschetta*, *Calzone*, *Ciabatta* or *Papamobil* (in DUDEN). If the dictionaries do not include a dating, it is searched for in corpora. The same process is carried out for Italianisms without datings in the OIM. Thanks to the DFWB, for example, it was possible to date *Futurismus* precisely to 1912, instead of the vague indication of 20<sup>th</sup> century in the OIM, *Fiasko* in the meaning of ‘failure’ to 1819 (instead of “sec. XIX” in the OIM) or *Furore machen*, partial calque of It. *far furore*, to 1830 (instead of 19<sup>th</sup> century, too) with the help of the DWDS corpus search.<sup>5</sup>

5 <[https://www.dwds.de/r?q=Furore&corpus=public&date-start=1473&date-end=1900&genre=Belletristik&genre=Wissenschaft&genre=Gebrauchsliteratur&genre=Zeitung&format=kwic&sort=date\\_asc&limit=50](https://www.dwds.de/r?q=Furore&corpus=public&date-start=1473&date-end=1900&genre=Belletristik&genre=Wissenschaft&genre=Gebrauchsliteratur&genre=Zeitung&format=kwic&sort=date_asc&limit=50)> [22/03/2018].

Another desideratum for German, on which the current phase is working, is a uniform labelling of its varieties, indicated for the single Italianisms not found in the whole German speaking area. The labels, on a first level, should classify those borrowings as belonging only to one macro area (Germany/Austria/Switzerland/South Tyrol; for the areal conception cf. Ammon, Bickel & Lenz 2016: XXV-XXVII) and then, on a second level, add a finer classification for those units we have more precise information about. An example is *Bassena* (< Fr. *bassin* < It. *baccina*), classified as Austrian and as used particularly in Vienna by DUDEN (s.v.).

In addition to the improvement and completion of the data itself, as mentioned above, the integration of some new lexicographic information is planned. At present it is still being worked out how the real use and diffusion of single Italianisms in the target languages and in their varieties can be represented according to a scheme. In a similar way, the reliability of their attestation (whether they are registered in dictionaries, in reference corpora, in other corpora or only in single texts and contexts) should be represented in a way that can be queried by means of the web interface and that is comparable for the different target languages. This information can be useful when applied to quantitative analysis.

However, they do not provide in-depth information on the real use of single borrowings. As an online tool, the OIM should take advantage of the possibility to add qualitative information to the Italianisms for which they are available. For example, there could be bibliographical references to research on single arguments, similar to the corpus studies done by Rovere (2012) on *Cappuccino*, *Galleria* and *Bambini*. There could be links to portals like OWID or DWDS, that can provide diachronic curves, more precise etymological explanations, authentic examples and corpus hints or information on syntagmatic combinations, word formation processes and products. Finally, there could be analyses coming from the project itself, reassumed in info boxes. For example, at the present a master's student from Salzburg, Katharina Kofler, is carrying out a sociolinguistic study on the real use of Italianisms in the German standard variety of Austria (supervisor of the thesis: M. Heinz). Using an online questionnaire, Kofler is asking if the single forms are known and if they are effectively being used by Austrian speakers, if their meanings are known and which geosynonyms are mostly used. Up to the present (end of March 2018) she has collected the answers and sociolinguistic data from more than 300 participants.

### 3.2 Addition of Italianisms in German

Parallel to the completion of the missing information on Italianisms already registered, the project phase intends to complete missing borrowings. This work is being done on three levels: borrowings in the common language, borrowings of certain domains of use and borrowings of certain diatopic varieties.

In the first group, Italianisms of the common language, mostly from the 20<sup>th</sup> and 21<sup>st</sup> centuries, have been added when they were not yet lexicographically described or not in use when the DIFIT was compiled. To collect them, by now all the entries marked as derived from Italian in the DUDEN *Online-Wörterbuch* and in the *Neologismenwörterbuch* have been extracted and compared to the stock in the OIM. The result was about 60 hits, almost all nouns, with the majority belonging to the semantic field of gastronomy: *Aceto balsamico* (with the variants *Balsamico*, *Balsamessig*, *Balsamicoessig*), *Caffè Latte*, *Chai Latte*, *Crema*, *Latte macchiato*, the Austrian *Melanzani*, *Pancetta*, *Pannacotta*, *Pecorino*, *Peperonata*, *Sugo*, *Vitello tonnato*; but there is also *Gabione* 'a cage of wire filled with rocks for use in civil engineering, road building and landscaping' or the adjective *papabel*, belonging to the use of the Catholic Church and referring to a cardinal with the possibility of being elected Pope.

In order to prepare the new entry, for every loan a lexicographical sheet with the following information is prepared: Italian etymon – borrowed form in the target language with possible variants – pronunciation – type of borrowing – reliability of attestation – reference to bibliography and further literature – meaning – grammatical category and genus – domain of use / semantic field – source – dating – first attestation – diffusion – register.

The second step of systematic research aims at balancing the semantic fields of Italianisms in German. As mentioned previously, at the moment of the compilation of the DIFIT, only musical terminology was considered in major terms; this valuable work of excerption from specific lexicons and texts should be replenished with collections of terms belonging to other semantic areas and domains known to have been particularly influenced by the Italian culture and language, especially the areas of gastronomy, economy and finance, art and architecture. Currently another master's student from Salzburg, Patricia Bagari, is writing her master thesis (supervisor: M. Heinz) on the Italianisms of the last mentioned area. She is collecting terms from Sandrart's historical *Teutsche Academie* (1675), from dictionaries and lexica on arts, such as the *Kunstlexikon* (Hartmann 1996), the *Lexikon der Kunst* (Olbrich <sup>2</sup>2004), the *Großes Bildwörterbuch der Architektur* (Koepf & Binding <sup>4</sup>2005) and Reclam's *Wörterbuch der Architektur* (<sup>15</sup>2015), retracing their etymology and adding corpus studies. She has, at the moment, prepared about 80 entries to insert into the database.

In relation to Italianisms in single standard varieties of German, the focus at present is on the South Tyrolean area. An important source is Abfalterer (2007), who has also worked on the South Tyrolean part of the *Variantenwörterbuch des Deutschen* (<sup>2</sup>2016), and whose work includes a glossary of 303 *primäre Südtirolismen*, 26 of which are to be classified as loanwords (e.g. *ACI*, *Aranciata*, *Dopolaro*, *Kondominium*) and 98 as calques (*Ausgeher*, *Autobüchlein*, *Berufsalbum*, *grüne Nummer*) from Italian, using the OIM's classification system (cf. Abfalterer 2007: 167-168). They have been integrated to the OIM database, using its fixed set of semantic labels, which sometimes differ from those used by Abfalterer, and have started to be checked for datings in the *Südtirol Korpus*. Many of the Italianisms, especially those created by loan translation, belong to the fields of administration, politics, finance, professions and the education system.

Ein Grund für das gehäufte Auftreten von Lehnbildungen bei Verwaltungsausdrücken ist sicher in der besonderen Situation Südtirols mit seinem institutionellen Überbau durch den italienischen Staat zu suchen. Mit dem Inkrafttreten der Durchführungsbestimmungen, die die Gleichstellung von deutscher und italienischer Sprache rechtlich festgelegt haben, ist es notwendig geworden, sämtliche Ausdrücke aus Verwaltung, Finanz, Politik und Bau- und Rechtswesen zu übersetzen. Dies stellt keine leichte Aufgabe dar und führt manchmal zu eigenartigen Wortgebilden. (ib.: 174)<sup>6</sup>

The possibilities to query the OIM database have to take into account this special situation of South Tyrol. In fact, mixing them with those of the other varieties without any remark could cause distortion in search requests. For example, it would create a much higher number of loan translations and administrative terminology. For this reason, it seems desirable to insert a filter which includes or excludes Italianisms used only in South Tyrol, or restricts the research only to them. In future studies, comparable solutions could have to be found for other polycentric languages among the target languages in OIM (cf. as an example Pierno 2017, with a study on Italianisms in the English and French used in Canada), and filters could work from top to bottom, making it possible to select entire target languages, entire standard varieties or single regional and dialectal varieties.

6 'One reason for the increased appearance of calques in administrative terminology is certainly to be found in the special situation of South Tyrol with its institutional superstructure by the Italian state. With the entry into force of the implementing regulations which established equality between German and Italian, it has become necessary to translate all terms from administration, finance, politics, construction and law. This is not an easy task and sometimes leads to peculiar word formations.'

## 4 Conclusion and Further Perspectives

For the target languages already included, such as German, which has been described here, the purpose of the revision can be summarized as follows:

- the completion of the supply of Italianisms, covering today's common language, relevant domains of use and diatopic (standard) varieties;
- the completion of missing attestations and datings for all Italianisms based on the integration of recent documentation;
- the possibility to make search queries using filters for single diatopic (standard) varieties of the target languages;
- a representation of entries that includes details on the effective use of a single Italianism in the target language and on the reliability of the given data (whether taken from dictionaries/corpora, etc.);
- the creation of a technical mechanism that points in a stable way to more information on single Italianisms, such as bibliographic references or persistent URLs that lead to more detailed online dictionary hits or information from studies carried out in the context of the project.

In parallel to the revision process on Italianisms in German, French and English, the OIM project is working on an extension of the target languages for borrowings from Italian in concentric circles. Therefore an international research network, composed of scholars from the universities of Florence, Rome, Salzburg, Dresden, Warsaw, Budapest, Malta, Seville, Toronto and New York, has been built up. By the summer of 2018, the mentioned work on Spanish, Portuguese, Catalan, Hungarian and Polish should be completed, and there should be the new design of the website and database structure. In the following step, the circle of languages shall be expanded to those for which collections of Italianisms are already available, such as Maltese, a language with a very strong Italian impact.

## References

- Abfalterer, H. (2007). *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht. Lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols*. Innsbruck: innsbruck university press.
- Ammon, U., Bickel, H. & Lenz, A. (eds.) (2016). *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2., völlig neu beab. u. erw. Aufl. Berlin & Boston: De Gruyter.
- Bagari, P. (in prep.). *Arte s.f. Studien zur Wortgeschichte und Entlehnungstypologie der kunstterminologischen Italianismen im Deutschen*. Master thesis. Paris Lodron Universität, Salzburg, A.
- BW = O. Bloch & W. von Wartburg (1996). *Dictionnaire étymologique de la langue française*. 11<sup>e</sup> éd. Paris: PUF.
- Christmann, H. H. (1992). Italienische Sprache und Italianistik in Deutschland vom 15. Jahrhundert bis zur Goethezeit. In K. Schröder (ed.) *Fremdsprachenunterricht 1500–1800. Vorträge gehalten anlässlich eines Arbeitsgesprächs vom 16. bis 19. Oktober 1988 in der Herzog August Bibliothek Wolfenbüttel*. Wiesbaden: Harrassowitz, pp. 43–55.
- DFWB = *Deutsches Fremdwörterbuch*, begr. von H. Schulz, fortgef. von O. Basler. 2. Aufl., völlig neu erarbeitet im Institut für Deutsche Sprache (2011 -). Accessed at: <http://www.owid.de/wb/dfwb/start.html> [19/03/2018].
- DIFIT = Stammerjohann, H. et al. (2009). *Dizionario di italianismi in francese, inglese, tedesco*. Firenze: Accademia della Crusca.
- DUDEN *Online-Wörterbuch*. Accessed at: <http://www.duden.de> [19/03/2018].
- DWDS = *Digitales Wörterbuch der Deutschen Sprache*. Accessed at: <http://www.dwds.de> [19/03/2018].
- elexiko. Accessed at <http://www.owid.de/wb/elexiko/start.html> [19/03/2018].
- Gärtig, A.-K. (2017). Italianismen im Deutschen: Potentiale und Grenzen der Analyse mithilfe der Datenbank OIM. In *Studi Germanici*, 12, pp. 349–381.



- Görlach, M. (ed.) (2001). *Dictionary of European Anglicisms. A Usage Dictionary of Anglicisms in Sixteen European Languages* (DEA). Oxford: Oxford University Press.
- GR = *Le Grand Robert de la langue française – Dictionnaire alphabétique et analogique de la langue française*. 2<sup>e</sup> édition entièrement revue et enrichie par Alain Rey. Paris: Le Robert 1989 [Réimpression 2001].
- Gusmani, R. (1986). *Saggi sull'interferenza linguistica*. Seconda edizione accresciuta. Firenze: Le Lettere.
- Hartmann, P. W. (1996). *Kunstlexikon*. Maria Enzersdorf: Hartmann.
- Heinz, M. & Gärtig, A.-K. (2014). What a multilingual loanword dictionary can be used for: searching the Dizionario di italianismi in francese, inglese, tedesco (DIFIT). In A. Abel, C. Vettori, N. Ralli (eds.). *Proceedings of the XVI EURALEX Congress: The User in Focus, Bolzano/Bozen 15-19 July 2014*. Bolzano/Bozen: EURAC research, pp. 1099-1107.
- Heinz, M. (2008). L'expérience du Dizionario di italianismi in francese, inglese, tedesco (DIFIT): objectifs, structure et aspects méthodologiques. In F. Pierno (ed.) *Aspects lexicographiques du contact entre les langues dans l'espace roman*. Strasbourg: Université Marc Bloch, pp. 165-180.
- Heinz, M. (2017). Dal DIFIT all'OIM: sfide lessicografiche e prospettive di implementazione. In M. Heinz (ed.) *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti, Atti dell'incontro OIM (Firenze, 20 giugno 2014)*. Firenze: Accademia della Crusca, pp. 21-38.
- Koepf, H. & Binding, G. (2005). *Großes Bildwörterbuch der Architektur. Mit englischem, französischem, italienischem und spanischem Fachglossar*. Stuttgart: Kröner.
- Kofler, K. (in prep.). Zur Vitalität von regionalspezifischen Italianismen im österreichischen Deutsch. Eine sozio-linguistische Untersuchung zu den Einträgen aus dem *Osservatorio degli Italianismi nel Mondo*. Master thesis. Paris Lodron Universität, Salzburg, A.
- Korpus Südtirol. Accessed at: <http://www.korpus-suedtirol.it/> [19/03/2018].
- Lehnwortportal Deutsch. Accessed at: <http://lwp.ids-mannheim.de> [19/03/2018].
- Marazzini, C. & Marelllo, C. (2011). Dizionario di italianismi in francese, inglese, tedesco, a c. di H. Stammerjohann e E. Arcaini, G. Cartago, P. Galetto, M. Heinz, M. Mayer, G. Rovere e G. Seymer [...]. In *Lingua e stile*, 46 (1), pp. 162-169 [recensione].
- Neologismenwörterbuch* = D. Herberg, M. Kinne, D. Steffens (2004). *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. Unter Mitarbeit von E. Tellenbach und D. al-Wadi. Berlin & New York: De Gruyter; D. Steffens & D. Al-Wadi (2013). *Neuer Wortschatz. Neologismen im Deutschen 2001–2010*. 2 vols. Mannheim: Institut für Deutsche Sprache. Accessed at: <http://www.owid.de/wb/neo/start.html> [19/03/2018].
- OED = *The Oxford English Dictionary*. Accessed at: <http://www.oed.com> [19/03/2018].
- OIM = *Osservatorio degli italianismi nel mondo*. Project hosted by Accademia della Crusca, coordinators: L. Serianni & M. Heinz. Online portal: <http://www.italianismi.org> [19/03/2018].
- Olbrich, H. (2004). *Lexikon der Kunst*. 7 vols. Leipzig: Seemann.
- Pierno, F. (2017). Gli italianismi nell'inglese di Toronto e nel francese di Montreal. Stato delle ricerche e del progetto sugli italianismi in Canada. In M. Heinz (ed.) *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti, Atti dell'incontro OIM (Firenze, 20 giugno 2014)*. Firenze: Accademia della Crusca, pp. 111-136.
- Pizzoli, L. (2017). Per un dizionario degli italianismi nel mondo: rilancio di un progetto. In *Testi e linguaggi*, 11, pp. 171-182.
- Rovere, G. (2012). Per una lessicografia del contatto linguistico. A proposito di *Cappuccino, Galleria e Bambini* in tedesco. In L. Cinato (ed.) *Intrecci di lingua e cultura. Studi in onore di Sandra Bosco Colettos*. Roma: Aracne, pp. 245-260.
- Sandart, J. v. (1675). *L'Academia Todesca. della Architectura, Scultura & Pittura: Oder Teutsche Academie der Edlen Bau- Bild- und Mahlerey-Künste*. 6 vols. Accessed at: [http://www.deutschestextarchiv.de/book/show/sandart\\_academie0101\\_1675](http://www.deutschestextarchiv.de/book/show/sandart_academie0101_1675) [19/03/2018].
- Serianni, L. (2017). L'italiano nel mondo. Intenti e propositi di un progetto editoriale sugli italianismi. In M. Heinz (ed.) *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti, Atti dell'incontro OIM (Firenze, 20 giugno 2014)*. Firenze: Accademia della Crusca, pp. 39-54.
- Stammerjohann, H. & Seymer, G. (2007). L'italiano in Europa: italianismi in francese, inglese e tedesco. In N. Maraschio (ed.) *Firenze e la lingua italiana fra nazione e Europa. Atti del convegno di studi, Firenze, 27-28 maggio 2004*. Firenze: Firenze University Press, pp. 41-55.



- TLF = *Trésor de la langue française. Dictionnaire de la langue du XIX<sup>e</sup> et du XX<sup>e</sup> siècle* (1789-1960). Publié sous la direction de P. Imbs & B. Quemada. 16 vols. Nancy: CNRS Editions / Paris: Gallimard 1971-1994.
- Van der Sijs, N. (2010). *Nederlandse woorden wereldwijd*. Den Haag: SDU Uitgevers. Accessed at: [https://pure.knaw.nl/portal/files/458170/Nww\\_compleet\\_archief.pdf](https://pure.knaw.nl/portal/files/458170/Nww_compleet_archief.pdf) [19/03/2018].
- Wiegand, H. E. (2001). Sprachkontaktwörterbücher: Typen, Funktionen, Strukturen. In B. Igla, P. Petkov, H. E. Wiegand (eds.) *Theoretische und praktische Probleme der Lexikographie. 1. internationales Kolloquium zur Wörterbuchforschung am Institut Germanicum der St. Kliment-Ohridski-Universität Sofia, 7. bis 8. Juli 2000*. Olms, Hildesheim & Zürich & New York: Olms (= *Germanistische Linguistik*, 161 - 162), pp. 115-224.
- Winter-Froemel, E. (2011): *Entlehnung in der Kommunikation und im Sprachwandel*. Berlin & New York: De Gruyter.
- Wörterbuch der Architektur* (<sup>15</sup>2015). Stuttgart: Reclam.



# The Treatment of Politeness Elements in French-Korean Bilingual Dictionaries

**Hae-Yun Jung, Jun Choi**

*Kyungpook National University*

*E-mail: haeyun.jung.22@gmail.com, c-juni@hanmail.net*

## Abstract

Expressions of politeness in French and Korean are lexically as well as conceptually different. For example, there is no equivalent word to the French *s'il te/vous plaît* ('please') in Korean. This particular expression of politeness can be translated in many ways in Korean, which are not necessarily lexical and can also correspond to syntactic elements. At the same time, polite terms of address in French such as *Monsieur/Madame* ('Sir/Madam') have a variety of equivalents which depend on the relationship between the interlocutors and denote the various levels of politeness in Korean. Because of these discrepancies, the lexicographical description of politeness related lexical items presents many issues and shortcomings with regard to their practical use from the perspective of French learners of Korean. The objective of this paper is thus to analyze how lexical items of politeness are treated in French-Korean lexicography and what issues the relevant entries present. The equivalents and examples analyzed are extracted from the Naver dictionary portal, which gathers information from existing published bilingual dictionaries. As a result of the analysis, this study proposes a solution to the issues which the entries of such expressions present by suggesting a model for the lexicographical treatment of politeness elements.

**Keywords:** Bilingual dictionary, French-Korean lexicography, Politeness, Pragmatic information, Micro-structural model

## 1 Introduction

Politeness research has yielded a number of theories, among which Brown and Levinson's Theory of Face is probably the most cited, whether as a supporting theory or one to refute. One of the major criticisms of the theory of politeness as 'facework' regards the Western bias of its treatment of politeness, despite its claim for universality (Choi 2002: 275; Leech 2005: 4-5; Brown 2011: 61; Culpeper 2011: 15). Despite divergences among opinions on theories of politeness, there is one common thread that guides the understanding of what politeness is. That thread is the notion that the fundamental function of politeness is to maintain good communicative relationships (Leech 2005: 7; Brown 2005: 1410; Sohn 1999: 407). What differs is the way this function is manifested behaviorally and linguistically.

French and Korean in particular are typologically at opposite ends, and it is not surprising that linguistic expressions of politeness are quite different in each language. Although both can express politeness through grammatical elements (e.g. sentence types, mood choice) and lexical items, one-to-one equivalent structures or words are hard to find. Another major difference is that the Korean language has a particularly rich system of honorifics which reflects the vertical polarity of linguistic politeness, as opposed to the horizontal *tu-vous* polarity in French. As contact between various cultures and languages becomes increasingly frequent, the teaching of fundamental communicative resources such as politeness expressions and markers has become likewise essential. While politeness constitutes an important part of KFL (Korean as a Foreign Language) syllabuses, the lexicographic treatment of politeness remains fairly scarce and, indeed, French-Korean/Korean-French dictionary

entries for politeness related headwords present many shortcomings which reflect the linguistic gap between the two languages. This study analyses these weaknesses as they currently exist in the Naver French-Korean dictionary online, which will be presented in Section 2. In particular, Korean equivalents of the French politeness marker *s'il-vous-plaît* (please) will be thoroughly examined in Section 4 after a brief cross-linguistic account of politeness markers (Section 3) in order to provide an alternative microstructural model (Section 5) that can help French learners of Korean use that headword in strategic communicative situations.

## 2 Naver dictionary

This study aims to analyze headwords denoting politeness as described in the Naver French-Korean dictionary (hereafter, NFKD). Naver is a Korean portal that provides free access to Korean and bilingual (to and from Korean and other languages) dictionary services, including a French-Korean and Korean-French dictionary. The French-Korean dictionary includes 73,000 headwords based on *Prime Dictionnaire Français-Coréen* published by Doosan Dong-a Press. As for the Korean-French dictionary, it consists of 118,000 headwords mainly based on *Nouveau Dictionnaire Coréen-Français* published by Hankuk University of Foreign Studies Press, but also supplemented by the Korean-French *Essence* published by Minjung Seorim Press. These dictionaries are the main printed Korean-French bilingual dictionaries in use. French headwords are complemented by monolingual definitions from Wiktionary.

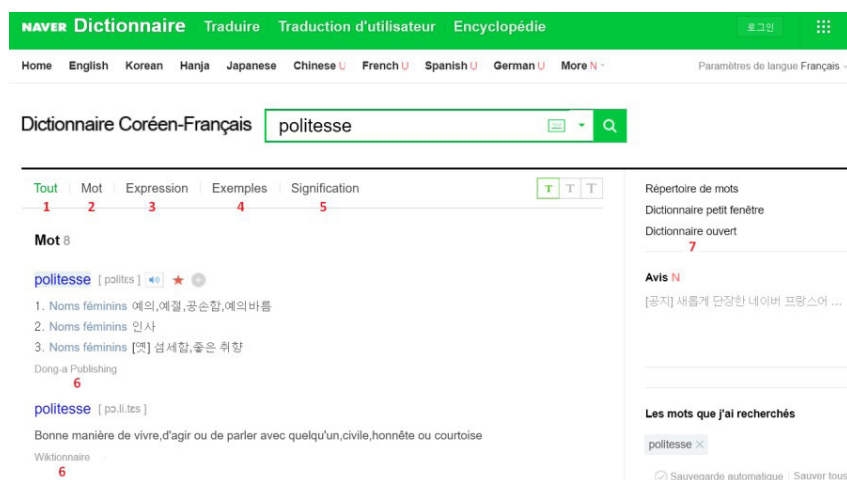


Figure 1: Interface of Naver Korean-French/French-Korean dictionary.

The entry is divided into five tabs. The first gives an overview of all sections (no. 1 in the figure above). The second tab (*Mot*) shows the equivalents for the searched headword and various expressions containing the headword (no. 2). The source is indicated under each subentry, as seen in no. 6. The third tab (*Expression*) gathers a number of idiomatic expressions and proverbs (no. 3). The fourth tab (*Exemples*) lists all examples containing the headword (no. 4). These examples are from both *Prime Dictionnaire Français-Coréen* and *Nouveau Dictionnaire Coréen-Français*. The last tab is called ‘Signification’ (no. 5) but in fact shows the *Nouveau Dictionnaire Coréen-Français* translations and Wiktionary’s descriptions of the equivalents given by *Prime Dictionnaire Français-Coréen*.

As a web dictionary, the contents provided by Naver have been expanded beyond what paper dictionaries could offer. Indeed, now it also includes a French open dictionary (no. 7) that is composed of new types of headwords, such as neologisms, buzzwords, and acronyms, updated with the participation of

students of the Korean Language Center at the University of Lyon 3. In addition, it also contains information from Wikipedia and the RMN (Réunion des musées nationaux) on about 170,000 headwords related to place names, people's names, works of art, titles and so on. With the digitalization of the Korean-French and French-Korean dictionaries, important improvements took place as they keep being updated. Today the Naver dictionary portal for the Korean-French language pair comprises a total of 1,754,850 headwords<sup>1</sup>. However, there are still a number of microstructural issues, in particular shortcomings related to pragmatic information which need to be addressed, including politeness markers.

### 3 Cross-linguistic Perspectives on politeness markers

Politeness, as mentioned in the introduction, can be behavioral and/or verbal. In the latter case, French politeness can be marked lexically, by using for example the polite pronoun *vous* (you), the polite form of address *Monsieur/Madame* (Sir/Madam), and/or the interjection *s'il-vous-plaît*, as well as syntactically, with the use of the conditional mode, for instance. For the purpose of this study, we will focus on lexical items and provide a detailed lexicographic analysis of the headword study *s'il-vous-plaît* in the following section.

If we compare French with Spanish, Italian, or English, we can find that these languages possess similar markers of politeness to French. Thus, the Spanish *usted* and the Italian *lei* are equivalent in usage to the French *vous*. *S'il-vous-plaît* can be translated into *please*, *per favore*, *por favor* in English, Italian, and Spanish respectively, and their uses are rather similar, too. *Monsieur* and *Madame* also find their equivalents in all three languages, namely *señor/señora* in Spanish, *signore/signora* in Italian, and *sir/madam* in English. Those are examples of equivalents that do not pose major lexicographic issues. Nonetheless, this can be explained by a certain typological and cultural closeness between French and English, and all the more so between French and Italian or Spanish. When considering two languages that are radically different from a typological and cultural point of view, just as French and Korean are, it seems rather obvious that the lexicographic treatment of cultural items such as politeness markers is not without problems.

Before addressing these, it is necessary to expose here the basics of the Korean honorific system. This includes two categories: the speech styles or hearer honorifics (Brown 2011: 23), which reflect the position of the speaker in relation to the hearer, and the referent honorifics, which show respect to the person the speaker is talking about. There are six speech styles in total, and each style has its own sets of verb endings, as seen in the table below.

Table 1. Hearer honorifics and their forms according to sentence mood (adapted from Brown and Yeon 2016: 95)

	Declarative	Interrogative	Imperative	Exhortative
Formal style	-(su)pnita <sup>2</sup>	-(su)pnikka	-(u)sipsio	-(u)sipsita
Polite style	-a/eyo			
Semiformal style	-o/so		-o	-psita
Familiar style	-ney	-na/-nunka	-key	-sey
Intimate style	-a/e			
Plain style	-ta	-(nu)nya	-(e)la	-ca

The speaker must choose the appropriate style according to whom he speaks. Formal and polite styles can be combined together and are mostly used to address adults, whether strangers or acquaintances,

<sup>1</sup> This information can be found on the Naver Dictionary homepage by clicking on the language of interest.

<sup>2</sup> All transcriptions of Korean follow the Yale Romanization.



as well as elders, including relatives and friends' relatives, and superiors in social and professional situations. Semiformal and familiar styles are seldom used, and intimate and plain styles, which are not honorific, can only be used towards children and very close friends. The polite style remains nonetheless the basic one for most interactions. Such an elaborate hierarchical system contrasts sharply with French politeness, wherein the opposition between *tu* and *vous* reflects the degree of intimacy between the speaker and the interlocutor, that is, a horizontal polarity.

The referent honorifics occur not only grammatically (i.e., in verb ending choices) but also lexically (i.e. word choices). To show respect to the person who is the subject of the sentence (whether the addressee or a third party) the particle *-si-* needs to be added between the verb stem and the verb ending. Nonetheless, in some cases there are alternative forms to some verbs, as shown in the examples below.

- (1) *issta* (to stay, non-honorific): *kyeysita* (to stay, honorific)
- (2) *cata* (to sleep, non-honorific): *cwumwusita* (to sleep, honorific)
- (3) *mekta* (to eat, non-honorific): *capswusita* (to eat, honorific)

A number of nouns and some case particles also have honorific equivalents.

- (4) *nai* (age, non-honorific): *yensey* (age, honorific)
- (5) *salam* (person, non-honorific): *pwun* (person, honorific)
- (6) *-ka/i*, (nominative case particle, non-honorific): *-kkeyse* (nominative case particle, honorific)

None of the honorific forms for the examples 1 to 5 are given as equivalents to the French headwords *rester* (to stay), *dormir* (to sleep), *manger* (to eat), *âge* (age), and *personne* (person). This is quite surprising considering the fact that honorification occupies a significant place in the Korean language.

Besides, Korean forms of address also constitute a complex network of terms showing the acknowledgement of the interlocutor's social and relational status by and with the speaker. While in France only the titles *Monsieur/Madame* have remained, it is common in Korea to use a professional title to which the deferential suffix *-nim* is attached to address someone, especially if that profession is considered honorable by Korean society<sup>3</sup>. This shows how stratified Korean society is, and, by consequence, how socially codified the Korean language is.

#### 4 A case study: *s'il-vous-plaît*

As seen in the example below, NFKD provides three equivalents (with a paraphrase in brackets for the last equivalent) to the interjection *s'il-vous-plaît*.

- (7) *S'il-vous-plaît*: *mianhaciman* [sorry but], *ese* [eagerly], *puti* [I beg you] (*puthakipnita* [it is a request])<sup>4</sup>

As example 7 shows, the equivalents are not synonyms and correspond to very different situations. Nonetheless, they are given without further information as to how, when, and to whom they should be used. The user has to cross-check him/herself with the examples. However, in the case of *s'il-vous-plaît*, there are 200 examples from both *Prime Dictionnaire Français-Coréen* and *Nouveau Dictionnaire Coréen-Français* (Table 2), which are distributed as 20 examples per page in no particular order.

3 The issue of address terms in French-Korean dictionaries was presented by Nam K. and Jung H-Y. at the Asialex Conference in June 2015.

4 See Appendix 1 for the original presentation. Personal translations into English in Examples are provided in square brackets.

Table 2: Distribution of examples by dictionary sources.

Sources	Number of examples
Prime Dictionnaire Français-Coréen	51
Nouveau Dictionnaire Coréen-Français	149

Since it is reasonable to assume that most dictionary users will not go through the 200 examples, it is essential to look into the validity of the equivalents before analyzing the them. The following examples show how the main Korean dictionary, namely the *Unabridged Dictionary of Standard Korean*, describes the usages of these equivalents.

- (8) *mian-hata* 2. (in the forms ‘*mian-haciman*’, ‘*mian-haoman*’) Used to request one’s consent in a humble way. ¶ *mian-haciman*, *kil com pikhye cwu-sipsio* [sorry but could you let me pass please?]
- (9) *ese* 2. Used to welcome or invite someone in a cheerful way. ¶ *ese o-key* [welcome]/*ese o-sipsio* [please welcome]/*ese tu-sipsio* [please welcome]
- (10) *puti*. Used to express one’s heartfelt feelings when requesting something or asking a favor from someone, with the meanings of ‘hopefully’, ‘at any cost’, or ‘by all means’. ¶ *emenim, puti mom-cosim-ha-sipsio* [mother, please take care of your health]/*ipon moim-ey puti chamsek-haye cusiki pala-pnita* [By all means, please attend this meeting]

The first equivalent *mian-haciman* is indeed a form of politeness, just as *s’il-vous-plaît* is, and both can serve to initiate a request, especially one that might impose on the interlocutor. However, while the meaning of the two words overlaps in this case, the former has a more restricted use and the latter encompasses other functions that *mian-haciman* does not. In that sense, this equivalent should be granted a usage information gloss, all the more so because it is illustrated in the examples provided by NFKD by only one instance. The second equivalent *ese* literally means ‘quickly, promptly’, but once combined with a verb meaning ‘come in’, such as *o-ta* (to come) and *tul-ta* (to go in), the phrase assumes the meaning of ‘welcome’. By no means does *ese* alone mean ‘please’ nor the verb alone ‘welcome’; *ese* is thus not an equivalent of *s’il-vous-plaît*. In fact, only one instance of *ese* figures among the 200 examples, and it is precisely the combination *ese* + verb meaning ‘come in’: *ese tule-o-seyyo* (please, come in). Finally, *puti*, which expresses the speaker’s feelings of hope and/or the urgency of a request, corresponds to the use of *s’il-vous-plaît* when it means ‘*je vous en supplie* (I beg you)’. Again, only one example of *puti*’s use is shown among the 200 examples.

It seems then rather surprising that the equivalents provided by NFKD are shown in context only once each, considering that there are 200 examples. Nonetheless, as Szende (1999: 200) pointed out, if one of the most basic roles of examples is to show the word in a sentence, thereby serving as “*proofs*” and “*models*”<sup>5</sup>, another function of examples is to complement equivalents. Therefore, it is now necessary to analyze the examples and examine to what extent they fulfill these roles. In order to achieve this, we classified the examples according to the situation wherein *s’il-vous-plaît* is used across the two dictionary sources. For this classification, we excluded the examples of *s’il-te-plaît* as they should appear under the headword ‘*s’il-te-plaît*’. As a result, there are 46 examples from *Prime Dictionnaire Français-Coréen* and 112 examples from *Nouveau Dictionnaire Coréen-Français* to be analyzed. The statistical results are shown in the following table.

5 In italic in the text.

Table 3: Number of examples according to the uses of *s'il-vous-plaît* and dictionary sources.

	<b>s'il-vous-plaît in context</b>	<b>Prime Dictionnaire Français-Coréen</b>	<b>Nouveau Dictionnaire Coréen-Français</b>	<b>Total</b>
1	used to attract the attention of someone or a group	-	11	11
2	used on sign boards	-	1	1
3	used over the telephone (professional or private calls)	1	3	4
4	used when ordering something at a restaurant/coffeeshop	18	11	29
5	used when buying something at a counter	1	3	4
6	request (services or private domain): imperative mood or nominal sentence	4	61	65
7	mitigated request (services or private domain): interrogative or declarative forms	3	19	22
8	used to ask permission	-	2	2
9	used ironically (not to convey politeness)	2	-	2
10	unclear or ambiguous example	2	1	3
11	translation does not correspond to the example	7	-	7
12	example does not correspond to <i>s'il-vous-plaît</i> but illustrates uses of verb <i>plaire</i> (please)	8	-	8
Total		46	112	158

It has to be noted that requests have been divided into two categories – simple and mitigated requests, on the basis of the sentence type in French, since the scope of this study is limited to the perspective of French learners of Korean and takes the position of French speakers who aim to produce in Korean. In this case, mitigated requests correspond to indirect orders, the order being mitigated by the interrogative or declarative forms, verbs such as *pouvoir* (can), *vouloir* (want), *aimer* (would like), and/or the use of the conditional mood as seen in the examples below selected from NFKD.

- (11) *Pourriez-vous parler plus fort s'il-vous-plaît?* [could you please speak louder?] à interrogative form + *pouvoir* + conditional mood. (source: *Nouveau Dictionnaire Coréen-Français*)
- (12) *Pouvez-vous me couper le quatre quarts s'il-vous-plaît?* [Can you please cut the pound cake for me?] à interrogative form + *pouvoir* (source: *Prime Dictionnaire Français-Coréen*)
- (13) *Voulez-vous me dire où se trouve la poste, s'il-vous-plaît?* [Could you (literally, do you want to) please tell me where the post office is?] à interrogative form + *vouloir* (source: *Nouveau Dictionnaire Coréen-Français*)
- (14) *J'aimerais une photocopie de ce document, s'il-vous-plaît.* [I would like a photocopy of this document, please] à declarative form + *aimer* + conditional mood (source: *Nouveau Dictionnaire Coréen-Français*)

The second step of the analysis was to examine the Korean translations and find out how *s'il-vous-plaît* was rendered. In this step, categories 9 to 12 are excluded since they are not relevant or useful to the study of polite expressions, and category 1 is treated separately. The table below gives an overview of various Korean politeness markers used to translate *s'il-vous-plaît* in the context of making request, and the number of examples found in *Prime Dictionnaire Français-Coréen* and *Nouveau Dictionnaire Coréen-Français*.

Table 4: Korean equivalents to *s'il-vous-plaît* according to its uses and dictionary sources.

	<b>s'il-vous-plaît in context</b>	<b>Korean expression of politeness to render s'il-vous-plaît</b>	<b>Prime Dictionnaire Français-Coréen</b>	<b>Nouveau Dictionnaire Coréen-Français</b>
2	used on sign boards	verb stem+ <i>-si-</i> +verb ending	-	1
3	used over the telephone (professional or private calls)	( <i>puthak-hapnita</i> ) <i>com</i> + verb stem+ <i>-a/e cwu-</i> + <i>-si-</i> +verb ending ( <i>camsimanyo</i> )	1 - -	- 2 1
4	used when ordering something at a restaurant/coffeeshop	verb stem <i>cwu</i> (give)+verb ending verb stem <i>cwu</i> (give)+ <i>-si-</i> +verb ending verb stem+ <i>a/e cwu-</i> + <i>-si-</i> +verb ending ( <i>puthak-hapnita/hayyo</i> ) no particular marker	- 15 - 2 1	1 8 1 1 -
5	used when buying something at a counter	verb stem <i>cwu</i> (give)+ <i>-si-</i> +verb ending: <i>cwuseyyo</i> no particular marker	- 1	3 -
6	request (services or private domain): imperative mood or nominal sentence	( <i>camsimanyo</i> ) ( <i>ese oseyyo</i> ) <i>com</i> <i>puti</i> ( <i>puthaktulipnita/puthakhayyo</i> ) verb stem+ <i>-si-</i> +verb ending <i>ceypal</i> + verb stem+ <i>-si-</i> +verb ending <i>com</i> + verb stem+ <i>-si-</i> +verb ending verb stem+ <i>a/e po-</i> + <i>-si-</i> +verb ending verb stem+ <i>a/e cwu-</i> +verb ending verb stem+ <i>a/e cwu-</i> + <i>-si-</i> +verb ending <i>com</i> + verb stem+ <i>a/e cwu-</i> +verb ending <i>com</i> + verb stem+ <i>a/e cwu-</i> + <i>-si-</i> +verb ending <i>ceypal</i> + verb stem+ <i>a/e cwu-</i> +verb ending <i>ceypal</i> + verb stem+ <i>a/e cwu-</i> + <i>-si-</i> +verb ending <i>ceypal</i> + <i>com</i> + verb stem+ <i>a/e cwu-</i> +verb ending no particular marker	- - - - - 1 - - - - 1 - 1 - - - - 1 1 - - - - 1	2 1 1 2 13 - 5 1 7 11 2 3 3 2 2 2 6
7	mitigated request (services or private domain): interrogative or declarative mood	<i>mianhaciman</i> verb stem+ <i>-si-</i> +verb ending <i>com</i> + verb stem+ <i>-si-</i> +verb ending verb stem+ <i>a/e cwu-</i> +verb ending verb stem+ <i>a/e cwu-</i> + <i>-si-</i> +verb ending <i>com</i> + verb stem+ <i>a/e cwu-</i> + <i>-si-</i> +verb ending verb stem+ <i>a/e cwu-</i> + <i>-si-</i> + <i>-keyss-</i> +verb ending <i>com</i> + verb stem+ <i>a/e cwu-</i> + <i>-si-</i> + <i>-llayyo</i> verb stem+ <i>a/e cwu-</i> + <i>-si-</i> + <i>-l swu iss-</i> + <i>-si -</i> +verb ending <i>com</i> ( <i>puthakhayyo</i> ) no particular marker	1 - - - - 1 - - - - 1 - - - - 1 - - -	- 2 1 1 2 2 4 3 - 2 1 1 1
8	used to ask permission	verb stem+ <i>a/e cwu-</i> + <i>-si-</i> + <i>-l swu iss-</i> +verb ending verb stem+ <i>a/eto toy(be)-lkkayo</i>	- -	1 1
	Total		27	100

The words in brackets correspond to formulaic expressions that contain the notion of *s'il-vous-plaît* without having a separate element that would translate literally to *s'il-vous-plaît*: *puthak-hapnita/-hayyo* means 'could you please do me this favor'; *camsimanyo* means 'one moment please'; and *ese tuleoseyyo* means, as mentioned earlier, 'please welcome'. The 'no particular marker' category encompasses two phenomena. First, politeness is rendered only by the verb endings chosen, that is, the use of deferential speech style *-(su)pnita*; *-(su)pnikka*; *-(u)sipsio*), with no other marker that would render the specific French marker *s'il-vous-plaît*, as seen in example 15 below. Second, the Korean translation is a literal one, thereby ignoring the pragmatic dimension of the French sentence and omitting any marker to render *s'il-vous-plaît*, as illustrated in example 16.

- (15) *Pourriez-vous m'indiquer les toilettes, s'il vous plaît madame?* [Madam, could you please tell me where the toilets are, please?]  
       à hwacangsil-i                      eti-i-pnikka? (source: Nouveau Dictionnaire Coréen-Français)  
       toilets-nominative                where-copula-deferential interrogative verb ending
- (16) *Je reprendrai une part de tarte tatin, s'il-vous-plaît.* [I will have another slice of tarte tatin, please]  
       à talutu tatayng-ul                te            mek-ul ke-yeyyo. (source: Prime Dictionnaire Français- Coréen)  
       tarte tatin-accusative            more    eat-future tense marker-polite verb ending

Although Table 4 seems to show too many different ways to translate *s'il-vous-plaît* to rationalize the equivalents, we can nonetheless observe some recurrent patterns.

- The quasi systematic use of the honorific particle *-si-*. Out of the 127 examples retained, 84 cases (66.1%) contain the pre-final verb ending *-si-* as politeness marker.
- The frequent use of the lexical bundle *-a/e cwu-*, wherein the verb *cwu-* does not have the literal meaning of 'give' but functions as an auxiliary, and which can be combined with the pre-final verb ending *-si-* (50 cases, 39.4%).
- The use of the adverb *com* (literally, a little) as a politeness marker, which can also be combined with the above (24 occurrences, 18.9%). This politeness marker appears to be less strong than the other two mentioned above, as it appears alone (that is, without the use of the pre-final verb ending *-si-* and/or the lexical bundle *-a/e cwu-*) in only one case. Nonetheless, it has a function of mitigating a request, defined in the *Unabridged Dictionary of Standard Korean* as follows: "used to sound less imperative when asking for a favor or seeking one's consent". Besides, *com* appears in the examples much more frequently than *puti* which is given as an equivalent.

There are, in addition, eight occurrences of the adverb *ceypal*, which is found to be combined with any or all of the above. In fact, *ceypal* means 'hopefully' (*Unabridged Dictionary of Standard Korean*) and does not mean 'please' *per se*. Rather, it functions as an intensifier when used in making requests. In that sense, *ceypal* could be translated by *s'il-vous-plaît* while the opposite is not necessarily true. *Ceypal* is very similar in meaning and usage to *puti*, with a difference of register, *ceybal* being less formal than *puti*. Finally, when *s'il-vous-plaît* is used as an interjection to attract the attention of someone (category 1 in Table 3) as in '*S'il-vous-plaît Monsieur/Madame/Mademoiselle!*', it can be translated by many forms and phrases in Korean, most of which contain an adverb of place such as *yeki* (here) or *ceki* (there) and/or the verb *po-* (see) and can be combined with the patterns mentioned above.

## 5 Suggestions for improving the lexicographic treatment of polite expressions in NFKD

As seen in the previous sections, NFKD has many shortcomings in entries related to politeness markers and expressions. These can be categorized as follows:



- Equivalents are incomplete due to the asymmetry in politeness-related markedness. Indeed, Korean has honorific equivalents to a number of common words which are polite *per se*. These honorific equivalents are in most cases not provided in the corresponding French headword. For example, when asking someone politely over the phone if he or she is in the office, it would be inappropriate to use the verb *issta* (to stay, to be, non-honorific); instead, *kyeysita* (to stay, to be, honorific) should be used. But as mentioned in Section 3, the honorific form is provided neither for the headword *rester* (to stay) nor the headword *être* (to be).
- Equivalents are inaccurate due to the lack of pragmatic information. For instance, *ese*, as seen in the previous section, can only render the notion of the French politeness marker *s'il-vous-plaît* when used with a semantically restricted set of verbs. Hence, this type of translation should be presented in the examples section or as phraseology, but not as an equivalent (i.e., in the equivalents section).
- Equivalents are inconsistent with the examples provided in the entry. Each of the three equivalents given for *s'il-vous-plaît* appears only once or twice out of the 200 examples. While one function of examples is to show other translation choices, there should be some coherence among all the parts of the entry. Conversely, some translation possibilities presented in the examples should be given as equivalents with appropriate explanation.
- Equivalents are inaccurate because of the tendency to provide equivalents that are on the same level as the headword. That is to say, the equivalents provided for the lexical item *s'il-vous-plaît* are likewise single lexical units. However, we have seen in the previous section that it was most likely rendered not only by syntactical elements, but also by the combination of syntactical and lexical elements, and by the speech level. While French politeness can also be rendered by grammatical patterns and phrases that colligate with the exclamation *s'il-vous-plaît*, such as *pouvoir* (can) in conditional mood and interrogative sentences, it is hardly conceivable to sound polite without using *s'il-vous-plaît*, no matter the sentence structure. In Korean, it is not a single word but the pattern itself that means *s'il-vous-plaît*, and this is precisely what should be reflected in the equivalents section.

Besides these weaknesses, additional issues can be raised. In the case of *s'il-vous-plaît* in particular, there are other patterns which are common expressions of politeness and can be used in requests to render the notion of *s'il-vous-plaît*. Nonetheless, they do not appear in the examples provided by NFKD. These patterns, which can combine with those identified in the previous section, fall under the category of mitigated requests (indirect requests) and can be divided in two groups: 1) *-myeon* (if) + *-keyss-* (future tense marker) patterns (Jeon 2004: 73) as seen in 17, and 2) volitive verb patterns (Choi 2002: 286) as seen in 18.

(17) (verb stem or verb stem+*a/e cwu+si*)-*myen coh-keyss-supnita* (it would be great if...), *kamsaha/komap-keyss-supnita* (I would be grateful if...)

(18) (verb stem or verb stem+*a/e cwu+si*)-*ki palapnita* (I/we would like you to...)<sup>6</sup>

Furthermore, many of the shortcomings addressed here could be linked to one crucial issue, which is the target user of French-Korean and Korean-French dictionaries. Until the last century, most were addressed to a French-learning Korean public (Choi 2016: 194). Despite some efforts to encompass both Korean and French users, the results remain somewhat unequal. Indeed, pragmatic information is provided for only two examples out of 200, and in the Korean language only. However, bilingual French-Korean lexicography needs to evolve to reflect today's socio-relational realities, such as the growing number of university partnerships between France and Korea and the increasing number of French learners of Korean<sup>7</sup>. As mentioned earlier, the NFKD microstructural content is mainly based

<sup>6</sup> Choi (2002) notes, however, that this pattern is mostly used on formal occasions.

<sup>7</sup> See the government-run French diplomacy website: <https://www.diplomatie.gouv.fr/en/country-files/south-korea/france-and-south-korea/>

on two printed dictionaries. While the macrostructure keeps being updated and supplemented, the microstructure remains almost unchanged from the printed version.

Now, if we are to update the equivalents section of the headword *s'il-vous-plaît* for instance, what challenges do we face and what caution do we need to take? Considering the number of combinational choices in Korean, the primary focus should be on the division of these various, unlike the single list currently provided by NFKD. Furthermore, providing appropriate amounts of equivalents and grammatical information is an important factor in the readability and thereby in the effectiveness of the dictionary entry. Finally, it is necessary to review and perhaps rethink existing “supplementary meaning-elucidating strategies” (Adamska-Sałaia 2016: 156), such as glosses in brackets, usage label, and explanatory notes. On the basis of these considerations and the patterns and usage analyzed above, we propose the following microstructural model for the headword *s'il-vous-plaît*.

Table 5: Proposed microstructural model for the headword *s'il-vous-plaît*.

Microstructural labels	Equivalents
I. <i>Général</i> (general)	
1. ( <i>requête directe</i> ) (direct request)	~ (쯘) V-아/어 주세요[주십시오]. ~ (com) V-a/e cwuseyyo[cwusipsio] ~ (쯘) V-세요[십시오]. ~ (com) V-seyyo[sipsio]
2. ( <i>requête indirecte/atténuée</i> ) (indirect/ mitigated request)	~ (쯘) V-아/어 주시면 좋[감사하/고맙]겠습니다. ~ (com) V-a/e cwusimyen coh[kamsaha/komap] keysssupnita ~ (쯘) 부탁드립니다[합니다/입니다/해요] ~ (com) pwutaktulipnita[hapnita/ipnita/hayyo] ~ (쯘) V-아/어 주실래요? ~ (com) V-a/e cwusillayyo?
II. <i>Situations particulières</i> (particular situations)	
1. ( <i>demander la permission de faire quelque chose</i> ) (asking permission)	~ (쯘) V-아/어도 될까요[괜찮을까요]? ~ (com) V- a/eto toylkkayo[koaynchanhulkkayo]?
2. ( <i>passer la commande au restaurant[café]; acheter quelque chose au comptoir</i> ) (ordering at a restaurant [coffeeshop]; buying something at a counter)	~을/를* 주세요 ~ul/lul* cwuseyyo * accusative marker
3. ( <i>interpeller quelqu'un pour attirer son attention</i> ) (calling out someone to attract their attention)	여[저]기요 ye[ce]kiyo 여기 쯘 보세요 yeki com poseyyo
4. ( <i>expressions figées</i> ) (phrases)	
<i>un instant s'il-vous-plaît</i> (one moment please)	잠시만요 camsimanyo
<i>entrez s'il-vous-plaît</i> (please come in)	어서 (들어) 오세요 ese (tule) oseyyo

In the ‘Equivalents’ column, *V* stands for ‘verb’ and round brackets signal optional items (i.e., may or may not be used), while square brackets indicate the various paradigms to choose from. As politeness markers such as *s'il-vous-plaît* have a crucial role in interpersonal discourse, the rationale of this model is the use of situation-based labels.

## 6 Conclusion

This study has addressed the issue of equivalents in regard to expressions of politeness, focusing on the French headword *s'il-vous-plaît* in the French-Korean online dictionary provided by Naver. The equivalents were found inaccurate, misleading, or incomplete, as no pragmatic information is provided for this discourse resource. An analysis of the 200 examples linked to this headword has allowed us to find and refine equivalents by identifying patterns, categorize them according to usage, and finally reorganize them into a new microstructural model based on situation labels so as to overcome the ineffectiveness of its current state. This study has also touched on an underlying issue, which is the unbalanced target users of French-Korean and Korean-French dictionaries. This work is only preliminary to a more thorough and systematic rethink of French-Korean and Korean-French lexicography, and still presents some limitations in scope. Nevertheless, we hope to extend the task so as to improve and redefine French-Korean/Korean-French lexicography in terms of pragmatics and effective communication.

## References

- Adamska-Sałaiak, A. (2016) Explaining meaning in bilingual dictionaries. In P. Durkin (ed) *The Oxford Handbook of Lexicography*, Chapter 9, pp. 144-162. Oxford: Oxford University Press
- Brown, L. (2011) *Korean Honorifics and Politeness in Second Language Learning*. Amsterdam: John Benjamins Publishing Company.
- Brown, L. And Yeon, J. (2016) *Speed up Your Korean. Strategies to Avoid Common Errors*. Abingdon: Routledge.
- Brown, P. (2005) Linguistic Politeness. In U. Ammon, N. Dittmar, K. J. Mattheier and P. Trudgill (Eds.) *Sociolinguistics: An International Handbook of the Science of Language and Society*, pp. 1410-1416. Berlin/ New York: W. de Gruyter
- Culpeper, J. (2011) Politeness and impoliteness. In K. Aijmer, G. Andersen (eds.) *Sociopragmatics*, Volume 5 of *Handbooks of Pragmatics* edited by Wolfram Bublitz, Andreas H. Jucker and Klaus P. Schneider. Berlin: Mouton de Gruyter, pp. 391-436.
- Choi, H. K. (2002) A Study of Politeness in Korean Requests. In *oykukeloseuy hankukekyoyuk*, 27(1), pp. 271-299.
- Choi, J. (2016) Aperçu Historique du Dictionnaire Bilingue en Corée: Le Cas des Dictionnaires Français-Coréen et Coréen-Français. In *International Journal of Lexicography*, 29(2), pp. 184-199
- Jeon, H-Y. (2004) Hankuke kongsonphyohyenuy uymi (On the meaning of Polite Expressions in Korean). In *hankuke uymihak*, 15, pp. 71-91.
- Jung, H-Y. (2015) Forms of Address in French-Korean and Korean-French Dictionaries from the Perspective of a French Learner of Korean. In *Proceedings of Asialex 2015, Words, Dictionaries and Corpora: Innovation in reference science, 25-27 June 2015*. The Hong Kong Polytechnic University, Hong-Kong.
- Leech, G. (2005) Politeness: Is There an East-West Divide? In *Journal of Foreign Languages*, 6, pp. 3-31.
- Naver French-Korean/Korean-French Dictionary online. Accessible at <https://dict.naver.com/frkodict/french/#/main>.
- Sohn, H. (1999) *The Korean Language*. Cambridge: Cambridge University Press.
- Szende, T. (1999) Problems of exemplification in bilingual dictionaries. In *Lexicographica. International Annual for Lexicography, de Gruyter*, 15, pp. 198-228.
- Unabridged Dictionary of Standard Korean online. Accessible at <http://stdweb2.korean.go.kr/search/View.jsp>.

## Appendix 1. Screenshot of the entry for *s'il-vous-plaît* in Naver French-Korean Dictionary





# Lexicography in the French Caribbean: An Assessment of Future Opportunities

**Jason F. Siegel**

*The University of the West Indies, Cave Hill Campus*

*E-mail: jason.siegel@cavehill.uwi.edu*

## Abstract

While lexicography in the Hispanophone Caribbean has flourished, and to a lesser extent in the territories of the Caribbean whose official language is English, dictionaries of the French-official Caribbean (except Haiti) have been quite limited. But for the rest of the French-official Caribbean, there remains much work to do. In this paper, I assess the state of lexicography in the French-official Caribbean, as well as the possibilities for future work. There are six principal areas of lexicographic documentation to be developed. The first, most urgent task is the documentation of the endangered St Barth French. The next priority is multilingual lexicography for the Caribbean region. The third priority is multilingual lexicography of French Guiana, home to endangered Amerindian, Creole and immigrant languages. Fourth, there is a largely pristine area of lexicographic work for the English varieties of the French Caribbean. The fifth area of work to be developed is monolingual lexicography of French-based Creoles. Lastly, there is exploratory work to be done on the signed language varieties of the French-official Caribbean. The paper concludes with a discussion of the role that the Richard and Jeannette Allsopp Centre for Caribbean Lexicography can play in the development of these areas.

**Keywords:** minority languages, bilingual lexicography, French Caribbean

## 1 Introduction

Overseas French (*le français d'outre-mer*) is a fairly important topic in French linguistics. But so far, the French of the Antilles and French Guiana have received less attention than French-based Creoles spoken in the same region. However, it is important, especially during this UN Decade for People of African Descent, to report not only on varieties of French spoken in Haiti, Guadeloupe, Martinique, St Barthelemy, St Martin and French Guiana, but to give a full account of the lexicographic work that remains to be done in these territories called the “French-official Caribbean” (Alleyne 1985).<sup>1</sup> Indeed, given a certain quantitative decreolization (Rickford 1987), a loss of creolophones (i.e. Creole-speakers) in the face of French glottophagy, it is important to know these varieties. In particular, there is much that remains to be done in the lexicographic field. While the Spanish-speaking Caribbean has bureaus of the Royal Spanish Academy dedicated in part to documenting the lexical particularities of each country or territory, the French-official Caribbean has no such body that operates over its whole territory. There are very few dictionaries in this region, despite the fact that there are hundreds of thousands of Francophones in the area. Here, I will review the lexicographic work, whether in the form of dictionaries or large glossaries, that has already been done in the Caribbean. I will also discuss the various territories and the kinds of dictionaries that would be useful there. I will conclude with an evaluation of the role that can be played by the Richard & Jeannette Allsopp Centre for Caribbean Lexicography at the University of the West Indies, Cave Hill Campus in Barbados.

<sup>1</sup> Alleyne (1985) criticizes the notion of “French-speaking Caribbean” (and similarly for “English-” and “Dutch-speaking”) because the people in this part of the world frequently do not speak French, but are likely monolingual in Creole.



## 2 Lexicography in the French-official Caribbean Through Today

Lexicography in the Caribbean dates from the 17<sup>th</sup> century, with the publication of Breton (1665). A missionary in Dominica, Father Breton knew the Amerindians who lived there very well. He therefore learned their language during his long evangelization there, which did not have much success except for achieving a state of relative peace between the missionaries and Caribbean people. He was able to learn the Caribbean language with the help of an indigenous interpreter (Pury 1999: XXVIII), and thus he started writing a dictionary of the language, hoping that the missionaries who arrived after him could continue to evangelize in the language of the people. It is clear, therefore, that the first dictionary in the Caribbean is a bilingual dictionary. From then on, this is the norm for lexicography in the French West Indies, and only bilingual dictionaries appear until 1997 (Telchid 1997).

This first dictionary does not have the form that we know today. It is an alphabetical list of words and sentences, but it is not clear that the sentences described are lexemes. Lemmas are often two or three words, and Breton often provides translations in the form of a complete sentence (see Figure 1).

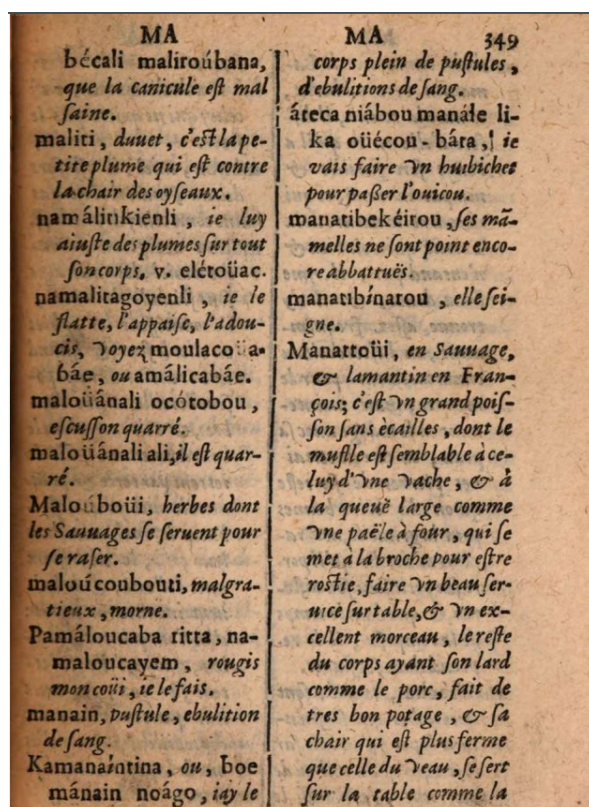


Figure 1: Extract from Breton (1665)

After France ceded control to the British of the islands where the Amerindians lived (and later died out), there was no great tradition of bilingual lexicography of Amerindian languages in the French-speaking Caribbean, because the purpose of such dictionaries or glossaries was to help missionaries evangelize those who did not know Christianity. Without access to this population, thanks to British conquests and genocide, the motivation to compile a dictionary disappeared.

There then appears almost a century later a dictionary of the Galibis of French Guiana by La Salle de l'Étang (1763). Unlike Breton (1665), this bilingual dictionary is bidirectional, so readers could search for the French word to express themselves in Galibi, or the word in Galibi to understand what the Native Americans said. There are also critiques of the language which indicate where the syntax

of Galibi is different from that of French, in addition to interlinear glosses for the examples. Another edition a century later adds Latin glosses for each French lemma.

We must wait until the 20<sup>th</sup> century for other French dictionaries in this region. Élodie Jourdain (1956) produced a vocabulary, organized by themes, of Martinican Creole. Pradel Pompilus (1958) made a lexicon of Haitian Creole. These two creoles, like the other creole languages, were long regarded as inferior to European languages. It is not surprising that French lexicographers have therefore ignored them. But the contrast between these two linguists is very important: Jordan was a *béké*, that is to say a white Martinican, and wanted to show the deformation of French by the blacks (Aub-Buscher 2003: 2). On the other hand, Pompilus was a black Haitian who was proud of his language, wanting to promote the knowledge of Haitian letters. This pride in the language of all Haitians plays an important role in the proliferation of Haitian Creole dictionaries. Today, it is the Caribbean Creole language that has by far the largest number of dictionaries (Skybina & Bitko 2014).

There is now at least one dictionary for every French Creole spoken in the French-speaking Caribbean (see Table 1), in addition to several dictionaries of the French Creoles spoken in Saint Lucia, Dominica, Louisiana and Amapá (Brazil). However, the French Creoles of Trinidad, Venezuela, Grenada and San Miguel (Panama) still do not have a dictionary. In addition, there are dictionaries in the French-official Caribbean of proverbs (Confiant 2004; Pinalie & Confiant 1994), a dictionary of Creole neologisms (Confiant 2003), and a two-volume etymological dictionary (Bollée 2017). However, the Atlantic zone is not monolingual in Creole. There are also varieties of French, many of which are strongly influenced by Creole, including Haitian French, Antillean French and Saint Barthelemy ‘Patois’. Yet, there is only one dictionary, of Antillean French, which is described by Aub-Buscher (2003: 4) as ‘not altogether satisfactory’. Haitian French, used by more speakers than several French Creoles, has no dictionary. In addition, there are dictionaries of Amerindian languages: Wayampi (Grenand 1989), Arawak/Lokono (Patte 2012) and Wayana (Camargo & Tom 2010) and of English creoles spoken in French Guiana such as the dictionary of Aluku (Maïs n.d.).

Table 1: Some of the dictionaries of the country/region’s French Creoles.

Region	Dictionaries
Guadeloupe	Bazerque (1969), Ludwig et al. (1990)
Martinique	Pinalie (1992), Confiant (2007)
Haiti	Valdman & Iskrova (2007), Valdman et. al (2017)
French Guiana	Contout (1992), Barthélemy (2007)
Dominica	Fontaine (1991)
St Lucia	Jones & Carrington (1992), Crosbie et al. (2001)
Louisiana	Valdman et al. (1998)
Amapá	Tobler (1987)

The dictionaries that exist are sometimes of excellent quality, sometimes of mediocre quality. For example, there is the huge *Haitian-English Creole Bilingual Dictionary* of Valdman and Iskrova (2007), which has been lauded by many Creole-speaking readers and has a much wider nomenclature than most Creole dictionaries, and its new complement English-Haitian Creole, reverse engineered from the former (Valdman et al. 2017). The difference between homonymy and polysemy, criticized in the dictionaries of the Lesser Antilles by Hazaël-Massieux (2002), is given sufficient attention here. On the other hand, in French Guiana, Barthélemy (2007) is a totally inadequate dictionary for this Creole. It does not distinguish between homonymy and polysemy, and there are many significant omissions of vocabulary, including ‘uncle’ and ‘sister’, and many errors of meaning and nomenclature (see Siegel 2009).

### 3 The Lexicographic Needs of the French-official Caribbean

There is obviously quite a bit of work to be done in the lexicographic field in the Caribbean. Here, I present six paths to pursue:

- 1) The documentation of Saint-Barthélemy French
- 2) Multilingual lexicography between the languages of the French-official Caribbean and the languages of the rest of the region
- 3) Lexicography of the languages of Guyana
- 4) English-speaking French-speaking varieties, including Gustavia English and St. Martin's English
- 5) The dialects of regional standard French
- 6) Signed language

#### 3.1 St Barthélemy French

The local variety of French spoken in St Barth is in urgent need of documentation, as it is probably moribund. The acculturation of the metropolis is replacing the 'patois' of the island with standard French (Maher 2013). The dialect stands out because it is the only one in the region that is not strongly influenced by the omnipresence of a Creole. Indeed, Maher indicates that those who speak what they themselves call Patois on the island do not belong to the same community as those who speak Creole, which came to the island from Guadeloupe with slaves in the 18<sup>th</sup> century. The Patois will most closely resemble colonial French, spoken by white settlers when French Creoles developed in the region (Valdman 1969-70: 78, cited in Maher 2013: 123). It is therefore of extreme importance from the point of view of Creolists to document as much as possible the 'patois' of this island, which may well evince some of the conservative traits of French-based Creole's lexicons, such as Haitian Creole's continued use of *pistach* to mean 'peanut', while French now uses that same form to mean 'pistachio' (employing *cacahuète* or *arachide* to denote 'peanut').

Until now, there is no lexicographic document freely available on this variety. There is a glossary by Gilles Lefebvre of the University of Montreal dating from the 1970s, but it is to be consulted only *in situ* at the Archives Nationales d'Outre-Mer in Aix-en-Provence. Thus the community does not even have access to the only lexicographic document of their own variety of French. The same goes for linguists, both Creole and dialectologists. Research on this variety also comes up against further pitfalls: the physical destruction of the island after the hurricanes of 2017 and the lack of trust felt by the people of St Barth for linguists. The hurricane displaced many of the island's residents, and consequently their knowledge of the dialect. The lack of trust is due to the arrival of a journalist who visited the island in the 1950s, and who stunned the friends he had made by denigrating the people as backward (Maher 2013). So it is quite difficult to do the necessary work before the disappearance of the dialect. If we successfully manage to convince the people of St. Barth who still remain to participate in such research, it will take a differential lexicon, a lexicon that shows the lemmas that are absent from the standard or are different there. It will also require a multilingual glossary, given the presence of a French-based Creole, a local variety of English and Standard French on the island.

#### 3.2 Transregional Multilingual Lexicography

The next task for the lexicographers of the French-speaking Caribbean is the multilingual lexicography of the region. Admittedly, there have been bilingual dictionaries for the region since the beginning, but there are quite few dictionaries that aim to connect the parts of the Caribbean that do not share the same official language. That is, there are many bilingual dictionaries for the languages of the same community – English-Creole in St. Lucia or Dominica, French-Creole in the French West Indies and

French-Wayana in French Guiana. On the other hand, Haiti has the only Caribbean Creole for which there is a lexicographic tradition between Creole and a language that is not official in its territory (English and Spanish), reflecting its status both as the largest Creole language of the region as well as its complex history with the nearby countries of the United States and Dominican Republic. Recently, Cocote (2017) has produced a thesis that translates regional Antillean French to Cuban Spanish. On the other hand, for Creoles, there is no dictionary that allows users to compare the meanings in the French Creoles of Dominica and Martinique, for example, since the official language of the former is English and that of the latter is French. Such a pan-Creole dictionary would be very useful for linguists (as the *Atlas linguistique des Petites Antilles* (Le Dû & Brun-Trigaud 2013) demonstrates). In addition, the quality of Creole dictionaries varies greatly (Hazaël-Massieux 2002), with excellent dictionaries for Haitian Creole, but poor ones for French Guianese Creole. A methodological rigor and a pan-Creole corpus could only improve the bilingual dictionaries produced in the region.

In addition, there is only one dictionary that attempts to span most of the Caribbean region: the *Caribbean Multilingual Dictionary of Flora, Fauna & Foods* (J. Allsopp 2003), an expansion of Jeannette Allsopp's contribution to the *Dictionary of Caribbean English Usage* (R. Allsopp 1996). It is the only dictionary that translates lemmas from Caribbean English into French, French Creole and Spanish. The dictionary is the result of decades of research, and is only the first volume of several regional dictionaries aimed at promoting knowledge of the cultural and linguistic foundation that is shared by the entire region, despite the official languages. There is a second volume in preparation expanding the scope to include religion, music and dance, and more efforts will be needed to keep the lexicography relevant to the entire region. For example, a subsequent edition could be produced that allows easy look-up in any language, not just in Caribbean English as the current dictionary requires.

### 3.3 Dictionaries of the Regional Languages of French Guiana

Bi- and multilingual lexicography in the French-official Caribbean cannot be limited to traversing the official-language boundaries of the Caribbean, but must also take place within the French Caribbean as well, namely in French Guiana. French Guiana is the department and region of France with the highest number of officially recognized regional and minority languages. Foremost among these is the French-based creole endemic to the territory, which enjoys a privileged position among the regional languages, serving as a lingua franca among the many different ethnic groups of the territory. There are also Amerindian languages such as Wayampi, Emerillon, Lokono, and Wayana, as well as English-based creoles such as Paramaka, Aluku and Ndyuka, shared with Suriname. Finally, the Hmong language of Southeast Asia has a well-established presence, its speakers having been resettled there as a form of asylum after fighting alongside the French in their loss of the War in Indochina.

Despite this rich multilingualism and a high value placed on diversity in the region, too few dictionaries of the area have been produced. There has been some recent effort on this front led by Bettina Migge and Isabelle Léglise, linguists at the Center for the Study of Indigenous Languages of the Americas in Paris (CELIA), with dictionaries of the English- and French-based creoles and Amerindian languages under production. These dictionaries will be bilingual, translated into either French or Dutch. Still, only five dictionaries are being produced, leaving half the languages of French Guiana without a dictionary. In light of the acculturation project that France continues to enforce within its territories, the documentation of these varieties is as urgent as ever.

### 3.4 English Dialects of the French-official Caribbean

Fourth, there is a largely pristine area of lexicographic work to be done on the English varieties of the French Caribbean. Because English is spoken by far fewer people in the Caribbean than Spanish,



French or French creoles (Allsopp 2003), it may be surprising that it has long-established communities of native speakers in all the official-language regions of the Caribbean, including groups in the Samaná peninsula of the Dominican Republic, the SSS islands (Saba, Saint Eustatius and Sint Maarten) and of course the French-official Caribbean. The English dialects are spoken on two islands of the French Caribbean, Saint Martin and Saint Barth. Saint Martin is mainly an Anglophone island, with people learning as second languages the European standard varieties of French or Dutch, depending on the side of the island on which they grow up. Still, this dialect remains poorly studied from a lexical perspective, and projects like the *Dictionary of Caribbean English Usage* exclude it from their scope, since it is not found in the English-official Caribbean. The dialect of English found in St Barth is spoken principally in Gustavia, and dates back to the colonial era when St Barth was owned by Sweden (Maher 2013). Sweden never seriously got into the establishment of Caribbean colonies, and used English as its language of trade in the region. The dialect persisted even after St Barth returned to French control. While some preliminary research (Decker 2004) has been conducted on this variety, showing some distinctive lexical elements such as the use of *day* as a locative copula, Gustavia English remains under-documented, and like the local French dialect, it remains under threat of extinction by physical displacement and French acculturation.

### 3.5 Regional Standard French Dialects

While regional languages of the French-official Caribbean are in need of lexical documentation, the regional French dialects mutually intelligible with European Standard French are similarly in need of dictionaries. There is currently a small differential dictionary of Antillean Regional French (Telchid 2007), but no similar dictionary for the standard French of Haiti, which has been shown by Étienne (2005) to be lexically distinctive from the French of Europe with words such as *maisonette* ‘any small house’, *souventes fois* ‘oftentimes’, *déchouage* ‘uprooting’ and *Primature* ‘Office of the Prime Minister’. Given the volume of French produced in Haiti on a daily basis in the press and in government communication, as well as French-language literary works, there is ample opportunity to quickly assemble a corpus on which to base a dictionary of the French of Haiti, as well as a larger dictionary of Antillean French. Similarly, the French of French Guiana is likely in need of documentation: while authors maintain that the French dialect spoken there is essentially just European Standard French, my own fieldwork in the region of only a few months has demonstrated the presence of some regionalisms such as *dégrad* ‘jetty’ (Standard French *débarcadère*), *bacove* ‘banana’ (Standard French *banane*), *boulin bouline* ‘duck duck goose’ (Standard French *chandelle, facteur*), and *maypouri* ‘tapir’ (Standard French *tapir*). There are likely many more regionalisms to be found under the influence of the local languages.

### 3.6 French-official Caribbean Signed Language Varieties

Lastly, there is exploratory work to be done on the signed language varieties of the French Caribbean. French Sign Language is taught everywhere in the French-official Caribbean, except for Haiti, which teaches American Sign Language. However, there is as yet no research into the lexical particularities of French Sign Language in the overseas departments and regions. Just as we saw a number of regionalisms in the overseas departments of French, given the differences between European and Caribbean realities, we must expect that a number of regionalisms would exist in the Caribbean varieties of French Sign Language. Similarly, we would expect that same difference to apply to the variety of American Sign Language taught in Haiti. Fieldwork is therefore needed among experts in these related signed languages.

Beyond regionalisms in the colonizer languages, there is an open question about the lexicons of any community signed languages. Some collaborative research carried out by the deaf university



Gallaudet University in the United States, the Organization of American States, and the Office of the Secretary of State for the Integration of Handicapped People in Haiti, has already started documenting the properties of Haitian Sign Language, an indigenous variety mutually unintelligible with the local American Sign Language variety (Bureau 2014). There is therefore ample opportunity to describe a large lexicon for a highly vulnerable population. Furthermore, because areas with small gene pools tend to develop deafness over time, it is suspected that St Barth, with its small white population, might have a community signed language to explore as well (Benjamin Braithwaite, p.c., August 5, 2016), and in principle the same arguments might apply to small communities in the rainforests of French Guiana.

Advancements in signed language lexicography are coming out of the English-official Caribbean, which will facilitate look-up strategies in signed languages to get the spoken language equivalent. Normally, signed language dictionaries are organized in a way that allows hearing people to find the signed language equivalent. However, at the University of the West Indies, St. Augustine Campus in Trinidad, there is research using motion-sensor technology that is being developed that will allow users of a signed language to sign a word in their native language. This technology will allow a much broader range of lexicographic projects to be pursued, including bilingual, transregional lexicography.

## 4 Conclusion

There is a wide variety of lexicographic projects that have yet to be attempted and completed, which does not even take into account improvements in the quality of projects that have already been carried out. The Richard & Jeannette Allsopp Centre for Caribbean Lexicography, of which I am the director, is prepared to assist with the execution of any of these projects and to lead a number of them. The Allsopp Centre is the only unit dedicated to the promotion and practice of lexicography that spans the entire Caribbean region. It is the successor to units that produced works such as the aforementioned *Dictionary of Caribbean English Usage* (R. Allsopp 1996) and *Caribbean Multilingual Dictionary* (J. Allsopp 2003). It also is actively producing regional dictionaries such as a bilingual culinary dictionary of Caribbean English with Costa Rican Spanish, a multilingual dictionary of French Guianese Creole, and a multilingual dictionary of medicinal plants of the region. We are well-placed to assist with any projects in the French-official Caribbean, from design to research to execution of the final product. From the urgent projects of documenting the endangered varieties of St Barth and French Guiana, to the longer term projects of multilingual dictionaries and dictionaries of regionalisms in Standard French, the Allsopp Centre is eager to fully document the lexicons of the French-official Caribbean.

## References

- Alleyne, M. C. (1985). *A Linguistic Perspective on the Caribbean*. Washington, D.C.: Woodrow Wilson International Center, Latin American Program.
- Allsopp, J. (2003). *The Caribbean Multilingual Dictionary of Flora, Fauna and Foods in English, French, French Creole and Spanish*. Kingston: Arawak.
- Allsopp, R. (1996). *Dictionary of Caribbean English Usage*. Oxford: Oxford University Press.
- Aub-Buscher, G. (2003). Linguistic Paradoxes: French and Creole in the West Indian DOM at the Turn of the Century. *The Francophone Caribbean Today: literature, language, culture*. Gertud Aub-Buscher and Beverly Ormerod Noakes (eds). Mona: UWI Press. pp. 1-15
- Barbotin, M. (1995). *Dictionnaire du créole de Marie-Galante*. Hamburg: Helmut Buske Verlag.
- Barthelemy, G. (2007). *Dictionnaire créole guyanais-français : suivi d'un index français-créole guyanais*. Matoury, Guyane: Ibis Rouge Editions.

- Bazerque, A. (1969). *Le langage créole*. Guadeloupe : ARTRA.
- Bollée, A. (2000). *Dictionnaire étymologique des créoles français de l'océan Indien*. Hamburg: Helmut Buske Verlag.
- Breton, R. (1665). *Dictionnaire caraïbe-françois: Meslé de quantité de remarques historiques pour l'esclaircissement de la langue*. A Auxerre: Par Gilles Bouquet.
- Bureau du Secrétaire d'Etat à l'Intégration des Personnes Handicapées (2014). Vers une meilleure compréhension de la langue des signes haïtienne. <http://www.seiph.gouv.ht/vers-une-meilleure-comprehension-de-la-langue-des-signes-haitienne-2/>. Accessed March 30, 2018.
- Camargo, E. & Tom, I. *Hakëne omijau eitop Dictionnaire bilingue Wajana-Palasisi Wayana-Français*. Cayenne: CELIA/DRAC-Guyane/TEKUREMAI.
- Cocote, E. (2017). *Création d'un lexique bilingue français régional des Antilles-espagnol cubain, et enjeux traductifs et interculturels*. PhD. thesis. Université des Antilles, Pointe-à-Pitre.
- Confiant, R. (2001). *Dictionnaire des néologismes créoles*. Matoury, French Guiana: Ibis Rouge Editions.
- Confiant, R. (2004). *Le grand livre des proverbes créoles: Ti-pawol*. Paris: Presses du Châtelet.
- Confiant, R. (2007). *Dictionnaire créole martiniquais-français*. (2 vol.) Matoury, French Guiana: Ibis Rouge Editions.
- Contout, A. (1992). *Le petit dictionnaire de la Guyane : Classé par thèmes, avec histoires de mots, tournures et conversations*. Cayenne: Self-published.
- Crosbie, P., Frank, D., Leon, E. & Samuel P. (2001). *Kweyòl Dictionary*. Castries: Ministry of Education, Government of Saint Lucia.
- Decker, K. (2004). Moribund English: The case of Gustavia English, St. Barthélemy. *English World-Wide*, 25, pp. 217-254.
- Étienne, C. (2005). "Lexical particularities of French in the Haitian press: Readers' perceptions and appropriation." *Journal of French Language Studies*, 15 (3), pp. 257-277.
- Fontaine, M. (1991). *Dominica's diksyonnè : Kwéyòl-Annglè = Dominica's English-Creole dictionary*. Roseau, Dominica: The Folk Research Institute, the Konmte Pou Etid Kweyol (KEK).
- Grenand, F. (1989). *Dictionnaire wayâpi-français, lexique français-wayâpi: Guyane française*. Paris: Peeters-Selaf.
- Hazaël-Massieux, M-C. (2002) A propos des dictionnaires créoles des Petites Antilles. <http://creoles.free.fr/Cours/diaporamas/dictionnairescreoles.pps>. Accessed March 30, 2018.
- Jourdain, É. (1956). *Le vocabulaire du parler créole de la Martinique*. Paris: Klincksieck.
- La Salle de L'Etang, Simon Philibert de (1763). *Dictionnaire galibi: Présenté sous deux forms; I° commençant par le mot français; II° par le mot galibi. Précédé d'un essai de grammaire*. Paris: Chez Bauche.
- Le Dù, J. & Brun-Trigaud, G. (2013). *Atlas linguistique des Petites Antilles*. Paris: CTHS.
- Ludwig, R., Montbrand, D., Poulet, H., & Telchid, S. (1990). *Dictionnaire Créole français : (Guadeloupe) : avec un abrégé de grammaire créole, un lexique français/créole, les comparaisons courantes, les locutions et plus de 1000 proverbes*. [Paris?]: Servedit/Editions Jasor.
- Maher, J. (2013). *The Survival of People and Languages: Schooners, Goats and Cassava in St. Barthélemy, French West Indies*. Leiden: Brill.
- Maïs, J-L. (2002). *Dictionnaire aluku tongo - français, français - aluku tongo*. Toulouse: Sedrap.
- Mondesir, J. E., & Carrington, L. D. (1992). *Dictionary of St. Lucian Creole*. Berlin; New York: Mouton de Gruyter.
- Patte, M. F. (2012). *La langue arawak de Guyane: Présentation historique et dictionnaires arawak-français et français arawak*. Marseille: IRD Editions.
- Pinalie, P. (1992). *Dictionnaire élémentaire français-créole*. Paris : L'Harmattan / Presse Universitaires Créoles.
- Pinalie, P., & Confiant, R. (1994). *Dictionnaire de proverbes créoles*. Fort-de-France: Ed. Désormeaux.
- Pompilus, P. (1958). *Lexique créole-français, thèse complémentaire*. PhD. thesis. Université de Paris, Paris.
- Pury, S. de. (1999). "Le Père Breton par lui-même." in M. Besada Paisa (ed.). *Dictionnaire caraïbe-français*. Paris: Karthala/IRD. pp. XV-XLV
- Rickford, J. R. (1987). *Dimensions of a Creole Continuum*. Palo Alto: Stanford University Press.
- Siegel, J. F. (2009). Barthelemi, Georges. 2007. *Dictionnaire créole guyanais-français : suivi d'un index français-créole guyanais* and Confiant, Raphael. 2007. *Dictionnaire créole martiniquais-français*. (2 vol.), *The French Review* 83 (2). pp. 463-65.
- Skybina, V. & Bytko, N. (2014). Caribbean creole lexicography as a cultural phenomenon. Paper presented at 20th Biennial Conference of the Society for Caribbean Linguistics (SCL) in conjunction with the Society for Pidgin and Creole Linguistics (SPCL) and the Associação de Crioulos de Base Lexical Portuguesa e Espanhola (AC-BLPE), Aruba.

- Telchid, S. (1997). *Dictionnaire du français régional des Antilles: Guadeloupe, Martinique*. Paris: Bonneton.
- Tobler, Alfred W. (1987) *Dicionário Crioulo Karipúna-Português/Português-Crioulo Karipúna*. Brasília: Summer Institute of Linguistics.
- Valdman, A., & Iskrova, I. (2007). *Haitian Creole-English Bilingual Dictionary*. Bloomington, IN: Indiana University, Creole Institute.
- Valdman, A., Klingler, T. A., Marshall, M. M., & Rottet, K. J. (1998). *Dictionary of Louisiana Creole*. Bloomington, IN: Indiana University Press.
- Valdman, A., Moody, M. D., & Davies, T. E. (2017). *English-Haitian Creole Bilingual Dictionary*. Bloomington, IN: Indiana University Creole Institute.



## Various Topics





# The Dictionary of the Learned Level of Modern Greek

*Anna Anastassiadis-Symeonidis<sup>1</sup>, Asimakis Fliatouras<sup>2</sup>, Georgia Nikolaou<sup>1</sup>*

<sup>1</sup>*Aristotle University of Thessaloniki, <sup>2</sup>Democritus University of Thrace*

*E-mail: ansym@lit.auth.gr, afliatou@helit.duth.gr, ngeorgia@smg.auth.gr*

## Abstract

The aim of this paper is to discuss the theoretical background and methodological tools for the elaboration of a specialized dictionary, the Dictionary of the Learned Elements of Modern Greek (DILLEMOG). The learned level of Modern Greek (MG), which originates from the natural diachronic inheritance and from the prototyping of Ancient Greek, includes segments, structures and processes which pertain to all levels of linguistic analysis. DILLEMOG will constitute an innovative lexicographical database which will provide the user with all the necessary information on the [+ learned] linguistic items of MG, such as definitions, collocations, degree of learnedness, lexical and morphological classification, functionality and usage.

**Keywords:** learned level, DILLEMOG, lexicographical project

## 1 Introduction

It is a fact that grammars, dictionaries and linguistic research concerning Modern Greek are restricted mainly to the description of the linguistic norms, that is of the neutral register zone, overlooking the larger part of its learned level. Moreover, Greek and/or foreign pupils/students and scholars/academics learning Modern Greek are not aware of the degree in which Ancient Greek (henceforth AG) has survived in the learned register of Modern Greek (henceforth MG), as well as in their own languages through internationalisms (such as scientific terms of Greek or Latin origin). As a result, they face difficulties in the formation of linguistic types. Therefore, in this paper we suggest the elaboration of a specific innovative scientific tool the *Dictionary of the Learned Level of Modern Greek*, in the form of an advanced digital product in a free-access repository. It comprises three parts (lexicographic protocol, electronic dictionary, and utility guide), and is designed to obtain a scientific patent. Finally, this product aims at projecting the learned level of Modern Greek in a linguistic (theoretical, historical and applied) as well as an educational framework<sup>1</sup>.

## 2 The learned level of contemporary Modern Greek: State of the art

The learned level includes the inherited segments, structures and processes from former periods of the Greek language on all levels of linguistic analysis (phonology, morphology, semantics, syntax and pragmatics) as well as lexicon, that are used mainly on the high/formal register. Therefore, learnedness is defined by etymology, mainly in terms of inheritance and grammatical/lexical deviation or peripherality<sup>2</sup>, and essentially by register, in terms of representation of the high/formal level (see Anastassiadis-Symeonidis & Fliatouras 2004; 2018), e.g.

1 The authors have already submitted two applications for sponsorship approval to two project sponsorship committees run by the Greek state (June 2017 and January 2018). The initial version of DILLEMOG will be soon available on the website of the linguistic laboratory *SYNMORPHOSE* (<http://synmorphose.compulaw.gr/index.php?lang=el>).

2 Etymology is a necessary but not sufficient condition. For example, the word *Θεός* 'God' is inherited from AG but it is not learned in MG.

(1) *ikia* ‘house’, *patir* ‘father’<sup>3</sup>.

The learned level is derived from natural diachronic inheritance, mainly through the language of administration, high oral/written registers, the scientific register and the language of church as language variation (see also Karantzola & Fliatouras, forthcoming), as well as the standardization of AG. The latter led to the re-introduction of learned elements, mainly in terminology internationalisms<sup>4</sup> (Anastassiadis-Symeonidis 1994), e.g.

(2) MG *osteoarθritida* < EN *osteoarthritis* < AG *ostoún* ‘bone’ + *arthron* ‘connection’ + *-itis* ‘suffix for diseases’

and the artificial revival of elements as an outcome of the “Language Question”<sup>5</sup>, mainly as Katharevousa fossils (see Papanastasiou 2010), e.g.

(3) *Trapeza tis Elaðos* ‘Bank of Greece’ (cf. the [-learned] *Trapeza tis Elað-as*)<sup>6</sup>.

A first attempt at a systematic cross-level classification in phonology, morphology, semantics, syntax and lexicon was made by Anastassiadis-Symeonidis and Fliatouras (2004; 2018) and Anastassiadis-Symeonidis (2015)<sup>7</sup> as shown in Table 1.

Table 1: Classification of the learned categories of MG

	Norm	Learned
<b>(A) Phonology</b>		
(a) Consonant Clusters	<b>ftoxós</b> ‘poor’	<b>ptoxós</b> (cf. <i>ptoxokomio</i> ‘poorhouse’)
(b) Stress	<b>asfálias</b> ‘safety’ (GEN)	<b>asfálias</b> (cf. <i>zoni asfálias</i> ‘safe belt’)
(c) Final -n	<b>usía</b> ‘substance’	<b>usían</b> (cf. <i>kat’ usían</i> ‘essentially’)
(d) Foreshortening	<b>klironomía</b> <sup>8</sup> ‘heritage’	<b>klironomía</b> (cf. <i>foros klironomías</i> ‘inheritance tax’)
<b>(B) Morphology</b>		
(a) Stem/Affixal allomorphy	<b>pali-os</b> ‘old’	<b>pale-os</b> (cf. <i>Palea Diaθiki</i> ‘Old Testament’)
(b) Word Construction	<b>miso-fegaro</b> ‘half-moon’	<b>imi-selinos</b> ‘half-moon’
(c) Fossilized elements	<b>(s)ta</b> ekato ‘percentage’ (ACC)	<b>tis</b> ekato ‘percentage’ (DAT <sup>9</sup> )
<b>(C) Syntax</b>		
(a) Prepositions	<b>stin</b> poli (<se tin poli) ‘in town’	<b>en ti</b> poli (in Christmas carols)
(b) Word order	<b>i kini apófasi</b> (Adj-Noun) ‘mutual decision’	<b>I apó kinú apófasi</b> (Adv Det.-Noun) ‘mutual decision’
(c) Infinitive	<b>to na kapnizeis</b> ‘to smoke’ (SUBJ)	<b>to kapniz-in</b> ‘to smoke’ (INF)
<b>(D) Lexicon</b>		
(a) Words	<b>kokalo</b> ‘bone’	<b>osto</b> ‘bone’
(b) Phrases	<b>apo tin pediki ilikia</b> ‘since childhood’	<b>eks apalon onixon</b> ‘since childhood’

3 The peripheral word *patir* is fully inherited from AG, whereas the morphologically changed form *pateras* (< AG *patir*) is not learned in MG.

4 Internationalisms are popular words which occur at the same time in many different languages through the process of borrowing.

5 The “Language Question” was a linguistic, political and social issue regarding the linguistic variant that should become the official language of the Greek state. There were two parties; the first one supported the Greek vernacular as it was spoken by the majority of people in continental Greece (‘dimotiki’), and the second one claimed that the official language should be a cultivated variant that did not include any “vulgar” expressions and adopted AG syntax and morphology as a sign of grandeur (‘katharevousa’). The issue was finally resolved in 1976, when dimotiki acquired the official language status under law.

6 The ending *-as* indicates that the noun is inflected according to the patterns of the vernacular, while the ending *-os* imitates AG inflection.

7 See also Browning (2008), Kamilaki (2009), Krimpas (2015).

8 This is the stress pattern in nouns where [i] is followed by a vowel.

9 In AG dative was a case which was used to indicate the manner or the tool with which something happened. It also served as an indirect object to the main verb.

As shown by Anastassiadis-Symeonidis and Fliatouras (2004; 2018), the degree of learnedness is defined by a non-static continuum, which involves a hyponymic representation of the formal/high register. As illustrated in Figure 1, it comprises two flexible zones, the learned and non-learned ones, which intersect in the overlapping distribution of the intermediate norm (unmarked zone). Every element is prototypically integrated into the learned or non-learned zone, but the degree of learnedness can differ from item to item as well as among speakers on the basis of sociolinguistic parameters, such as proficiency in Ancient Greek, age and so on (see also Kambakis-Vougiouklis & Fliatouras, forthcoming).

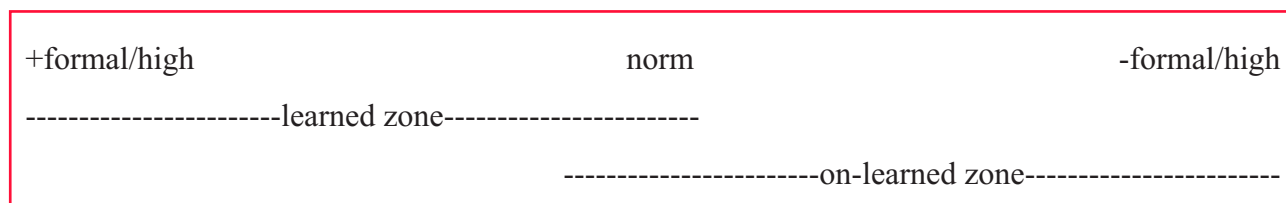


Figure 1: Continuum of learnedness

For example, the suffix *pan-/pant-* ‘pan-’ falls prototypically into the learned category, but it can be found in norm or even in non-learned words, e.g.

- (4) *Pant-anassa* ‘queen of all’ as an epithet of Mary (learned)
- (5) *pan-ellinies* ‘panhellenic’ (norm)
- (6) *pan(t)-ermos* ‘very lonely’ (non-learned).

The learned level nowadays occurs mainly in the administration/legal vocabulary, in inscriptions/names/signs, in place/street names and surnames, in terminology, in academic vocabulary, in official military terminology, in church language, in commercial language (mainly in titles, brand names), and in announcements, e.g.,

- (7) *o katothi ipoyeyramenos* ‘the undersigned’ (in applications)
- (8) *par’ Ario Payo* ‘in the Supreme Court’ (title of lawyers)
- (9) *othisate/elksate* ‘push/pull’ (in door signs)
- (10) *Δukisis Plakentias* ‘duchess of Plakentia’ (toponym)
- (11) *aiθales* ‘evergreen’ (terminology)
- (12) *Xristos Anesti* ‘Christ is risen’ (ecclesiastic phrase used at Easter)
- (13) *alt, tis i?* ‘halt, who are you?’ (military phrase)
- (14) *anθos aravositu* ‘corn flour’ (brand name)
- (15) *iserxete ston staθmo tu Plateos* ‘the train arrives at the station of Platy’ (announcement in trains).

Moreover, the learned level is not fossilized, but is still functional, as it is productive in primary sources (terminology, etc.), e.g.,

- (16) *siriaki θira* ‘serial port’

in playful neologisms (see also Kamilaki 2012; Fliatouras & Koukos forthcoming), e.g.,

- (17) *yiaurtoskorδion* instead of *tzatziki* ‘tzatziki’<sup>10</sup>

in morphology, in the form of allomorphic analogy and/or orthographic rules (see Papanastasiou 2008; Fliatouras, in press), e.g.,

<sup>10</sup> It was used in the extremely popular comic series *Konstantinou and Elenis*, the first episode of which aired in October 1998. Although the series ended in June 2000, it was re-broadcast several times due to the extraordinary audience reception.

(18) *anθ-elinas* ‘anti-Greek’<sup>11</sup> (cf. the +/-learned *anti-*)

and following diaphasic markers of reformulation (see Anastassiadis-Symeonidis, forthcoming), e.g.,

(19) *γία να το πο αρχεοπροπος* ‘to say it like in Ancient Greek’.

In other words, after the establishment of Demotic Greek, the learned register has segued from the frame of natural creativity into the sphere of analogicality (see Anastassiadis-Symeonidis & Fliatouras forthcoming; Martzoukou et al., forthcoming<sup>12</sup>) and occasionally into a form of neo-Katharevousa<sup>13</sup> (see Petrounias 1984).

Usually no learned variety is an absolute equivalent to the non-learned, on the language system and/or register level, e.g. *lefkos* ‘white’ (learned) is not deployed in all the uses of *aspros* ‘white’ (norm). The competitiveness between the learned/non-learned morphological and lexical segments and the consequent variety is extensive, and needs further study on the basis of text corpora (Fliatouras, forthcoming), e.g.,

(20) *ikos* (learned) vs *spiti* (norm) ‘house’ (cf. *Lefkos Ikos* ‘White House’ vs *aspro spiti* ‘white house’)

(21) *θira* (learned) vs *porta* (norm) ‘door’ (cf. *θira 7* ‘gate 7 of a stadium’ vs *porta 7* ‘a door with number 7’).

Apart from total and fossilized remnants (see Coffey 2013) of AG, perceptible to a greater or lesser degree as peripheral types, formally and/or semantically marked (Anastassiadis-Symeonidis & Fliatouras 2004; 2018), e.g.,

(22) *o ayon* (learned) vs *o ayonas* (norm) ‘match’

(23) *estali* (learned) vs *stalθike* (norm) ‘was sent’

(24) *uδis* (learned) vs *kanis* (norm) ‘nobody’.

There are many learned elements that do not have competitive selection, and this is what seems to keep the learned element strong in the Greek language, as seen, for instance, with the following:

(25) lexical phrases: *Nekra Thalasa* vs *\*Nekri Thalasa* ‘Dead Sea’

(26) allomorphs: *siδiro-δromos* vs *\*siδero-δromos* ‘railway’

(27) substantivized words by conversion or ellipsis: *nekra (taxitita)* vs *?nekri (taxitita)* ‘neutral gear’

(28) verb types: *parenevi* vs *?parenevice* ‘intervened’

(29) affixes (inflectional and derivational): *stroma thalas-isINF* vs<sup>14</sup> *\*stroma thalas-asINF* ‘sea mattress’

(30) fixed phrases/collocations: *eyine tis kolaseos* vs<sup>15</sup> *\*eyine tis kolasis* ‘there were fireworks’ (about fierce disagreement).

Finally, there may be observed a tendency of change in learned elements, either as a natural procedure or as a result of language and at times political ideology. For instance, many learned elements tend to be expanded analogically in the Demotic, e.g.,

(31) *Kathara Aeftera* à *Kathari Aeftera* ‘Ash Monday’

11 The change of the stop consonant [t] into a fricative [θ] is attributed to the presence of the rough breathing (*spiritus asper* in Latin), indicated by a special symbol above the initial vowel.

12 The authors conducted a psycholinguistic experiment using pupils of the last grade of primary school (age 11 or 12) as subjects. According to the findings, some of the subjects created analogical pseudo-neologisms, i.e. new lexical units which do not exist but follow a certain existing pattern, e.g. *\*chapi thalasis* ‘sea pill’ which was formed by analogy with the existing *stroma thalasis* ‘sea mattress’.

13 For many language specialists in Greece (see Petrounias 1984), neo-Katharevousa constitutes a “new disease” of the Greek language, where the effort of the speakers to sound more cultivated leads them to extremely sophisticated and sometimes erroneous imitations of AG vocabulary and syntax.

14 The ending *-as* indicates that the noun is inflected according to the patterns of the vernacular, while the ending *-is* imitates AG inflection.

15 See above.



and, inversely, many non-learned elements are often led to hypercorrection (see Krimpas, forthcoming), e.g.,

- (32) *apopiúme ton eφθinón mu* ‘I abdicate my responsibilities’ (Object in Genitive = AG syntax) instead of *apopiúme tis eφθínes mu* (Object in Accusative = MG syntax).

### 3 DILLEMOG: Necessity, innovation and impact

The distinction between learned and non-learned/colloquial/vernacular registers is prevalent in a variety of academic works (see, among others, Triandaphyllidis 1963; Setatos 1969, 1992; Tobaidis 1978; Babiniotis 1982; Petrounias 1984; Charalambakis 1999; Ralli 2005; Panaretou 2006; Xydopoulos 2007; Kamilaki 2009; Anastassiadis-Symeonidis 2015; Krimpas 2015). However, as documented in the literature, the use of this criterion does not correspond to a systematic theoretical framework; rather, it is usually empirical, descriptive or circumstantial, and is mainly related to etymological criteria. Now that approximately forty years since the establishment of Demotic have passed, it is necessary to redefine the learned register as a product of the natural learned evolution, as well as of Katharevousa, that in some cases can be understood as variation and in other cases as a compulsion version without competitive variety.

Furthermore, research in lexicography is usually limited to marking certain elements as learned (cf. *Dictionary of Standard Modern Greek* by the Triandaphyllidis Foundation, *Dictionary of Modern Greek Language* by Babiniotis, and *Utilitarian Dictionary of Modern Greek Language* by the Academy of Athens, edited by Ch. Charalambakis) or the collection of archaic lexical units (cf. the *Dictionary of the most Advanced Words* by Babiniotis, the *Dictionary of Scholarly Expressions in Contemporary Greek* by Iordanidou and the *Dictionary of Difficulties and Common Errors in the Use of Modern Greek* by Babiniotis), while the scientific terminology as part of the learned register is also studied fragmentarily (see the ELETO website<sup>16</sup>). Regarding metalinguistic tools, only hints at certain features of the learned level are recorded in the grammars<sup>17</sup> by Holton, Mackridge and Philippaki-Warbuton (1999), Kleris and Babiniotis (2005) and Chatzisavvidis and Chatzisavvidou (2013).

Therefore, the only way to safely demark, describe, analyze and implement (e.g. in language teaching) the learned level passes through the organized and systematized collection and classification of the material. The DILLEMOG will allow further investigation of the following issues of the thus far theoretical and historical research in relation to the necessity and deployment of the learned in MG, in its relation to the AG, paving the way to its utilization on an academic level and its implementation in educational frameworks (see Fliatouras, forthcoming).

Specifically, the goal of the DILLEMOG is the exhaustive collection, linguistic analysis and implementation proposals of the learned level of MG (including scientific terminology) as part of holistic research based on a representative corpus of texts, such as written texts in scientific journals, conference proceedings, university teaching materials, public administration and legal documents, texts written in the religious language, as well as in data that derive from important electronic sources of textual bodies, such as the CGT<sup>18</sup> (Corpus of Greek Texts), the HNC<sup>19</sup> (Hellenic National Corpus), and the Portal for the Greek Language<sup>20</sup>. The objectives of the DILLEMOG include: (a) On a

16 <http://www.eleto.gr/gr/reception.htm>.

17 E.g. in the cases of certain expressions and collocations where [+ learned] prepositions are used (ἐπί vs πάνω ‘on top of, above’), the authors indicate that the prepositions originate from AG and have acquired a special use in MG (see Kleris & Babiniotis 2005: 944).

18 <http://sek.edu.gr/>.

19 <http://hnc.ilsp.gr/>.

20 <http://greeklanguage.gr/>.

scientific and academic level, to contribute towards the more precise description of MG and the more effective teaching of MG, AG and the special languages<sup>21</sup> to Greek and foreign pupils, students and instructors (see Katsogiannou & Xenophontos 2015; Anastassiadis-Symeonidis & Fliatouras, 2018), as well as to point out the history of the Greek language, granting significant linguistic tools for the scientific community in a free-access electronic repository. (b) On a creative innovation level, being an original scientific project, to claim a Greek and international patent, as there does not exist, to the best of our knowledge, an analytical dictionary of a language register – specifically the learned level – which can be found neither in the Greek nor international literature.

As a result, the DILLEMOG is a digital product which will provide free data access to all potential users via the official website of the research program. It is designed to be a public repository of national value which will bring benefits in many different fields:

- **Scientific field:** It is an innovative product that attempts for the first time in Greek scientific research to present the learned elements of the Greek language (items, lexical units, supralexic combinations, collocations, phrases, idioms). It will shed light on all the following parameters of the subject: definition, criteria of learnedness, degree of learnedness, lexical and morphological categorization, interpretation, usage, functionality, competitiveness and consolidation. It offers clear and complete answers to questions like: which are the learned elements of Modern Greek (both written and vernacular), what is their origin, what are their syntagmatic relationships and their most frequent grammatical patterns and what is their actual use, e.g. what are their occurrences in native speakers' utterances? It is based on a multilevel exploitation of both the written resources that already exist (dictionaries, encyclopedias, thesauri), as well as the most important digital tools that are currently used for linguistic description.
- **Academic field:** It prepares the way for pursuing new scientific goals and can become a useful basis for the elaboration of language user guides and the development of curricula in language teaching.
- **Cultural field:** It constitutes a valuable scientific tool for the community of Greek scholars as well as for the international academic community, as it offers useful insights into the history of the Greek language, the survival of Ancient Greek in modern language (see Symeonidis, Xenis & Fliatouras 2007) and the evolution of linguistic culture.
- **Educational field:** It can be equally efficient in Greek language teaching as a first and second/foreign language, thanks to the large amount of information that the user is provided with: information about the spelling, the standard pronunciation of the linguistic items and their place and role in the universe of communication. For the first time diachrony (e.g. historical evolution of language) becomes a strong factor in the synchronic description of the language system (langue) and its realization (parole) (see Anastassiadis-Symeonidis & Fliatouras, forthcoming).
- **Social field:** It can become a compensation tool in the hands of speakers and users who do not enjoy easy access to academic/high profile language use.

## 4 Methodology/Compilation process

The DILLEMOG is a digital lexicographic information system (see Müller-Spitzer 2009) which is designed within the frame of the *function theory* (Müller-Spitzer *ibid*; Tarp 2009; Schierholz 2015). This considers all lexicographic tools (printed as well as electronic dictionaries) as *social culture-specific products* (Tarp 2009: 22) which “target the specific needs of specific users in specific social situations” (Fuertes-Olivera 2009: 99). In the paragraphs below, we shall outline the basic parts of the compilation process.

21 Greek language variants which are differentiated from others on social grounds, e.g. language of the law and legal documents.

## 4.1 Lexicographic Protocol

After having taken into account the basic principles and current standards of international lexicography, we began the compilation of the DILLEMOG with the elaboration of a lexicographic protocol, a detailed scientific description of all the data which will be included in our project. More specifically, the DILLEMOG protocol will determine the selection criteria of the macrostructural units, the nature and length of the microstructural units, the type of units which will have cross-reference function (mediostructural units, according to Müller-Spitzer *ibid*), the order of the information which will appear in a typical full-length article<sup>22</sup>, the typographic indications of the various elements of the articles, the particular symbols which will introduce data of a different kind (e.g. audio files or graphs), the navigational elements (e.g. the list of headwords) and the handling of certain theoretical issues concerning linguistic descriptions and metalanguage. The protocol will also provide information related to the manipulation of the three electronically published corpora of Modern Greek and the RSS feeds which will be used for the detection of the current language use.

## 4.2 The Electronic Dictionary of the Learned Element of Modern Greek

The selection of the macrostructural units will be based on the indexing of the existing reference works; e.g. dictionaries from the previous decades (*The Dictionary of Modern Greek Language of Proia Editions*, *The Dictionary of the Greek Medieval Literature*, *The Dictionary of Dimitrakos*, *The Greek-English Dictionary Georgakas*) and current dictionaries of Standard Greek (three of them being General Dictionaries and the fourth being the *Reverse Dictionary of Modern Greek*), as well as other tools (e.g. encyclopedias) which contain a large amount of metalinguistic information. Their selection criteria will be thoroughly discussed in the lexicographic protocol. The headwords (lemmas) can be lexical units, sublexical units or supralexical combinations, and will be accompanied by their allomorphic variations as reported in current language use (e.g. *nixt/nixth*- 'night'), so that the search process is facilitated. Regarding the metalinguistic data of each headword that will appear in a particular order, the following information will be indicated: the grammatical category of the word (or its morphological status if the lemma is a sublexical unit), its concise definition and a number of *sublemmata* (e.g. collocations) and their respective meanings and most popular field of use (e.g. church language). A link will connect each sublemma to an authentic utterance drawn from the Greek corpora. At the end of the lexicographic article there will be information about the origin of the headword (etymology) and the degree of its [+learnedness]. In cases where pronunciation issues arise, a special symbol will indicate access to an audio file.

Regarding the underlying structure of our product, a tailor-made XML DTD (Document Type Definition) will determine the content-related value of our data. XML tags will define the scope of our elements, and will insert instructions concerning their special presentation characteristics. This method will secure the consistency and homogeneity of the data viewed by potential users. In the final step of the elaboration, an open-source content management system (CMS) will be manipulated so that the data is electronically published and ready to be used by those who are interested. In this final step all the necessary multimedia applications will be included in the product (e.g. the audio files).

The DILLEMOG complies with the standards of modern lexicography. It is developed within a specific theoretical framework with a strict specification of its layout, typographic indications, internal architecture and search options. Although it encompasses different kinds of information, the average dictionary user can easily access its content. As such, it responds to all possible search situations that average users find themselves in (Tarp 2009: 25-26): (i) Attempting to add something to their

22 As far as electronic dictionaries are concerned, there are different views of the lexicographical data depending upon the search of the potential user. A typical full-length article is the kind of view that would appear in a printed dictionary, where the structure is fixed and non-modifiable.

previous knowledge; (ii) attempting to become more efficient in particular communicational circumstances; (iii) looking for instructions and/or advice for completing a task (mental or manual); and (iv) handling a guide for decoding the symbols and signs of the surrounding world. Its underlying data structure allows the modification and/or enrichment of its content; its data is thus not static, and can be updated according to new research findings. As a result, the DILLEMOG is a lexicographic product which adopts certain linguistic trends and options and aims to offer a consistent description of current language use regardless of the attitudes expressed by the speech community or the linguistic stereotypes that linger on.

### 4.3 Utility Guide for the Exploitation of the DILLEMOG

The Guide for the Exploitation of the DILLEMOG will contain thorough teaching scenarios (Iordanidou & Papaioannou 2014; Kosegian, Papathanasiou & Fliatouras 2012) within the scope of Task-Based Learning (Willis 1996). TBL involves language learning through tasks of data searching and analyzing, comparing language use and handling tools for the description of the target language. A part of the guide will be dedicated to CLIL (Content and Language Integrated Learning), which involves the development of academic vocabulary through different subject teaching (e.g. history or science) for all learners who are interested in acquiring the academic variety of the target language (Anastassiadis-Symeonidis et al. 2014). Proposals will also be made as to the exploitation of the learned level in creative writing.

## 5 Conclusions

The aims of this paper are (i) to provide an outline of the theoretical issues that arise from the long-term parallel use of the Greek vernacular and the [+ learned] elements which originate from AG, and (ii) to present the basic elaboration principles of an innovative lexicographical project which will address not only native speakers of Greek, but also non-Greek-speaking users who are interested in the processes of language borrowing and/or language standardization. The Dictionary of the Learned Elements of Modern Greek is designed to include various information on the segments, forms, structures and processes that belong to the learned zone of the continuum of learnedness, such as definitions, collocations, degree of learnedness, lexical and morphological classification, functionality and usage. The DILLEMOG will be in digital form and will constitute a valuable tool for both researchers as well as teachers of Greek. The compilation of DILLEMOG will be based upon a lexicographic protocol according to the current international standards. The project will be completed with the Utility Guide for the Exploitation of the DILLEMOG within the classroom, with the presentation of teaching scenarios and task examples.

## References

- Anastassiadis-Symeonidis, A. (1994). *Neological borrowing in Modern Greek. Direct Loans form French and English/American*. Thessaloniki: Estia editions [in Greek].
- Anastassiadis-Symeonidis, A. (2003). *Reverse Dictionary of Modern Greek*. Thessaloniki: Institute of Modern Greek Studies [Manolis Triandaphyllidis Foundation].
- Anastassiadis-Symeonidis, A. (2015). Grammar of the Modern Greek academic language within the scope of Linguistics. In *Proceedings of 10<sup>th</sup> ELET Conference, 12-14 November*, University of Athens, Greece [in Greek].
- Anastassiadis-Symeonidis, A. (Forthcoming). Diaphasic discourse markers and Lexicography. In *Honorary Volume* under preparation: Athens, Greece.



- Anastassiadis-Symeonidis & Fliatouras, A. (2004). The distinction learned and colloquial in Modern Greek: definition and classification. In *Proceedings of ICGL6, 18-21 September 2003*, Department of Philology of the University of Crete, CD-ROM [in Greek].
- Anastassiadis-Symeonidis, A., Zagka, E. & Mattheoudaki, M. (2014). Combinational approaches in language teaching: the case of *CLIL*. In *Proceedings of ICGL11, 26-29 September 2013*, University of the Aegean, Rhodes, Greece, pp. 102-113.
- Anastassiadis-Symeonidis, A. & Fliatouras, A. (2018). From the learned register of Modern Greek to Ancient Greek: Research proposals and educational perspectives. In *Studies for Greek Language* (April 2017). Thessaloniki: Aristotle University of Thessaloniki [in Greek].
- Anastassiadis-Symeonidis, A. & Fliatouras, A. (Forthcoming). [+/- learned] process. In *Proceedings of ICGL13, 7-9 September 2017*. Department of Greek Language, London-Westminster, UK [in Greek].
- Babiniotis, G. (2005). *Dictionary of Modern Greek Language*. Athens: Centre of Lexicology.
- Babiniotis, G. (1982). Katastasis enadion katastaseos. In *Glossologia* 1, pp. 119-127 [in Greek].
- Babiniotis, G. (2014). *Dictionary of Difficulties and Common Errors in the Use of Modern Greek*. Athens: Centre of Lexicology.
- Babiniotis, G. (2015). *Dictionary of the most advanced words of Modern Greek*. Athens: Centre of Lexicology.
- Browning, R. (2008). *The Medieval and Modern Greek language* (6<sup>th</sup> edition) (Translation: M. Konomis). Athens: editions Papadimas [in Greek].
- Charalampakis, Ch. (1999). The “correct” and the “incorrect” usage of Modern Greek: Theoretical and practical problems. In *Conference on Greek Language 1976-1996, 29 November-1 December 1996*, Department of Linguistics, University of Athens, Greece, pp. 261-275 [in Greek].
- Chatzisavvidis, S. & Chatzisavvidou, A. (2013). *Grammar of Modern Greek Language- A', B', C' Grades of Gymnasio*. Athens: Ministry of Education, Religious Affairs, Culture and Sports, Diofantos [in Greek].
- Coffey, S. J. (2013). Lexical fossils in present-day English: Describing and delimiting the phenomenon. In R. W. McConchie et al. (eds.) *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3)*. Somerville: Cascadia Proceedings Project, pp. 47-53.
- Dictionary of Modern Greek Language* (1932). G. Zevgolis (ed.). Athens: Proia.
- Dictionary of Standard Modern Greek*. (1998). Thessaloniki: Institute of Modern Greek Studies [Manolis Triandaphyllidis Foundation].
- Dictionary of the Greek Medieval Literature (1100-1669)*. (1968-). E. Kriaras (ed.). Thessaloniki: Centre for the Greek Language.
- Fliatouras, A. (In press -copyright 2015). *The morphological change in the Greek language: A brief presentation*. Athens: Patakis editions [in Greek].
- Fliatouras, A. (Forthcoming). The learned register of Modern Greek as language use and educational goal. Is it a wish or a curse? In *Philologos*, The Department of Philology alumni: Thessaloniki.
- Fliatouras, A. & Koukos, Th. (Forthcoming). Fatseika as kind of Modern Greek joke slang. In *Proceedings of ISTAL 23, 31 March-2 April 2017*, Department of Linguistics, Aristotle University of Thessaloniki [in Greek].
- Fuertes-Olivera, P. A. (2009). The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a Crossroads, Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pp. 99-134.
- Great Dictionary of the Greek Language*. (1936-1950). D. Dimitrakos (ed.). Athens: Dimitrakos editions.
- Greek-English Dictionary Georgakas Online*. Accessed at: [http://www.greek-language.gr/greekLang/modern\\_greek/tools/lexica/georgakas/index.html](http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/georgakas/index.html) [31/03/2018].
- Holton, D., Mackridge P. & Filippaki-Warbuton, I. (1999). *Grammar of the Greek Language*. Athens: Patakis editions [in Greek].
- Iordanidou, A. (2001). *Dictionary of the scholarly expressions in contemporary Greek*. Athens: Patakis editions [in Greek].
- Iordanidou, A. & Papaioannou, P. (2014). The teaching scenario in Language Teaching. In *Studies for the Greek Language* 34, Thessaloniki: Aristotle University of Thessaloniki, pp. 562-574 [in Greek].
- Kamilaki, M. (2009). The learned elements in communication amongst young people: sociopragmatological investigation of the variety [+LEARNED]. PhD Thesis, National and Kapodistrian University of Athens, Athens, Greece. [in Greek].
- Kamilaki, M. (2012). Learned elements as a strategy of verbal humor: Evidence from young speakers of Standard Modern Greek. In G. Fragaki, T. Georgakopoulos & C. Themistocleous (eds.), *Current Trends in Greek Linguistics*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 124-147.



- Kambakis-Vougiouklis, P. & Fliatouras, A. (Forthcoming). The application of the “bar method” on the [+/- learned] variety in the Greek language. In *Proceedings of ICGL13, 7-9 September 2017*. Department of Greek Language, London-Westminster, UK.
- Karantzola, E. & Fliatouras, A. (Forthcoming). Towards a new perspective of the remodeling of the nominal system in Greek: Evidence from Early Modern Greek. In *Language Variety* (honorary volume for A. Ralli). Patras: University of Patras [in Greek].
- Katsogiannou, M. & Xenophondos, K. (2015). The teaching of Mathematics in a multilingual classroom. In *Proceedings of 10<sup>th</sup> ELETO Conference, 12-14 November*, University of Athens, Greece [in Greek].
- Kleris, Ch. & Babiniotis, G. (2005). *Grammar of Modern Greek. Structural-Functional-Communicational*. Athens: Ellinika Grammata [in Greek].
- Kosegian, Ch., Papathanasiou, V. & Fliatouras, A. (2012). Curriculum of Ancient Greek from original texts for Junior High School. In Research project *New School/School of the 21<sup>st</sup> Century*, Greek Ministry of Education, Lifelong Learning and Religious Affairs. Accessed at: [www.ebooks.edu.gr/new/ps.php](http://www.ebooks.edu.gr/new/ps.php) [31/032018] [in Greek].
- Krimpas, P. (2015). Towards the elaboration of the Grammar of Modern Greek law language. In *Proceedings of 10<sup>th</sup> ELETO Conference, 12-14 November*, University of Athens, Greece [in Greek].
- Krimpas, P. (Forthcoming). Pseudo-learned forms and hypercorrection in Modern Greek: a linguistic level-based analysis. In *Proceedings of ICGL13, 7-9 September 2017*. Department of Greek Language, London-Westminster, UK.
- Martzoukou, M., Selimis, St, Katsalirou, A. & Fliatouras, A. (Forthcoming). The formal register of Modern Greek as a basis for learning Ancient Greek: Experimental data and educational considerations. In *Proceedings of 2<sup>nd</sup> SSU, 28-30 September 2017*, Department of Greek Philology, Democritus University of Thrace, Komotini, Greece.
- Müller-Spitzer, C. (2009). Textual structures in electronic dictionaries compared with printed dictionaries: a short general survey. In R. W. Gouws, U. Heid, H. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent developments with special focus on computational lexicography*. Berlin/New York: De Gruyter, pp. 367-381.
- Panaretou, E. (2006). Text sub-genres: Court Rulings. In D. Goutsos, A. Bakakou-Orfanou, S. Koutsoulelou & E. Panaretou (eds.) *The world of texts. Studies in honor of George Babiniotis*. Athens: Ellinika Grammata, pp. 127-139 [in Greek].
- Papanastasiou, G. (2008). *Modern Greek Spelling*. Thessaloniki: Institute of Modern Greek Studies [Manolis Triandaphyllidis Foundation] [in Greek].
- Papanastasiou, G. (2010). Katharevousa: Its nature and contribution to Modern Greek. In C. Karagounis (ed.), *Greek: A language in evolution*, pp. 227-248.
- Petrounias, E. (1984). *Modern Greek Grammar and Comparative Analysis*. Thessaloniki: University Studio Press [in Greek].
- Ralli, A. (2005). *Morphology*. Athens: Patakis editions [in Greek].
- Schierholz, S. (2015). Methods in Lexicography and Dictionary Research. In *Lexikos* 25, pp. 323-352.
- Setatos, M. (1969). The Etymological-Semantics Pair of Learned and Demotic words of Standard Greek. PhD thesis, Aristotle University of Thessaloniki, Thessaloniki, Greece [in Greek].
- Setatos, M. (1992). The functional exploitation of variety in Standard Greek. In *Scientific Yearbook of the Faculty of Philosophy* V. II, Aristotle University of Thessaloniki, Thessaloniki, Greece [in Greek].
- Symeonidis, Ch., Xenis, G. & Fliatouras, A. (2007). *Dictionary of Ancient Greek for Junior High School*. Organisation of Publication of School Books. Athens: Ellinika Grammata [in Greek].
- Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in The Information Age. In H. Bergenholtz, S. Nielsen & S. Tarp (eds) *Linguistic Insights* 90, Bern: Peter Lang AG., pp. 17-32.
- Tobaidis, D. (1978). *The synonym pairs of the learned and the colloquial words of Standard Greek*. Athens: Grigoris editions [in Greek].
- Triandaphyllidis, M. (1963). The dynamics of the non-integrated learned types. In *Collective Words of Manolis Triandaphyllidis, Erevnitika B'*, Thessaloniki: Institute of Modern Greek Studies, [Manolis Triandaphyllidis Foundation], pp. 216-226 [in Greek].
- Utilitarian Dictionary of Modern Greek Language*. (2014). Ch. Charalampakis (ed.). Athens: Academy of Athens.
- Willis, J. (1996). *A framework for task-based learning*. London: Longman.
- Xydopoulos, G. (2007). *Lexicology: An introduction to word and lexicon analysis*. Athens: Patakis editions [in Greek].

# In Praise of Simplicity: Lexicographic Lightweight Markup Language

**Vladimír Benko**

*Slovak Academy of Sciences, L. Štúr Institute of Linguistics*

*E-mail: vladimir.benko@juls.savba.sk*

## Abstract

Our paper presents a simple markup language – *Lexicographic Lightweight Markup Language (LLML)* that has been used for almost the last three decades in the framework of two dozen lexicographic projects carried out by our Institute, as well as in several projects carried out in co-operation with commercial dictionary publishers. While initially trying to solve the problem of insufficient computing power of early *MS-DOS*-based personal computers in early 1990's only, *LLML* is even today the central component of lexicographic workstations our lexicographers work with. Central components of the *LLML* syntax are introduced and exemplified by a sample entry from the *Dictionary of the Contemporary Slovak Language (SSSJ)*. The final part of the paper describes in short some components of the *LLML*-aware toolbox, i.e., programs that are used in our Institute during compilation, validation, proofreading and typesetting of the respective entries. Some of these tools, however, are just a “bonus”, and “low-cost” projects could do even without them.

**Keywords:** lexicographic data representation, lightweight markup language, XML

## 1 Introduction

One of the typical features of lexicographic projects is that they usually take many years – in the case of multivolume works, even decades – to complete. The developments in the area of information technologies, on the other hand, are extremely fast. This usually means that several generations of IT components may change during the life cycle of a project.

Today it is mostly taken for granted that dictionary data (at least in the framework of large-scale lexicographic projects) should be represented as “structured text”, i.e., encoded in XML and complying to some standard, such as ISO 1951 (2007) or TEI-P5 (2018). The advantages of this approach have described in several works (cf. Derouin & Le Meur 2008).

Nevertheless, we are aware of many projects that – for various reasons – do not use XML and represent dictionary data as “formatted text”, i.e., using a standard word processor, such as Microsoft Word (e.g., Apresyan, 2014). Some of them do so just because they are continuing to use the same method as when the project was started years ago, and do not have the resources to change it. The main argument in such a case is usually that “XML is too expensive”, having in mind not only the price of the appropriate software – an XML-aware text editor, or even a full-fledged Dictionary-Writing System (*DWS*) –, but also additional “human costs”, i.e. salaries for IT specialists necessary to support the software, as well as training costs for the lexicographic team. The Microsoft Word format, on the other hand, seems to be “cheap” – the necessary software is usually available anyway, and almost no additional education for the lexicographers is necessary.

There are, of course, many disadvantages to such an approach, with probably the most important being that it is difficult to enforce uniformity in dictionary entry structure, and such data is almost impossible to validate.

On the other hand, it is also worth noting that traditional lexicographers' "mental model" of a dictionary entry maps directly to typefaces and font styles, and working with a *DWS* requires "mental switching" between two models: a "tree-structured" and a "formatted" one. This involves additional mental burden that – especially the older members of lexicographic terms – by not be easily accepted easily.

In our paper we thus introduce a type of dictionary data representation that may be considered a compromise between fully structured XML format and typographical-only format – using a markup language that is nowadays referred to as *Lightweight Markup Language (LML)*. The most important feature of such languages is that their syntax is very simple, the data is readily comprehensible in source form, and no special software (besides a generic text editor) is needed.

## 2 Historical Background and Related Work

*"Lightweight markup languages were originally used on text-only displays which could not display characters in italics or bold, so informal methods to convey this information had to be developed. This formatting choice was naturally carried forth to plain-text email communications. ... In 1986 international standard SGML provided facilities to define and parse lightweight markup languages using grammars and tag implication. The 1998 W3C XML is a profile of SGML that omits these facilities. However, no SGML DTD for most of the LMLs is known."* (Wikipedia, 2018).

From this perspective, we can say that it was the conventions developed in e-mail (and USENET) that evolved into languages like *Markdown*<sup>1</sup> & *reStructuredText*<sup>2</sup>.

Our markup language, now called *Lexicographic Lightweight Markup Language (LLML)*, has also a fairly long history, and its first version was developed in 1990 during the project of retro-digitalization of a one-volume monolingual Slovak dictionary that was later republished (KSSJ, 1997). Despite its history, no (English) paper on *LLML* has yet been published. In 1992 this system was introduced internationally, at the Budapest *COMPLEX '92* Conference (Benko, 1992). However, as it did not appear in the Proceedings, and so only the Conference participants were informed about our efforts. Our paper at the *Slovko 2001* Conference (Benko, 2001), on the other hand, was in Slovak only, so became "hidden" to the international lexicographic community.

Meanwhile, the language (with only minor modifications) has been used in the preparation of more than 20 monolingual and bilingual dictionaries, and is currently used in the framework of the multivolume *Dictionary of the Contemporary Slovak Language* (three volumes already published, five more to come; SSSJ 2006, 2011, 2016).

## 3 LLML

We believe that the main point of *LLML* can be described by the keyword "simple". The language elements can be learned within the first day of use, even by novice lexicographers, and a DIN A5 "cheat sheet" typically contains almost everything they need to know. Moreover the *LLML* type of markup can also be considered "natural", as punctuation marks are traditionally used to enhance the structure of highly complex texts.

<sup>1</sup> <https://daringfireball.net/projects/markdown/>

<sup>2</sup> <http://docutils.sourceforge.net/rst.html>

The main elements of the *LLML* syntax can be summarized as follows:

- A dictionary entry is represented a single block of text, entries are separated by a blank line. Though the length of individual lines is not specified by the language itself, it is recommended to keep lines relatively short.
- A line starting with an exclamation mark is used as an entry identifier; its syntax is project dependent. For our retro-digitization projects this has carried information on page and column numbers; in some early projects where dictionary entries had first been compiled on traditional paper slips, these slip numbers were indicated.
- A line starting with a question mark (optionally preceded by whitespace) is considered as a “comment”, i.e., will not appear in the final output. Comments are useful for communication between the entry author and editor(s), and provide a device to record editorial decisions.
- “Structural breaks”, such as new sense, phraseology zone or run-on, begin on a new line indented by two spaces.
- The respective “information fields” of the entry are indicated by a small set of punctuation and special characters. The actual syntax may slightly differ from one project to another. Table 1 shows the actual syntax used within the *SSSJ* project.

Table 1: Main *LLML* Syntax Elements (*SSSJ* Dialect)

<i>LLML Element</i>	<i>Default rendering</i>
"headword"	<b>headword</b>
"headword^1"	<b>headword<sup>1</sup></b> (headword with index)
"%substandard headword"	<b>substandard headword</b>
"*incorrect headword"	*incorrect headword
"~crossref headword"	<b>crossref headword</b>
[pronunciation]	[pronunciation]
*PoS label	pos label
other label	other label
<etymology>	<u>etymology</u>
'example text'	<i>example text</i>
[*reference]	[reference]
{1}, {2}, ...	<b>1., 2., ...</b> (sense numbers)
{M}, {T}, ...	□, □, ... (special symbols indicating “structural breaks” in entry structure)
(unmarked)	(unmarked) ... definitions, explanations, etc.

As the *LLML* syntax is very similar to that of programming languages, by using text editor featuring user definable syntax highlighting the respective information fields in colors, the lexicographer’s work becomes even more user-friendly. We hope that the reader can appreciate its legibility in Figure 1, showing a screenshot of an example entry as displayed by the Notepad++ editor using a custom “language definition”.

Identification and comment lines are displayed in gray, so that the entry text itself is highlighted. The cyan vertical line at the right margin indicates the suggested line length, and other colors highlight the respective structure elements.



```

309 |10260
310 "légia" -ie |p1. G| -ií |D| -iám |L| -iách |*ž.| <lat.>
311 {1} |hist., voj.| najväčšia a hlavná bojová jednotka
312 rímskej armády, ktorú tvorilo niekoľko kohort: 'rímske
313 légie'; '1. cisára Marca Aurelia'; 'veliteľ légiám'; 'poraziť
314 légie'; 'privolať na pomoc légie z Východu'; 'Rím,
315 zákonodarca a vládca sveta, sídlo nepremožiteľných
316 légii.' [*Anton Hlinka]
317 {2} |voj.| dobrovoľná vojenská jednotka: 'veliteľ,
318 príslušník légie'; 'v roku 1916 vstúpil do
319 československých légii v Rusku'
320 {T} |hist.| 'Biela légia' protikomunistické ilegálne
321 hnutie pôsobiace na Slovensku v r. 1948 - 1955; 'légia
322 Kondor' nacistická vojenská jednotka vyslaná Hitlerom
323 do Španielska počas španielskej občianskej vojny
324 {M} 'Cudzinecká légia' francúzska špeciálna jednotka
325 tvorená zahraničnými dobrovoľníkmi, súčasť francúzskej
326 armády, v súčasnosti nasadzovaná v rámci mierových
327 humanitných operácií vojenských síl OSN
328 {3} |expr.| veľký počet niečoho, množstvo ľudí, dav:
329 '1. básnikov'; 'Mačiek sa vrátila celá légia, lebo sa
330 niekde nakotili a spätnásobili.' [*Š. Žáry]; 'Federálne
331 a miestne vlády zamestnávajú celé légie inšpektorov,
332 ktorí udeľujú vysoké pokuty.' [*HN 2003]
333 ?a IKPMV
334 ?b NK
335 ?? Anton Hlinka - Ozvena slova 1 - Blahozvešť v horizonte
336 ?? ľudskej skúsenosti 2, 1996
337 ?? Štefan Žáry - Apeninský vzduch, 1984
338 ?? Teofil Klas - Putovanie do Loreta, 1999

```

Figure 1: Lexicographic Lightweight Markup Language (LLML) text editor screenshot (Notepad++)

## 4 Data Validation

The LLML approach to lexicographic data representation does not allow for full-scale data validation, but essential data checks can be performed. A special validation parser had to be written from the very beginning of using LLML in order to detect errors with regard to the “well-formedness” of the dictionary data, such as unmatched syntactic structures and non-sequential appearance of sense numbers. With the advent of text editors with color syntax highlighting the former problem became less acute, as unmatched “tags” are usually immediately apparent by “spoiled” colors. The tool, nonetheless, proved to be quite useful and is being gradually enhanced to include checking for the presence and/or absence of whitespace around punctuation, presence of suspicious special characters, etc. The error report produced by the validation parser always contains a detailed error message, including the respective excerpt from the input file indicating the affected line numbers, such as those at Figure 2.

```

157 ***** po čiarku má byť medzera
157: {1} |lit.| literárny žáner o živote svätých, v ktorom

228-231 ***** číslo významu mimo poradia
228: {1} vlastnosť toho, čo sa zakladá na legendách,
229: vymyslených, neskutočných veciach: '1. príbehu';
230: 'literatúra s črtami legendárnosti'
231: {3} vlastnosť toho, čo je nezabudnuteľné, výnimočné,

```

Figure 2: Error report generated by the Validation Parser



The first message (line 157 of the source file) indicates a missing space after a comma, and the second message (lines 228 to 231) states that a sense number out of sequence has been encountered. The lexicographer can usually detect the exact cause of the issue from the error report itself, without the need to study the larger context of the source file.

## 5 LLML Toolbox

In this section we want to mention some other parts of our *LLML*-aware toolbox that are needed to cover the whole process of dictionary creation:

- (1) “Paragraph grep” – an open-source *perl* script used to extract dictionary entries. This provides for the creation of ad-hoc lists of entries based on regular expressions,
- (2) “Paragraph sort” – custom sort using marked headword as sort keys. A simple modification of a standard (quicksort-based) sort.
- (3) Converter to proofreading format (an MS-Word document in RTF format retaining original line breaks and comments, and indicating entry identifiers and line numbers (Figure 3). The line numbers are convenient in subsequent editing of the dictionary data.

```

310 10260 légia -ie pl. G -íí D -iám L -iách ž. <lat.>
311      {1} hist., voj. najväčšia a hlavná bojová jednotka
312      rímskej armády, ktorú tvorilo niekoľko kohort: rímske
313      légie; l. cisára Marca Aurelia; veliteľ légiám; poraziť
314      légie; privolať na pomoc légie z Východu; Rím,
315      zákonodarca a vládca sveta, sídlo nepremožiteľných
316      légii. [Anton Hlinka]
317      {2} voj. dobrovoľná vojenská jednotka: veliteľ,
318      príslušník légie; v roku 1916 vstúpil do
319      československých légii v Rusku
320      {T} hist. Biela légia protikomunistické ilegálne
321      hnutie pôsobiace na Slovensku v r. 1948 – 1955; légia
322      Kondor nacistická vojenská jednotka vyslaná Hitlerom
323      do Španielska počas španielskej občianskej vojny
324      {M} Cudzinecká légia francúzska špeciálna jednotka
325      tvorená zahraničnými dobrovoľníkmi, súčasť francúzskej
326      armády, v súčasnosti nasadzovaná v rámci mierových
327      humanitných operácií vojenských síl OSN
328      {3} expr. veľký počet niečoho, množstvo ľudí, dav:
329      l. básnikov; Mačiek sa vrátila celá légia, lebo sa
330      niekde nakotili a späťnásobili. [Š. Žáry]; Federálne
331      a miestne vlády zamestnávajú celé légie inšpektorov,
332      ktorí udeľujú vysoké pokuty. [HN 2003]
333      ?a IKPMV
334      ?b NK
335      ?? Anton Hlinka — Ozvena slova 1 — Blahozvešť v horizonte
336      ?? ľudskej skúsenosti 2, 1996
337      ?? Štefan Žáry — Apeninský vzduch, 1984
338      ?? Teofil Klas — Putovanie do Loreta, 1999

```

Figure 3: Proofreading format (line breaks and comments retained, line numbers indicated)

- (4) Converter to typesetting format (entry identifiers and comments deleted, Figure 4). This is in fact the only “compulsory” part of the toolbox. It works in two phases: firstly the *LLML* data is converted to “presentation type” of *XML* format, i.e., indicating the respective typefaces the information fields are mapped into. The second step can use any standard tool for *XML* conversion, in our case generating an *RTF* format that can be imported into the respective publishing system.

**légia** -ie pl. G -íí D -iám L -iách ž. ⟨lat.⟩ 1. hist., voj. ▶ najväčšia a hlavná bojová jednotka rímskej armády, ktorú tvorilo niekoľko kohort: *rímske légie; l. cisára Marca Aurelia; veliť légiám; poraziť légie; privolať na pomoc légie z Východu; Rím, zákonodarca a vládca sveta, sídlo nepremožiteľných légií*. [Anton Hlinka] 2. voj. ▶ dobrovoľná vojenská jednotka: *veliteľ, príslušník légie; v roku 1916 vstúpil do československých légií v Rusku* □ hist. *Biela légia* protikomunistické ilegálne hnutie pôsobiace na Slovensku v r. 1948–1955; *légia Kondor* nacistická vojenská jednotka vyslaná Hitlerom do Španielska počas španielskej občianskej vojny □ *Cudzinecká légia* francúzska špeciálna jednotka tvorená zahraničnými dobrovoľníkmi, súčasť francúzskej armády, v súčasnosti nasadzovaná v rámci mierových humanitných operácií vojenských síl OSN 3. expr. ▶ veľký počet niečoho, množstvo ľudí, dav: *l. básnikov; Mačiek sa vrátila celá légia, lebo sa niekde nakotili a spätňásobili*. [Š. Žáry]; *Federálne a miestne vlády zamestnávajú celé légie inšpektorov, ktorí udeľujú vysoké pokuty*. [HN 2003]

Figure 4: Final format (typeset entry)

## 6 Conclusion and Further Work

From the early 1990s, when our first *LLML*-based projects started, we considered it as something temporary that should (and will) eventually be replaced by a more sophisticated representation. During the *SGML* period, however, our computing equipment was firstly not powerful enough to implement it, and secondly the *SGML*-aware software was also far beyond what we could afford to buy. With the advent of XML, the situation has changed dramatically, both in terms of the computing power of our equipment and availability of affordable software tools.

We have observed the efforts of introducing the *XML* technology at our partner institutions that, surprisingly, turned out to be not as straightforward and easy as we would imagine. Though the computing power of modern workstations is no longer the main problem, another scarce resource appeared: the *XML*-based projects require much more (human) IT support. This is probably why we are still reluctant to switch our main project to it. We realize, however, that the day is approaching, and that better interoperability will most likely be the motivating main reason.

One of our anonymous reviewers noted that “... *I accept that you use the framework what you describe, but I doubt that it could be recommended for other (new) dictionary projects as a standard instead of XML*”. This is naturally difficult to argue against, yet there is at least one area where the *LLML*-based approach can be of an advantage even today: the dictionary retro-digitization projects involving manual proofing of OCR-ed material. According to our experience, it is convenient here to split the process into two separate phases, with the first aimed to achieving only the “typographical identity” – an explicit, *simple* markup can ease the whole process significantly.

## References

- Benko, V. (1992). *Late Computational Support for a Dictionary Project*. Presentation at the COMPLEX '92 International Conference. Budapest, Hungary. (unpublished).
- Benko, V. (2001). *Počítačová podpora lexikografických projektov – retrospektívny pohľad*. (Computational Support of Lexicographic Projects – A Retrospective View). In: Jarošová, A. (ed) *Slovenčina a čeština v počítačom spracovaní*. Proceedings of the Slovko 2001 Conference. Bratislava: VEDA.
- ASRYa (2014). Apresyan, Yu. (Ed.) *Aktivnyj slovar' russkogo yazyka*. Tom 1. A – B. Yazyki slavyanskoj kul'tury, Moskva.
- Derouin, Marie-Jeanne and Le Meur, André (2008). *ISO-Standards for Lexicography and Dictionary Publishing*. Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, 2008. pp. 663 – 668, ISBN 978-84-96742-67-3.
- KSSJ (1997). *Krátky slovník slovenského jazyka*. Red. J. Kačala – M. Pisárčiková. 3. dopl. a preprac. vyd. Bratislava: Veda 1997. 943 s. ISBN 80-224-0464-0
- ISO (2007). ISO 15924:2006(en), Presentation/representation of entries in dictionaries — Requirements, recommendations and information.
- SSSJ I (2006). *Slovník súčasného slovenského jazyka. A – G*. Hl. red. K. Buzássyová – A. Jarošová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied 2006. 1134 p. ISBN 978-80-224-0932-4
- SSSJ II (2011). *Slovník súčasného slovenského jazyka. H – L*. Ved. red. A. Jarošová – K. Buzássyová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied 2011. 1087 p. ISBN 978-80-224-1172-1
- SSSJ III (2016). *Slovník súčasného slovenského jazyka. M – N*. Ved. red. A. Jarošová, Bratislava: Veda, vydavateľstvo SAV 2015. 1100 s. ISBN 978-80-224-1485-2.
- TEI (2018). TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>
- Wikipedia (2018). Lightweight markup language. [https://en.wikipedia.org/wiki/Lightweight\\_markup\\_language](https://en.wikipedia.org/wiki/Lightweight_markup_language). [15/06/2018]

## Acknowledgement

This work has been, in part, financially supported by the Slovak VEGA Grant Agency, Project No. 2/0017/17.



# Corpus-based Cognitive Lexicography: Insights into the Meaning and Use of the Verb *Stagger*

**Thomai Dalpanagioti**

Centre for the Greek Language

E-mail: dalpanagiotith@yahoo.gr

## Abstract

Situated within the framework of “cognitive lexicography”, this paper aims to demonstrate how lexical meaning and usage patterns can be represented in a coherent and principled manner by applying cognitive semantic theories to corpus data. The focus of attention is on the lexicographic tasks of establishing lexical units, capturing usage patterns and providing definitions. The proposed corpus-based and cognitively-oriented approach is applied to a lexical item from the semantic field of motion, the verb *stagger*. Monolingual learner’s dictionaries (MLDs) are examined as to their *stagger* entry in order to specify in what respects this approach can improve EFL lexicography. The paper is not restricted to a theoretical discussion of lexicographic issues or a critical review of existing entries; rather, a new version of the *stagger* entry is offered.

**Keywords:** frame semantics; conceptual metaphor and metonymy theory; word sense disambiguation; usage patterns

## 1 Introduction

The application of cognitive linguistics to lexicography has led to the emergence of a new interdisciplinary research field called “cognitive lexicography” (Ostermann 2015). This paper aims to contribute to this field by demonstrating how cognitive linguistic theories can be applied to corpus data in building up a dictionary entry. To this end, the polysemous manner-of-motion verb *stagger* is used as a case study.

The study builds on the idea that the compilation of dictionary entries can be systematized by structuring corpus-derived lexical information in a pre-lexicographic database (Atkins & Rundell 2008: 100-101). To interpret the data, we draw on Corpus Pattern Analysis (Hanks 2013a: 404), Frame Semantics (Fillmore 1982), the Conceptual Metaphor and Metonymy Theory (Lakoff & Johnson 1980) and the Principled Polysemy approach (Evans & Green 2006: 342-352).

After outlining this theoretical background, the paper presents the corpus-based and cognitively-oriented analysis of the meaning and use of the verb *stagger*. Then, we briefly examine the *stagger* entry in the “Big Five” MLDs (i.e. online editions of OALD, LDOCE, COBUILD, CALD, MEDAL) with respect to sense distinctions, usage patterns and definitions. By way of conclusion, we present a new version of the *stagger* entry, which is claimed to be more enlightening to users.

## 2 Corpus-based Cognitive Lexicography

“Corpus-based cognitive lexicography” brings together corpus linguistics, cognitive linguistics and lexicographic practice. Corpus linguistics has revolutionized lexicography by providing access to vast amounts of authentic language data. The empirical approach to the study of language has promoted



the contextual view of meaning, changing not only the source of data but also its presentation in dictionaries, so as to demonstrate use in context; EFL lexicography has been a pioneer in this respect (Rundell 1998: 320, 330). Cognitive linguistics can contribute to lexicography in a different way, i.e. by making dictionary entries more reasonable and streamlined; as Geeraerts (2007: 1168) has pointed out, what cognitive linguistics can offer to lexicography is a more realistic conception of semantic structure.

The combination of corpus linguistics and cognitive linguistics is not something new, in particular with regard to the issue of distinguishing senses (e.g. Gries & Stefanowitsch 2006). In lexicographic practice, we can find dictionaries which explicitly combine some cognitive insights with corpus-based lexicography (e.g. Moon 2004); however, this can be done more widely and systematically by applying a variety of cognitive linguistic theories and corpus approaches to various parts of dictionary entries and aspects of lexicographic work. In this light, the present study aims to demonstrate how cognitive semantics can operate in combination with a corpus approach to improve the treatment of lexical meaning and use in EFL dictionaries. This section summarizes the approaches upon which this study relies most.

The power of the corpus to foreground the syntagmatic aspect of lexis has led to a contextual theory of meaning expressed in statements such as “every distinct sense of a word is associated with a distinction in form” (Sinclair 1987: 89), and “context disambiguates” (Moon 1987: 87). Viewing meaning as function in context, Sinclair (1998: 14-23) has proposed four categories of co-selection (i.e. “collocation”, “colligation”, “semantic preference”, and “semantic prosody”) as components of an extended unit of meaning, which can be identified only by observing “the cumulative effect of usage” in corpora (Tognini-Bonelli 2002: 73). Building on this idea, Hanks (2004a: 246-251) argues that the lexicographer’s task is to identify the normal patterns of the usage of words by means of a Corpus Pattern Analysis (CPA), and to present these norms rather than their exploitations in dictionaries. As Hanks (2004b: 88) explains, in CPA “concordance lines are grouped into semantically motivated syntagmatic patterns”. A “pattern” in the Pattern Dictionary of English Verbs (PDEV) – a CPA product – includes information on the semantic types of arguments that are relevant for distinguishing between different senses.

The present study combines CPA with Frame Semantics and takes account of their application in PDEV and FrameNet, respectively. Both of these pioneering projects “seek to identify stereotypical contexts”; while PDEV focuses on the “phraseological context”, FrameNet concentrates on “the context of situation in which words are used” (Hanks 2013b: 729). The main assumption of Frame Semantics is that words must be grouped and explained in relation to a “(semantic) frame”, i.e. a structured background of experience which constitutes a kind of prerequisite for understanding the meaning of a word (Fillmore 1985: 224). Every semantic frame consists of specific “frame elements” (FEs), i.e. the “various participants, props, and other conceptual roles” involved in the schematic representation of a situation (Fillmore & Petruck 2003: 359). Frame semantics links these situation-specific semantic roles to their syntactic realizations (grammatical functions and phrase types), thus specifying valence in both semantic and syntactic terms. Work in the FrameNet project involves developing frame descriptions, establishing lexical units (LUs, i.e. words in one of their senses) as annotation targets, extracting example sentences from the BNC, and annotating them in terms of FEs, phrase types, and grammatical functions (Ruppenhofer et al. 2016: 7-8).

CPA and Frame Semantics may be regarded as complementary approaches, as “each CPA pattern can in principle be plugged into a FrameNet semantic frame” (Hanks 2018); in fact, PDEV includes direct links between its verb patterns and FrameNet frames. FrameNet is claimed to be valuable for lexicography, because it provides principles for identifying what is lexicographically relevant in the overwhelming amount of corpus data (Atkins, Fillmore & Johnson 2003; Atkins, Rundell & Sato

2003). However, it is also criticized for randomly selecting the frames and the LUs to be analyzed and for adopting a “top-down” (as opposed to “bottom-up”) corpus approach (Hanks 2013b: 729; Johnson & Lenci 2013: 45). Therefore, our lexicographic study of *stagger* can benefit from combining CPA’s methodology of identifying normal patterns with FrameNet’s in-depth analysis of semantic frames in discovering usage patterns and distinguishing senses. However, this theoretical background needs to be complemented by a framework for organizing senses and uses into a coherent and motivated network; the Conceptual Metaphor and Metonymy Theory and the Principled Polysemy approach are relevant in this respect.

Cognitive semanticists view metaphor and metonymy as phenomena fundamental to the structure of the conceptual system rather than mere stylistic features of language (Lakoff & Johnson 1980; Croft 2000; Kövecses 2002; Evans & Green 2006). The cognitive mechanisms of metaphor and metonymy account for lexical semantic extension, as they can show the relationship between multiple synchronic uses of a given form. Metaphor and metonymy differ in terms of the function of the conceptual relationship (imagistic reasoning in metaphor vs. shift of reference in metonymy) and its nature (similarity in metaphor vs. contiguity in metonymy). Mapping occurs across two separate domains (domain mapping) in metaphor, whereas in metonymy it occurs within a single domain (domain highlighting). Nevertheless, the distinction between metaphor and metonymy is not absolute, but rather scalar (Radden 2003) and metaphorical mappings often have a metonymic basis.<sup>1</sup>

The metaphorical/metonymic extensions that are of particular lexicographic interest are those that have achieved conventional status, as opposed to ad-hoc coinages (Hanks 2004a: 272). While no one would disagree about the type of linguistic metaphors/metonymies to be included in dictionaries, there are opposing views with regard to their arrangement within a dictionary entry. According to the frequency-based approach, highly frequent metaphorical extensions should precede less frequently occurring literal meanings (Hanks 1987: 133-134). In contrast, according to the semantic order approach, the core meaning should precede extended uses. In particular, Van der Meer (1999) argues that making users – and especially learners – aware of the metaphorical extensions of words, by ordering senses in the dictionary from literal to figurative, facilitates vocabulary learning and especially understanding of subtle shades of meaning. Apart from making metaphorical extension implicit in the entry structure, dictionaries may explicitly mark it with a label (“figurative”). However, this is not recommended because labels are considered to be a “blunt instrument”, as they mean more to the lexicographer than the user (Atkins & Rundell 2008: 496, 498); instead, it is better to show the relation of senses in the wording of definitions. Against this background, we will attempt to reflect the metaphoric/metonymic relation between senses in both the structure and the definitions of the proposed entry.

To this end, it is also useful to take account of the Principled Polysemy approach which seeks to “develop clear decision principles that make semantic network analyses objective and verifiable” (Evans & Green 2006: 342). In brief, according to this approach, the semantic network of a polysemous word must be organized around the synchronically prototypical sense, from which other senses are naturally derived with varying degrees of relatedness; distinct senses must contain additional meaning, and manifest specific collocational patterns and/or grammatical structures (Evans 2005: 38-44). In this light, in the present study the core motion meaning of the verb *stagger* is used as a basis on

1 Two instances of metonymy-based metaphors, which are mentioned in the case study below, are ACTION IS MOTION and EMOTION AS MOTION. With regard to the ACTION IS MOTION metaphor, Kövecses and Radden (1998: 61) explain that since “MOTION may be seen as a subcategory of ACTION”, this metaphor may be understood “as ultimately deriving from the conceptual metonymy MEMBER OF A CATEGORY FOR THE CATEGORY”, i.e. a part-whole relationship. Similarly, Niemeier (2003: 195, 209) argues that metaphors related to emotions are dependent on a conceptually prior metonymic relationship; for example, the EMOTION AS MOTION metaphor is experientially grounded on the metonymy PHYSIOLOGICAL/ BEHAVIORAL EFFECT FOR EMOTION.

which cognitive processes (metaphor, metonymy) are applied to justify semantic extension. Evans's (2005) meaning criterion is defined here in frame-semantic terms as involving additional or different FEs, while lexicogrammatical patterns are identified by means of CPA.

### 3 Case Study: *To Stagger*

The aim of this section is to implement the corpus-based and cognitively-oriented framework described above in the analysis of the verb *stagger*.

#### 3.1 Corpus Data

The data for the study is drawn from two corpora, i.e. the BNC and the ukWaC, accessed through the Sketch Engine interface. The combined use of these corpora provides a more representative basis for the analysis, as they differ in size (759 vs. 3,565 *stagger* occurrences, respectively) and coverage of text types. As Ferraresi et al. (2008: 7) point out, the BNC has a higher proportion of narrative fiction texts and spoken texts, and is characterized by a stronger historical perspective, whereas the ukWaC contains comparatively more texts dealing with the Web, education and public sphere issues, and is characterized by a stronger concern with the present time.

The Word Sketches derived for *stagger* from the two corpora help us make some preliminary remarks about its usage patterns. Both Word Sketches indicate that most of the times the verb is followed by a prepositional phrase (e.g. *to, into, from* + NP), a particle (e.g. *off, along, out, home*), or an adverb (e.g. *backwards, drunkenly*) – contextual cues that point to the basic motion sense of the verb. However, the ukWaC Word Sketch shows more evidently that *stagger* also occurs in non-motion contexts; for example, collocates in object position include *joint, start, hour, time*, and patterns such as *staggered by the amount of, staggered to find/learn/see* are explicitly recorded. Word Sketches are hence used as a springboard for a thorough analysis of concordances; they make it easier to identify separate senses when scanning corpus examples.

In short, the process followed in the study includes first examining the Word Sketches derived for *stagger* from the two corpora and then analyzing all occurrences of the verb in a large random sample (i.e. 60% of the *stagger* uses in each corpus). Uses are clustered together and LUs are established in the way illustrated in the following section.

#### 3.2 Word Sense Disambiguation

To interpret the corpus data, we apply the integrated approach to word sense disambiguation outlined in Section 2. The first step in the process of establishing LUs involves identifying the frame evoked by *stagger* in each corpus sentence and annotating its predicate-argument structure in terms of FEs, phrase structure and grammatical structure, following FrameNet's practice. Separate senses generally correspond to different semantic frames and assign different frame elements (FEs) (Atkins 2008: 256-257; Atkins, Rundell & Sato 2003: 335-337).

Table 1 demonstrates how sample corpus sentences are clustered under semantic frames. In an attempt to exhaustively analyse the polysemy of the verb under study, we develop a frame-semantic analysis of LUs currently missing from FrameNet. Whereas FrameNet records *stagger* under two frames (i.e. [Self\_motion] and [Stimulate\_emotion]), we identify four relevant frames. The FE annotation may generally demonstrate the straightforward applicability of the existing FrameNet descriptions; yet, metaphorical uses under 1b seem to pose a subtle problem for frame assignment. In light of the Principled Polysemy approach, we have decided to distinguish these uses from the literal ones in 1a, and

mark them as evoking the [Self\_motion]<sub>figurative</sub> sub-frame, because they exhibit distinct co-occurrence patterns that affect aspects of meaning (see Section 3.3).<sup>2</sup>

To lend further support to the frame-based sense distinctions, we consider how they are motivated by the cognitive mechanisms of metaphor and metonymy. In this respect, Table 2 points out the non-arbitrary relationship between the semantic extensions of *stagger* and proposes a rational arrangement of the LUs. More precisely, the core [Self\_motion] LU is related, on the one hand, to the [Self\_motion]<sub>figurative</sub> frame-evoking LU by means of the EVENT STRUCTURE metaphor (MANNER OF ACTION IS MANNER OF MOTION), and, on the other hand, to the [Cause\_motion] frame-evoking LU by means of the ACTIVITY FOR CAUSED EVENT metonymy. Similarly, different conceptual metonymies and metaphors account for the [Stimulate\_emotion] and [Arranging] frame-evoking LUs, which extend from the [Cause\_motion] one. As the fourth column of Table 2 shows, LUs can be organized into a tiered structure with two main clusters of related senses.

At this point, we should also note that the [Cause\_motion] LU poses the following dilemma: to record it in the dictionary entry because it motivates other senses or not to record it because it is relatively infrequent? The decision relies on the purpose of the dictionary and the principle (coherence vs. frequency) adopted. The issue is further discussed in Sections 4 and 5.

Table 1: Assigning semantic frames to corpus examples.

1. Frame: [Self_motion] <sup>3</sup>
<i>a</i>
(1) <u>Chopra</u> <sub>SELF_MOVER</sub> <b>staggered</b> , and slumped on to the floor.
(2) There were a lot of young drunks <u>staggering about</u> <sub>AREA</sub> .
(3) Now, as he watched <u>him</u> <sub>SELF_MOVER</sub> limping and <b>staggering up the slope</b> <sub>PATH</sub> , it occurred to hint that he might actually be wounded in some way.
(4) <u>She</u> <sub>SELF_MOVER</sub> <b>staggered home</b> <sub>GOAL</sub> and called help.
(5) <u>The porter</u> <sub>SELF_MOVER</sub> <b>staggered drunkenly</b> <sub>MANNER</sub> <u>to his feet</u> <sub>GOAL</sub> .
(6) She had not finished exhorting Dr Neil about this when <u>McAllister</u> <sub>SELF_MOVER</sub> , who could hear every word in the kitchen, returned with the tea-tray, <b>staggering under its weight</b> <sub>EXTERNAL_CAUSE</sub> .
(7) <u>He</u> <sub>SELF_MOVER</sub> could have been shot at close range to the device and <b>staggered 20 yards</b> <sub>DISTANCE</sub> <u>away</u> <sub>SOURCE</sub> before collapsing.
<i>b</i>
(8) <u>Both of them</u> <sub>SELF_MOVER</sub> recovered, and <b>staggered on</b> <sub>DIRECTION</sub> <u>through the year</u> <sub>PATH</sub> .
(9) When Gilda heard what happened, she said, ‘ <u>A man who</u> <sub>SELF_MOVER</sub> ’s just <b>staggered out of a nasty relationship</b> <sub>SOURCE</sub> wants a bloody nursemaid at first, and then he wants to play the field for a bit’.
(10) And, as <u>his administration</u> <sub>SELF_MOVER</sub> <b>staggered through its winter of discontent</b> <sub>PATH</sub> in the first two months of 1979, Callaghan’s famed skills as a crisis-manager seemed to desert him.

2 For further evidence on the [Self\_motion]<sub>figurative</sub> sub-frame, see Dalpanagioti (2013: 19-20). Metaphorical uses are not treated systematically in FrameNet, since “such work is worthy of an entire research project in itself” (Ruppenhofer et al. 2016: 101). Similarly, with regard to the metaphorical LU of *stagger* evoking the [Stimulate\_emotion] frame, no annotated sentences are currently available in FrameNet and there is no indication of the relation between the source and target frames.

3 Frame definition: “The SELF\_MOVER, a living being, moves under its own direction along a PATH. Alternatively or in addition to PATH, an AREA, DIRECTION, SOURCE, or GOAL for the movement may be mentioned” (FrameNet).



- (11) A lot of marriages<sub>SELF\_MOVER</sub> **staggered** along<sub>PATH</sub> with less.
- (12) Rarely has so important a constitutional bill<sub>SELF\_MOVER</sub> **staggered** towards enactment<sub>DIRECTION</sub> so inelegantly.
- (13) The company<sub>SELF\_MOVER</sub> was a victim of the mania for leveraged buyouts of the late 1980s and has been **staggering** under its burden of debt<sub>EXTERNAL\_CAUSE</sub>.
- (14) Agriculture and the transport system<sub>SELF\_MOVER</sub> were likewise soon **staggering** under the strains<sub>EXTERNAL\_CAUSE</sub> imposed by war.
- (15) Now the Boston Museum of Fine Arts<sub>SELF\_MOVER</sub>, struggling to service debt on its 1980s expansions, is **staggering** under the weight of a nearly \$5 million deficit<sub>EXTERNAL\_CAUSE</sub>.
- (16) The economy<sub>SELF\_MOVER</sub> continued to **stagger** from crisis<sub>SOURCE</sub> to crisis<sub>GOAL</sub>.

## 2. Frame: [Cause\_motion]<sup>4</sup>

- (17) The collision<sub>AGENT</sub> **staggered** her<sub>THEME</sub> and she fell.
- (18) He<sub>AGENT</sub> slapped Bicker on the back, **staggering** him<sub>THEME</sub>, then turned to Riven.
- (19) The vicious slap<sub>AGENT</sub> on Polly's soft cheek had for a moment **staggered** the little girl<sub>THEME</sub>, but Polly was the stuff that heroines are made of.
- (20) So unprepared was Ryan that the blow<sub>AGENT</sub> **staggered** him<sub>THEME</sub> sideways<sub>DIRECTION</sub>, his head smashing into the corridor.
- (21) Christy<sub>AGENT</sub> reacted instinctually, firing her other fist into Laura's face and **staggering** her<sub>THEME</sub> backwards<sub>DIRECTION</sub>.

## 3. Frame: [Stimulate\_emotion]<sup>5</sup>

- (22) The sense of effort<sub>STIMULUS</sub> in his conversation **staggered** her<sub>EXPERIENCER</sub>, and she watched him with pity, for he laboured as though he were to try to write a sonnet.
- (23) The howl of dissent<sub>STIMULUS</sub> that came from the entire room **staggered** me<sub>EXPERIENCER</sub>.
- (24) This<sub>STIMULUS</sub> **staggered** Mrs Funnell<sub>EXPERIENCER</sub> into silence<sub>RESULT</sub>, and her feelings were registered on her grim face as she watched this husband and wife whom she had never liked.
- (25) Rory<sub>EXPERIENCER</sub> **was staggered** by his answer<sub>STIMULUS</sub>.
- (26) All-who met him<sub>EXPERIENCER</sub> **were staggered** by his easy command of innumerable languages, by his polished manners, by his superb musicianship and by his mastery of courtly pursuits<sub>STIMULUS</sub> like hunting and playing chess – not to mention his astonishing good looks.

4 Frame definition: "An AGENT causes a THEME to move from a SOURCE, along a PATH, to a GOAL. Different members of the frame emphasize the trajectory to different degrees, and a given instance of the frame will usually leave some of SOURCE, PATH and/or GOAL implicit" (FrameNet).

5 Frame definition: "Some phenomenon (the STIMULUS) provokes a particular emotion in an EXPERIENCER" (FrameNet).



4. Frame: [Arranging]<sup>6</sup>*a*

- (27) The two side groups<sup>THEME</sup> are no longer eclipsed but become slightly **staggered**.
- (28) The framework is then clad with two layers of 12mm plasterboard, the second layer<sup>THEME</sup> being **staggered** so that the joints do not coincide.
- (29) Many of the tables<sup>THEME</sup> were screened from one another, and **staggered** over different levels<sup>CONFIGURATION</sup>.

*b*

- (30) This led to the evacuation of 250.000 people<sup>THEME</sup> which, although **staggered** over several days<sup>CONFIGURATION</sup>, led to great pressure on transport arteries.
- (31) We will be able to do this because, for the first time in a World Cup, it seems there will be three different kick-off times each day with all three daily matches<sup>THEME</sup> being **staggered**.
- (32) I just wish they<sup>AGENT</sup>'d **stagger** lecture finishing times<sup>THEME</sup> because the corridors just get so totally packed.
- (33) During the season, the NEDC group<sup>AGENT</sup> sees further benefits in amending the traditional school summer holiday, **staggering** the starting dates<sup>THEME</sup> to avoid the sudden rush of tourist traffic.

Table 2: Building a motivated semantic network.

Corpus-attested examples	Frame	Motivation	Structure
<i>There were a lot of young drunks <b>staggering</b> about.</i>	[Self_motion]	core meaning: to walk unsteadily	<b>1a</b>
<i>The economy continued to <b>stagger</b> from crisis to crisis.</i>	[Self_motion] <sub>figurative</sub>	EVENT STRUCTURE metaphor (MANNER OF ACTION IS MANNER OF MOTION)	<b>1b</b>
<i>The blow <b>staggered</b> him sideways, his head smashing into the corridor.</i>	[Cause_motion]	ACTIVITY FOR CAUSED EVENT metonymy	<b>2a</b>
<i>The howl of dissent that came from the entire room <b>staggered</b> me.</i>	[Stimulate_emotion]	BEHAVIORAL EFFECT FOR EMOTION metonymy, EMOTION AS MOTION metaphor	<b>2b</b>
<i>The tables were <b>staggered</b> over different levels. I wish they'd <b>stagger</b> lecture finishing times because the corridors just get so totally packed.</i>	[Arranging]	MANNER OF MOTION ALONG THE PATH FOR CONFIGURATION OF THE PATH metonymy (fictive motion), TIME IS SPACE metaphor	<b>2c</b>

<sup>6</sup> Frame definition: "An AGENT puts a complex THEME into a particular CONFIGURATION, which can be a proper order, a correct or suitable sequence, or a spatial position" (FrameNet).

### 3.3 Usage Patterns

The results of applying CPA to the verb under study are summarized in Table 3. It presents the typical co-occurrence patterns identified in the two corpora examined, and makes it clear that each LU exhibits distinct patterns. Following CPA, we specify the semantic type of the verb's arguments and demonstrate that different semantic types in the same syntactic slot can give rise to different senses. At the same time, each pattern is connected with a FrameNet frame, and semantic types of argument fillers are associated with FEs.

The combination of CPA and FrameNet features constitutes a powerful tool for representing the combinatorial behavior of the LUs and thus accurately discriminating between senses. This becomes obvious if we compare Table 3 to PDEV *stagger* entry (comprised of four patterns) and FrameNet's *stagger* LUs (two patterns); Table 3 seems to complement both resources by providing a more detailed and coherent picture of the senses and usage patterns of *stagger*. At this point, we should also note that, although PDEV attempts to link its verb patterns to FrameNet frames, only one link actually works in its *stagger* entry, directing to the [Self\_motion] frame; the rest of the patterns are not connected to FrameNet.

Table 3: Combining CPA and FrameNet.

LU	Frame	Corpus Patterns
1a	[Self_motion]	SELF_MOVER <b>collocate type</b> : human (less prototypically: four-legged animal, such as a horse, a deer, a dog) <b>colligation</b> : <i>stagger</i> + PP or AVP of DIRECTION, PATH, SOURCE, GOAL, AREA, EXTERNAL_CAUSE <b>collocation</b> : <i>stagger</i> + NP of DISTANCE <b>collocation</b> : <i>stagger to one's feet</i> <b>semantic prosody</b> : it implies that the lack of balance is due to being drunk, ill or under a great weight
1b	[Self_motion] <sub>figurative</sub>	SELF_MOVER <b>collocate type</b> : 1. human (acting, not moving), 2. business enterprise, institution (e.g. <i>company, colony, sports team, marriage, bill, economy</i> ) <b>colligation</b> : <i>stagger</i> + figurative PP or AVP of DIRECTION, PATH, SOURCE, GOAL, AREA, EXTERNAL_CAUSE - <i>stagger on</i> - <i>stagger through</i> + time period (TIME IS STATIONARY AND WE MOVE THROUGH IT) - <i>stagger from... to...</i> + unpleasant situation - <i>stagger under</i> + <i>weight/ burden/ load/ strains/ debt</i> (the [+ heavy] NP used literally in LU1a is used here metaphorically to indicate obstacles to actions) <b>semantic prosody</b> : it implies continuation in difficult circumstances
2a	[Cause_motion]	AGENT <b>collocate types</b> : 1. human, 2. event (e.g. <i>collision, blow, slap</i> ) THEME <b>collocate type</b> : human
2b	[Stimulate_emotion]	STIMULUS <b>collocate type</b> : event EXPERIENCER <b>collocate type</b> : human <b>semantic prosody</b> : it implies an unexpected or unusual happening
2c	[Arranging]	AGENT <b>collocate type</b> : human THEME <b>collocate type</b> : 1. artifact, 2. activity/ event CONFIGURATION <b>colligation</b> : PP- <i>over</i> (denoting space or time respectively)

### 3.3 Definitions

The wording of the proposed definitions for the *stagger* LUs is presented in Table 4. In devising definitions, we have a twofold aim. On the one hand, we try to reveal the interrelationship between the senses (outlined in Table 2); consider, in this respect, the repetition of “unsteadily” in 1a and 2a, the repetition of “cause” in 2a, 2b and 2c, as well as the lumping of “spatial” and “temporal” arrangement in 2c. On the other hand, we try to reflect implications revealed in the wider context of corpus examples, such as the semantic prosody in 1a and 1b (outlined in Table 3). At the same time, we need to use a defining vocabulary that intermediate-level language learners can understand.

Table 4: Devising definitions.

LU	Definition
<b>1a</b>	walk or move unsteadily as if you are going to fall over (e.g. because of being drunk, ill, or under a weight)
<b>1b</b>	continue or carry on with great difficulty
<b>2a</b>	cause someone to lose their balance and walk unsteadily
<b>2b</b>	cause someone to feel surprised/ shocked
<b>2c</b>	cause things or events to be at different levels in space or time

## 4 The *Stagger* Entry in MLDs

The aim of this section is to briefly examine how the *stagger* senses and uses are treated in the “Big Five” MLDs (i.e. online editions of OALD, LDOCE, COBUILD, CALD, MEDAL). In particular, we are interested in the treatment of the features analyzed above, i.e. sense distinctions, usage patterns and definitions. Table 5 indicates whether the LUs identified above have been entered in the dictionary entries, under which sense they have been recorded and what usage patterns are used to illustrate them.

As the numbers in Table 5 indicate, all five *stagger* entries use a “flat” structure to present the meanings of the polysemous verb, as opposed to the “tiered” structure used in the analysis above. All entries generally cover the senses distinguished above, with the exception of the [Cause\_motion] LU, which is not recorded in any of the entries, most probably due to its relatively low frequency. They all record the core [Self\_motion] LU first, and most of them choose to record the [Stimulate\_emotion] LU second due to its high frequency. What is interesting to note is that there is considerable variation in the treatment of the [Self\_motion]<sub>figurative</sub> LU; entries seem to disagree as to whether it should be presented as a distinct sense (LDOCE, COBUILD, MEDAL) or as an example under the literal [Self\_motion] sense (OALD, CALD), and even those which choose the first option disagree as to the position of this sense (2<sup>nd</sup> in COBUILD vs. 3<sup>rd</sup> in LDOCE and MEDAL). Lastly, Table 5 indicates that most entries have two sense divisions corresponding to the [Arranging] LU; they either separate the two theme semantic types of the [Arranging] LU, i.e. artifacts and events (MEDAL), or assign the status of a distinct sense to the special event of the start of a race (LDOCE, CALD).

With regard to usage patterns, we do not expect to find great differences in coverage, since all dictionaries are corpus-informed. However, some variation may be observed with regard to the grammatical structure of the [Self\_motion] LU; while all entries record the “+ adverb or preposition” structure, only OALD records the transitive use, only COBUILD indicates the non-complement option, and LDOCE and CALD differ as to the “always” vs. “usually” specification of the “+ adverb or

preposition” structure.<sup>7</sup> Similarly, the entries seem to complement each other with regard to the lexicogrammatical patterns of the [Self\_motion]<sub>figurative</sub> and [Stimulate\_emotion] LUs. In contrast, they all record almost the same collocations for the [Arranging] LU; in this respect, it is worth noting that MEDAL’s treatment is very close to PDEV’s recording of both artifact and event collocate types, while LDOCE’s and CALD’s decision to particularly highlight the use of *stagger* in the context of a race is not verified by PDEV or our corpus analysis. Lastly, we should mention that the usage patterns entered in Table 5 have been retrieved from various parts of the entries, i.e. examples, definitions, highlighted figures.

Definitions have not been included in Table 5 due to space limitations; yet, we have examined whether the wording of the definitions in the five entries reflects (a) the interrelationship between the senses, and (b) implications/ associations revealed by corpora. Our conclusion is that whereas subtle aspects of meaning can be detected in some definitions (e.g. “If you stagger, you walk very unsteadily, for example because you are ill or drunk,” in COBUILD, “to continue doing something when you seem to be going to fail and you do not know what will happen” in LDOCE, “to cause someone to feel shocked or surprised because of something unexpected or very unusual happening” in CALD), no serious attempt is made to indicate the conceptual link between the senses of the polysemous headword. There are only two cases in which some traces of cognitive reasoning can be found, i.e. OALD’s and CALD’s use of the label “figurative” to specify [Self\_motion]<sub>figurative</sub> examples which are entered under the literal [Self\_motion] sense, and MEDAL’s similar wording in the following definitions: “to arrange for events or activities to start at different times” and “to arrange objects so that they are not at the same height or are not in a straight line”.

Table 5: Senses and uses in the *stagger* entry of the “Big Five” MLDs.

LUs	OALD	LDOCE	COBUILD	CALD	MEDAL
<b>1a</b> [Self_motion]	1 • something • + adv./prep. • <i>stagger to your feet</i>	1 always + adv./prep.	1 • verb • verb + adv./prep.	1 usually + adv./prep.	1 intransitive • <i>stagger backwards/ towards/ into/ out of</i> • <i>stagger to your feet</i>
<b>1b</b> [Self_motion] <sub>figurative</sub>	example under 1 <i>under the weight</i>	3 • <i>stagger on</i> • <i>from something to something</i>	2 <i>someone or something staggers on</i>	example under 1 <i>under a debt</i>	3 intransitive • <i>stagger on</i> • <i>stagger under debts</i>

<sup>7</sup> As Levin & Rappaport Hovav (1995: 197) note, through the addition of complements (particle, prepositional phrase, noun phrase) marking goals, English verbs of manner of motion are mapped onto the class of verbs of directed motion “in a completely productive way, and, therefore, the availability of the multiple meanings does not have to be listed in the lexical entry of any individual verb”. Therefore, we do not assign separate LUs to the atelic and the telic readings; however, this productive pattern should be clearly and systematically indicated in all relevant entries of a dictionary.

LUs	OALD	LDOCE	COBUILD	CALD	MEDAL
<b>2a</b> [Cause_motion]	-	-	-	-	-
<b>2b</b> [Stimulate_emotion]	2	2	3	2	2
	<ul style="list-style-type: none"> <li>• <i>stagger somebody</i></li> <li>• <i>it staggers somebody that</i></li> </ul>	-	<i>something staggers you</i>	-	transitive <i>be staggered by</i>
<b>2c</b> [Arranging]	3 (time)	4, 5	4 (time)	3, 4	4, 5
	<i>stagger something (events)</i>	<ul style="list-style-type: none"> <li>• <i>working hours, holidays</i></li> <li>• <i>race</i></li> </ul>	<i>holidays, hours of work</i>	<ul style="list-style-type: none"> <li>• <i>hours of work, holidays, events</i></li> <li>• <i>start of a race</i></li> </ul>	definition: <ul style="list-style-type: none"> <li>• events, activities</li> <li>• objects</li> </ul> example: <i>staggered working hours</i>

## 5 A New Version of the *Stagger* Entry

This section presents a new entry for the verb *stagger* compiled on the basis of the corpus-informed and cognitively-oriented analysis described in Section 3, while taking account of the comparative review of MLDs in Section 4.

The new entry, which is shown in Figure 1, has the following basic design characteristics:

- a signpost indicating the core semantic feature running through all uses (i.e. UNSTEADY, IRREGULAR MANNER)
- a tiered structure with two clusters of senses ordered in a logical manner (based on Table 2)
- definitions reflecting both the interrelationship between senses and the implications revealed by corpora (based on Table 4)
- corpus-attested examples (based on Table 1)
- tables indicating frequent usage patterns, the semantic types of core FEs, and/or their typical lexical realizations (based on Table 3).

The proposed entry looks quite different from the ones reviewed in Section 4; the differences are not so much due to its corpus-informed basis, but rather due to its cognitive orientation. For example, contrary to the other entries, which follow the frequency principle, we have decided to record the [Cause\_motion] LU, though relatively infrequent, because it motivates other senses. We apply the coherence principle to the whole entry by displaying the interrelationship between senses in their arrangement and definitions. Similarly, the signposts preceding sense divisions are edintend to help users realize how the senses of the verb are linked together. Lastly, indicating the semantic types of core FEs and recording usage patterns in distinct tables are features which can facilitate findability and usability.





<b>stagger</b> verb  		
UNSTEADY, IRREGULAR MANNER		
MOVE/ ACT		
<b>1a]</b> walk or move unsteadily as if you are going to fall over (e.g. because of being drunk, ill, or under a weight) <i>She <b>staggered</b>, and slumped on to the floor.</i> <i>He <b>staggered</b> 20 yards away before collapsing.</i> <i>There were a lot of young drunks <b>staggering</b> about.</i>		
human/ animal	<i>stagger</i>	<i>about/ back/ backwards/ away/ into home</i> <i>to one's feet</i> <i>from side to side</i> <i>under the weight of</i>
<b>1b]</b> continue or carry on with great difficulty <i>Both of them recovered, and <b>staggered</b> on through the year.</i> <i>The economy continued to <b>stagger</b> from crisis to crisis.</i>		
human business institution	<i>stagger</i>	<i>on</i> <i>through + time period</i> <i>from... to... + unpleasant situation</i> <i>under + weight/ burden/ load/ strain/ debt</i>
CAUSE MOTION/ EMOTION/ ARRANGEMENT		
<b>2]</b> cause someone to lose their balance and walk unsteadily <infrequent> <i>The blow <b>staggered</b> him sideways, his head smashing into the corridor.</i>		
<b>3]</b> cause someone to feel surprised/ shocked <i>The howl of dissent that came from the entire room <b>staggered</b> me.</i> <i>Rory was <b>staggered</b> by his answer.</i>		
event	<i>stagger</i>	human
human	<i>be staggered</i>	<i>by + event</i> <i>to find/ learn/ see</i>
<b>4]</b> cause things or events to be at different levels in space or time <i>The tables were <b>staggered</b> over different levels.</i> <i>The evacuation of 250,000 people was <b>staggered</b> over several days.</i> <i>I wish they'd <b>stagger</b> lecture finishing times because the corridors get packed.</i>		
human	<i>stagger</i>	event (the start/ starting dates/ closing times/ working hours/ holidays)
event	<i>be staggered</i>	over + time period

Figure 1: The proposed dictionary entry for *stagger*.

## 6 Conclusion

This study illustrates how cognitive linguistic theories can be systematically applied to corpus data in building up a dictionary entry for an English manner-of-motion verb.<sup>8</sup> In particular, we have demonstrated how an integrated methodology that draws on CPA, Frame Semantics, the Conceptual Metaphor and Metonymy Theory and the Principled Polysemy approach can facilitate the lexicographic tasks of establishing LUs, capturing usage patterns and providing definitions. The parts of the pre-lexicographic analysis have been used in compiling a new dictionary entry, which, if compared to the corresponding entry in MLDs, can be characterized as more reasonable and streamlined.

## References

- [BNC] British National Corpus. Accessed at: <https://www.sketchengine.co.uk/british-national-corpus> [30/03/2018].
- [CALD] *Cambridge Advanced Learner's Dictionary*. Accessed at: <https://dictionary.cambridge.org/dictionary/english> [30/03/2018].
- [COBUILD] *Collins COBUILD Advanced Learner's English Dictionary*. Accessed at: <https://www.collinsdictionary.com/dictionary/english> [30/03/2018].
- [LDOCE] *Longman Dictionary of Contemporary English*. Accessed at: <https://www.ldoceonline.com> [30/03/2018].
- [MEDAL] *Macmillan English Dictionary for Advanced Learners*. Accessed at: <https://www.macmillandictionary.com> [30/03/2018].
- [OALD] *Oxford Advanced Learner's Dictionary*. Accessed at: <https://www.oxfordlearnersdictionaries.com> [30/03/2018].
- [PDEV] *Pattern Dictionary of English Verbs*. Accessed at: <http://www.pdev.org.uk/#browse?q=&f=C> [30/03/2018].
- [ukWaC] British English Web Corpus. Accessed at: <https://www.sketchengine.co.uk/ukwac-corpus> [30/03/2018].
- Atkins, B. T. S. (2008). Then and now: Competence and performance in 35 years of lexicography. In T. Fontenelle (ed.) *Practical Lexicography: A Reader*. Oxford: Oxford University Press, pp. 247-272.
- Atkins, B. T. S., Fillmore, C. & Johnson, C. (2003). Lexicographic relevance: Selecting information from corpus evidence. In *International Journal of Lexicography*, 16(3), pp.: 251-280.
- Atkins, B. T. S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Atkins, B. T. S., Rundell, M. & Sato, H. (2003). The contribution of FrameNet to practical lexicography. In *International Journal of Lexicography*, 16(3), pp. 333-357.
- Croft, W. (2000). The role of domains in the interpretation of metaphors and metonymies. In B. Peeters (ed.) *The Lexicon-Encyclopedia Interface*. Oxford: Elsevier, pp. 219-256.
- Dalpanagioti, Th. (2012). Incorporating corpus data and semantic theory in Modern Greek lexicography: A special reference to the self-motion uses of *πετάω*. In Z. Gavrilidou, A. Efthymiou, E. Thomadaki & P. Kambakis-Vougiouklis (eds.) *Selected Papers of the 10th International Conference of Greek Linguistics, ICGL 10*, 1-4 September 2011. Komotini: Democritus University of Thrace, pp. 235-242.
- Dalpanagioti, Th. (2013). Frame-semantic issues in building a bilingual lexicographic resource: A case study of Greek and English motion verbs. In *Constructions and Frames*, 5(1), pp. 5-38.
- Evans, V. (2005). The meaning of *time*: Polysemy, the lexicon and conceptual structure. In *Journal of Linguistics* 41, pp. 33-75.
- Evans, V., Green M. (2006). *Cognitive Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.
- Ferraresi, A., Zanchetta, E., Bernardini, S. & Baroni, M. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgariff & S. Sharoff (eds.) *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?* 1 June 2008. Marrakech, Morocco. Accessed at: [http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wac4\\_2008.pdf](http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wac4_2008.pdf) [30/03/2018].
- Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing, pp. 111-137. Reprinted in D. Geeraerts (ed.) (2006) *Cognitive Linguistics: Basic Readings*. Berlin: Mouton de Gruyter, pp. 373-400.

<sup>8</sup> In fact, the data presented here is part of a large corpus-informed and cognitively-oriented pre-lexicographic database compiled for English and Greek polysemous manner-of-motion verbs (Dalpanagioti 2012; 2013).

- Fillmore, C. (1985). Frames and the semantics of understanding. In *Quaderni di Semantica* 6(2), pp. 222-254.
- Fillmore, C., Petruck, M. (2003). FrameNet glossary. In *International Journal of Lexicography* 16(3), pp. 359-361.
- FrameNet. Accessed at: <https://framenet.icsi.berkeley.edu/fndrupal> [30/03/2018].
- Geeraerts, D. (2007). Lexicography. In D. Geeraerts, H. Cuyckens (eds.) *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, pp. 1160-1174.
- Gries, St. Th., Stefanowitsch, A. (eds.) (2006). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter.
- Hanks, P. (1987). Definitions and explanations. In J. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins, pp. 116-136.
- Hanks, P. (2004a). The syntagmatics of metaphor and idiom. In *International Journal of Lexicography* 17(3), pp. 245-274.
- Hanks, P. (2004b). Corpus Pattern Analysis. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, 6-10 July 2004, Lorient, France*.
- Hanks, P. (2013a). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.
- Hanks, P. (2013b). English and American II: Synchronic lexicography. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: Mouton de Gruyter, pp. 720-730.
- Hanks, P. (2018). Corpus Pattern Analysis. CPA Project Page. Accessed at: <http://nlp.fi.muni.cz/projects/cpa> [30/03/2018].
- Johnson, M., Lenci A. (2013). Verbs of visual perception in Italian FrameNet. In M. Fried, K. Nikiforidou (eds.) *Advances in Frame Semantics*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 13-50.
- Kövecses, Z. (2002). *Metaphor. A Practical Introduction*. Oxford: Oxford University Press
- Kövecses, Z., Radden, G. (1998). Metonymy: Developing a cognitive linguistic view. In *Cognitive Linguistics* 9(1), pp. 37-77.
- Lakoff, G., M. Johnson (1980). *Metaphors We Live By*. Chicago and London: The University of Chicago Press.
- Levin, B., M. Rappaport Hovav (1995). *Unaccusativity: At the Syntax Lexical-Semantics Interface*. Cambridge, MA, London: MIT Press.
- Meer, G. van der. (1999). Metaphors and dictionaries: The morass of meaning, or how to get two ideas for one. In *International Journal of Lexicography* 12(3), pp. 195-208.
- Moon, R. (1987). The analysis of meaning. In J. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins, pp. 86-103.
- Moon, R. (2004). On specifying metaphor: An idea and its implementation. In *International Journal of Lexicography* 17(2), pp. 195-220.
- Niemeier, S. (2003). Straight from the heart – metonymic and metaphorical explorations. In A. Barcelona (ed.) *Metaphor and Metonymy at the Crossroads. A Cognitive Perspective*. Berlin: Mouton de Gruyter, pp. 195-213.
- Ostermann, C. (2015). *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*. Lexicographica. Series Maior 149. Berlin: Mouton de Gruyter.
- Radden, G. (2003). How metonymic are metaphors? In A. Barcelona (ed.) *Metaphor and Metonymy at the Crossroads. A Cognitive Perspective*. Berlin: Mouton de Gruyter, pp. 93-108.
- Rundell, M. (1998). Recent trends in English pedagogical lexicography. In *International Journal of Lexicography* 11(4), pp. 315-342.
- Ruppenhofer, J., Ellsworth, M., Petruck M., Johnson C. & J. Scheffczyk. (2016). *FrameNet II: Extended Theory and Practice*. Accessed at: <https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf> [30/03/2018].
- Sinclair, J. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.
- Sinclair, J. (1998). The lexical item. In E. Weigand (ed.) *Contrastive Lexical Semantics*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 1-24.
- Tognini-Bonelli, E. (2002). Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach. In B. Altenberg & S. Granger (eds.) *Lexis in Contrast: Corpus-based Approaches*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 73-95.

# Polysemy and Sense Extension in Bilingual Lexicography

**Janet DeCesaris**

*Institute for Applied Linguistics, Universitat Pompeu Fabra*

*E-mail: janet.decesaris@upf.edu*

## Abstract

Polysemy often poses problems for the dictionary representation of word meaning, because the discrimination of senses is seldom clear-cut. In the past twenty years, corpus linguists, notably Kilgariff (1997) and Hanks (2000), have argued that the concepts of “word sense” and “word meaning” are problematic to the extent that they invite a “checklist” view of meaning that is not borne out by corpus evidence, although precisely that “checklist” view is encouraged by dictionary representation in that dictionaries describe meaning as discrete items in lists (Fontanelle 2016). The challenges associated with representing polysemy are particularly acute for bilingual dictionaries, because patterns of polysemy associated with cross-linguistic equivalents display differing degrees of what has been called “overlapping polysemy” (Alsina & DeCesaris 2002; Boas 2009). This paper considers the treatment in bilingual dictionaries of two small sets of words in English and their equivalents in Spanish, French and Italian which display varying degrees of overlapping polysemy. We suggest ways of incorporating sense extension and partial parallelisms into bilingual dictionary entries, specifically by subdividing senses according to the semantic types as found in corpora.

**Keywords:** bilingual lexicography, overlapping polysemy, sense extension

## 1 Introduction

The representation of polysemy has been discussed by many, particularly in the context of monolingual dictionaries and the two opposing approaches informally known as “lumping” or “splitting” (Landau 2004). Polysemy often poses problems for the dictionary representation of word meaning, because the discrimination of senses is seldom clear-cut. Leading figures in corpus lexicography, such as Hanks (2000) and Kilgariff (1997), have noted that the presentation of word meaning in lists with a hierarchical structure that is fundamental to traditional dictionary representation is actually not borne out by the evidence: the notions of “word sense” and “word meaning” without further context are problematic. Yet, as Fontanelle (2016) cogently points out, dictionary representation as we know it today encourages a “checklist” view of word meaning, because meaning is described as discrete items on lists. In this context, the issue of how to deal with the intricacies of word meaning across two languages presents an especially difficult challenge. A situation which frequently arises in the comparison of two languages is that of partial correspondence of sense extension. In this paper, I discuss the dictionary representation of a few cases of partial correspondence or “overlapping polysemy” between English and three Romance languages, Spanish, French and Italian, in light of corpus evidence. This paper is structured as follows. After a brief introduction to the nature of polysemy and its representation in dictionaries, two small sets of words in English that are known to exhibit varying degrees of sense extension are discussed. The entries of these words in several large bilingual dictionaries in the three aforementioned language combinations are discussed, and the information found in the dictionaries is compared with that from corpus analysis. The paper concludes by suggesting how bilingual dictionaries could improve their representation of overlapping polysemy.



## 2 Polysemy

Polysemy may be defined as “having or characterized by many meanings”.<sup>1</sup> As has often been noted (Landau 2004), the multiplicity of meaning associated with a single written form is usually represented in a dictionary by means of a single headword encompassing several senses which are typically displayed as a list. However, exactly what constitutes multiplicity of meaning (polysemy) as opposed to independent meanings the forms of which converge (homonymy), is an issue that is open to debate. Before discussing the dictionary representation of polysemy, it is worthwhile noting that in the fields of philosophy of language and theoretical semantics, polysemy is an oft-treated topic. Sennet (2016) offers a convenient discussion of the importance of polysemy in the philosophy of language, and Falkum and Vicente’s (2015) introduction to a specially themed issue of *Lingua* on approaches to polysemy in semantics brings together different theoretical views. Nonetheless, for those of us who work with dictionaries, which in a very direct way must deal with senses and words, it is rather disheartening to see that most discussion of polysemy in lexicographic research has gone unnoticed by these two research communities, and that which has been noticed is sometimes met with a cavalier attitude (e.g. Sennet’s comment “And in general, taxonomy based on the relatedness of distinct meanings is a pretty dull affair for anyone but the committed lexicographer” (2016)).

### 2.1 Polysemy and Monolingual Dictionaries

In many monolingual dictionaries, etymology is a determining factor in the representation of polysemy: if senses are known to have been derived from a single source, the senses are included under a single headword, even though the current meanings of the historically related senses may seem unrelated to the modern eye. Let us look at an example. Some current monolingual dictionaries of English, such as *The American Heritage Dictionary of the English Language* (see Figure 1<sup>2</sup>), *Merriam-Webster’s Unabridged Dictionary*, and the *Macmillan Dictionary*, assign both the sense of the word *crane* referring to a kind of bird with long legs and a long neck, and the sense referring to machinery used to move heavy objects, to the same headword, because the sense referring to a kind of machinery is historically the result of sense extension from that referring to a kind of bird.

The *Oxford Dictionary of English*, in contrast, assigns those two same senses to two different headwords, *crane*<sup>1</sup> (the machinery) and *crane*<sup>2</sup> (the bird), as seen in Figure 2.

Grammar sometimes plays an important role in lexicographers’ representation of polysemy (or homonymy): the English verb *crane*<sub>[verb]</sub> is listed as a separate word in the *Macmillan Dictionary* and in *Merriam-Webster’s Unabridged Dictionary*, presumably because the morphological forms and syntactic behavior of *crane*<sub>[verb]</sub> are different from those of *crane*<sub>[noun]</sub>. The *Oxford Dictionary of English* provides a different treatment, placing the verb *crane* under *crane*<sup>1</sup> (i.e. along with the sense referring to machinery used to move heavy objects, as seen in Figure 2), and the *American Heritage Dictionary of the English Language* gives all senses of both the noun and the verb under a single headword (Figure 1). This brief example points up the different approaches that monolingual dictionaries can take to representing the complex relationships existing across related senses and words. The task facing bilingual dictionaries is arguably even more daunting, given the fact that patterns of polysemy associated with cross-linguistic equivalents coincide only partially.

1 Definition taken from the *American Heritage Dictionary of the English Language*, 5<sup>th</sup> edition.

2 The entry in the dictionary also includes the etymology and two photos, one of a bird and one of heavy machinery, which I have not included here.



**crane** (krān)*n.*

1.

**a.** Any of various large wading birds of the family Gruidae, having a long neck, long legs, and a long bill.

**b.** A similar bird, such as a heron.

**2.** A machine for hoisting and moving heavy objects by means of cables attached to a movable boom.

**3.** Any of various devices with a swinging arm, as in a fireplace for suspending a pot.

*v. craned, cran·ing, cranes**v.tr.*

**1.** To hoist or move with or as if with a crane.

**2.** To strain and stretch (the neck, for example) in order to see better.

*v.intr.*

**1.** To stretch one's neck toward something for a better view.

**2.** To be irresolute; hesitate.

Figure 1: Entry for *crane*, *The American Heritage Dictionary of the English Language*, Fifth Edition.

**crane**<sup>1</sup>

► *noun* a large, tall machine used for moving heavy objects by suspending them from a projecting arm or beam: *a dockside crane* | [as modifier] : *a crane driver*.

■ a moving platform supporting a television or film camera: *a very long tracking shot done with dolly and crane* | [as modifier] : *the opening crane shot*.

*verb*

**1.** [no object, with adverbial of direction] stretch out one's body or neck in order to see something: *she craned forward to look more clearly*.

■ [with object] stretch out (one's neck) so as to see something: *she craned her neck to see past me*.

**2.** [with object and adverbial] move (a heavy object) with a crane: *the wheelhouse module is craned into position on the hull*.

**crane**<sup>2</sup>

► *noun* a tall, long-legged, long-necked bird, typically with white or grey plumage and often with tail plumes and patches of bare red skin on the head. Cranes are noted for their elaborate courtship dances.

● Family *Gruidae*: four genera, in particular *Grus*, and several species, including the Eurasian **common crane** (*G. grus*).

Figure 2: Entries for *crane*<sup>1</sup> and *crane*<sup>2</sup>, *Oxford Dictionary of English*.

## 2.2 Overlapping Polysemy

Overlapping polysemy occurs when cross-linguistic equivalents share a literal sense and some, but not all, extended senses, and may be described as a situation when the sense extension or extensions of one word belonging to one language correspond in part to those of an equivalent word in another language. The polysemy of a word rarely coincides with that of a word in another language because the derived senses associated with a given word are the result of a combination of internal and external forces in a specific language, and thus the potential for differences across languages is very high. Partial correspondence of patterns of polysemy, however, is not rare because many languages share metaphors. In addition, and specifically in the cases under consideration in this paper, overlapping polysemy is widespread in languages sharing etymological sources and sociocultural development. Certainly, given the contact between English and vocabulary originally from Latin over centuries, some overlapping polysemy between English on the one hand, and Spanish, French, and Italian on the other, is to be expected.

In the context of bilingual dictionaries, correctly recognizing and representing the complexity and varying degrees of overlapping polysemy is particularly important to the user if the dictionary is being consulted to produce text in a non-native language. In terms of language comprehension, it is quite possible that understanding cases of overlapping polysemy unknown to the non-native speaker at the time may not be problematic at all: for example, if a native speaker of English reads the Spanish phrase “*sembró el pánico entre la población*” (‘it spread/caused panic among the population’), decoding “*sembró el pánico*” (literally, ‘sowed panic’) is not difficult even if the speaker is unaware that *sembrar* can take direct objects other than types of seeds or fields because the equivalent of *sembrar*, the English verb *sow*, is frequently used in conjunction with nouns such as *confusion*, *terror*, *dissension*, and *distrust*, and thus the sense extension displayed in “*sembró el pánico*” is quite similar to that displayed in English. As Taylor (2003) noted, speakers are rarely troubled by polysemy (thus resulting in what Taylor famously called “polysemy’s paradox”). Production tasks in a non-native language, in contrast to comprehension tasks (at least for some speakers), can be quite troublesome if speakers simply assume that sense extension in the non-native language parallels that of the native language. To continue with data from English and Spanish, in English one might be tempted to translate the phrase “a flood of complaints” as “*una inundación de quejas*” when in fact the Spanish noun *inundación* is rarely used with an extended sense; a better translation would be “*una avalancha de quejas*” (literally, ‘an avalanche of complaints’) or “*una oleada de quejas*” (literally, ‘a huge wave of complaints’). To the extent that the extended sense of the expression in English (“flood of” meaning ‘a huge amount of X, often appearing without warning’) is standard and frequent enough to warrant inclusion in a monolingual dictionary, it is a candidate for inclusion in a comprehensive bilingual dictionary; in this case, the dictionary would need to show that *inundación* is not a good equivalent for all senses of flood<sub>[noun]</sub>.

## 3 Description of the Study

### 3.1 Words Studied

In order to study the representation of overlapping polysemy in several English-Spanish, English-French and English-Italian bilingual dictionaries, two small sets of words which display varying degrees of overlapping polysemy were considered. One set of words (*avalanche*, *flood*, *mountain*, *storm*, and *stream*) consists of nouns the literal sense of which refers to a natural phenomenon; these nouns commonly display an extended sense when they are the heads of a prepositional phrase (e.g. *an avalanche of publicity*, *a storm of protest*, *a stream of visitors*). The choice of nouns referring to natural phenomena also allowed us to assume that, for the purposes of this study, the literal meaning of the

noun was essentially the same in all the languages under consideration in this paper. This assumption cannot be made for all semantic classes of nouns; for example, abstract nouns (e.g. *privacy*, *friendship*, *freedom*), the meaning of which often involves a significant cultural component, would not allow clear comparison of extended senses. The other set of words studied consists of verbs (*cultivate*, *fabricate*, *forge*, *plough/plow*, *sow*) that are used in conjunction with direct objects belonging to very different semantic classes (e.g. *cultivate the land* vs. *cultivate the arts*; *forge iron* vs. *forge a career*; *sow seeds* vs. *sow hatred*). Such obvious differentiation of semantic classes of direct objects can be detected through corpus analysis, specifically by using the Word Sketch function in Sketch Engine®, which allows one to quickly identify nouns in the direct object position. This sort of display of information is desirable, because the grouping of direct objects into semantic classes is a possible way of identifying differences in equivalents, and thus a way of organizing entries in a bilingual dictionary.

### 3.2 Dictionaries Consulted

In order to compare the representation of overlapping polysemy in the dictionaries, the degree of sense extension of the words in English as represented in *The American Heritage Dictionary of the English Language* and shown by the Word Sketch feature of Sketch Engine while consulting the English Web 2015 (enTenTen15) corpus was examined. The following online bilingual dictionaries were then consulted. To the extent possible, using online bilingual dictionaries without pay walls was the goal; however, this was not always possible, and some dictionaries with pay walls were consulted and this information is noted.

- English-Spanish:  
*Collins Spanish Dictionary*; *Oxford Spanish Dictionary* online; English-Spanish combination on [www.diccionarios.com](http://www.diccionarios.com) (subscription required);
- English-French:  
*Collins French Dictionary*; English-French combination on [www.diccionarios.com](http://www.diccionarios.com) (includes Larousse dictionaries; subscription required);
- English-Italian:  
*Collins Italian Dictionary*; *Grande Dizionario Hoepli Inglese* by F. Picchi (available online at *La Repubblica* newspaper website).

## 4 Analysis and Discussion

The first observation to note about the nouns chosen for this study ((*avalanche*, *flood*, *mountain*, *storm*, and *stream*) is that they do not all behave in the same way with respect to sense extension. *Storm* and *stream* have a wide variety of complements (e.g. *storm of XX*, *stream of XX*), whereas *mountain*, once geographic references are removed (e.g. *mountains of California*; *mountains of the Alps*), has fewer possible figurative complements. *Mountain* is used both in the singular and plural in its extended sense (e.g. *mountain of evidence*, *mountains of paperwork*). Of the top 25 possible complements of “*avalanche of*”, only five (*mud*, *rubble*, *ash*, *snow*, and *dust*) are related to the literal sense; the remaining 20 (e.g. *criticism*, *publicity*, *lawsuit*, *propaganda*) are all related to the extended sense. Of the 25 most frequent complements of the expression “*flood of*”, only one (“*flood of lava*”) is a manifestation of the literal sense. This surely says something about the meaning of *flood*, which must be of water and thus the repetition of a complement referring to water would be superfluous; *avalanche*, on the other hand, while prototypically involving snow, can also be applied to other substances, and thus can occur with a prepositional phrase with its non-extended sense.

The verbs chosen yield equally interesting observations. The verb *fabricate* has two clear meanings, one referring to manufacturing and the other to producing falsified statements, and both are widely attested in current usage. In English, one can *sow seeds, wheat, oats* or other grains, fields or other surface areas, and in its extended sense a wide variety of nouns with negative resonance (of the top 25 direct object nouns of *sow* according to the Word Sketch, 12 were nouns belonging to that category and all were uncountable: *confusion, terror, chaos, dissension, distrust, fear, strife, hatred, panic, division, doubt, mistrust*). The verb *forge* is far more frequent in its extended sense of *forging an allegiance* or *friendship* than in its original sense of *forging steel*, but this verb has a seemingly unrelated sense of falsifying documents (*forge passports* or *documents*). The dictionaries all register this latter sense, of course, but interestingly do not label it as a figurative sense (possibly because *forged documents* exist and are not imagined). The extended sense of *plough/plow* occurs with the preposition *through*: one frequently *ploughs through books*, but one can also *plough through snow*.

Let us now look at a few specific examples: the representations of *stream* (Figures 3, 4, 5, and 6) and *sow* (Figures 7, 8, 9, and 10) in each combination of languages.

1. (= *brook*) **arroyo m ♦ riachuelo m**

2. (= *current*) **corriente f**

to go with/against the stream (*literal, figurative*) **ir con/contra la corriente**

3. (= *jet, gush*)

[*of liquid*] **chorro m**

[*of light*] **raudal m**

[*of air*] **chorro m ♦ corriente f**

[*of lava*] **río m**

[*of insults, abuse*] **sarta f**

[*of letters, questions, complaints*] **lluvia f**

a thin stream of water **un chorrito de agua**

she exhaled a thin stream of smoke **lanzó or exhaló un chorrillo de humo**

a steady stream of cars **un flujo constante or ininterrumpido de coches**

people were coming out of the cinema in a steady stream **había una continua hilera de gente que iba saliendo del cine**

we had a constant stream of visitors **recibíamos visitas continuamente or sin parar**

he let out a stream of insults **soltó una sarta de insultos**

stream of consciousness **monólogo m interior**

4. (*British*) (*Education*) **grupo de alumnos de la misma edad y aptitud académica**

the top/middle/bottom stream **la clase de nivel superior/medio/inferior**

5. (*Industry*)

to be on/off stream [*machinery, production line*] **estar/no estar en funcionamiento; [oil well] estar/no estar en producción**

to come on stream [*machinery, production line*] **entrar en funcionamiento; [oil well] entrar en producción**

Figure 3: *stream* in the Collins Spanish Dictionary

1

1.1 (small river)

arroyo (masculine)riachuelo (masculine)

1.2 (current)

corriente (feminine)2 (flow) *a thin stream of water issued from the fountain***un chorrito de agua salía de la fuente***a stream of lava***un río de lava***a stream of sunlight entered the room***el sol entró a raudales en la habitación***she poured out a stream of abuse at him***le soltó una sarta de insultos***the affair generated a stream of books and articles***el caso generó un torrente de libros y artículos***there is a continuous stream of traffic***pasan vehículos continuamente****el tráfico es ininterrumpido***streams of people were coming out of the theater***un torrente de personas salía del teatro**

3 (British) (School)

*(conjunto de alumnos agrupados según su nivel de aptitud para una asignatura)*

Figure 4: Stream in the Oxford Spanish Dictionary

1. (= brook) **ruisseau m**2. (= current) **courant m**3. (= continuous flow) [of smoke, air, liquid] **flot m**4. (= moving line) [of people, vehicles] **flot m**5. (= large number) [of letters, jokes, complaints, visitors] **flot m**6. (Britain) (SCHOOL) **niveau m**

7. (INDUSTRY)

to be on stream [new power plant, computer system] **être en service**to come on stream [new power plant, computer system] **être mis en service**

transitive verb

(Britain) (SCHOOL) [PUPILS, CLASSES] **répartir par niveau**

Figure 5: Stream in the Collins French Dictionary



**1** (= *brook*) torrente, ruscello, fiumiciattolo, corso d'acqua

◇ **stream bed** alveo o letto di torrente; **a mountain stream** un ruscello di montagna; **as fresh as a mountain stream** fresco come un torrente di montagna; **to go up/down stream** idiomaticamente andare a monte/a valle di un corso d'acqua; **the boy leapt over the stream** il ragazzo superò il torrente con un salto; **a lovely stream flows through the park** un grazioso ruscello attraversa il parco

**2.1** (= *current*) corrente (*di liquido o gas*)

◇ **the wind stream quickly spread the radioactivity** la corrente del vento propagò rapidamente la radioattività

**2.2** mar corrente, filo di corrente/di marea

◇ **stream-anchor** ancora di tonnellaggio; ancora di corrente; **stream cable/chain** catena dell'ancora di tonnellaggio/di corrente; **stream tide** marea delle sizigie, marea sizigiale, grande marea; **tidal stream** corrente di marea; **in the stream** (*di nave*) ancorato al largo; **to swim with/against the stream** idiomaticamente nuotare a favore di/contro la corrente

**3** +of (= *trickle*) flusso, rivolo, rivoletto

**a stream of blood was flowing from his nose** dal naso gli colava un rivoletto di sangue

**4** (*di persone, auto, ecc*) flusso, corrente, afflusso, deflusso, fila, massa

◇ **an endless stream of tourists** un afflusso continuo di turisti; **a steady stream of traffic** un flusso costante di traffico; **to come/to be brought on stream** figurativamente entrare in funzione; **the assets built up by Japan are generating a stream of income** le attività accumulate dal Giappone generano un flusso di reddito

**5** (= *flood*) figurativamente fiumana, marea, massa, fiotto, flusso

◇ **a stream of abuse/insults** una marea di insulti; **a steady stream of phone calls** un flusso costante di telefonate; **a stream of questions** una massa di domande; **streams of immigrants in search of employment arrived in the country** arrivarono nel paese masse di immigrati in cerca di occupazione

**6** scol gruppo omogeneo di studenti formato in base alle loro capacità

◇ **the top stream** il gruppo più bravo.

Figure 6: *Stream* in the *Grande Dizionario Hoepli Inglese*

Interestingly, the bilingual dictionaries do not structure their entries in the same ways, but the differences do not appear to be a result of differences in contrastive analysis. All the dictionaries indicate that *stream* corresponds to the notions of “brook” and “current”, but these two notions are listed as subsenses of a single sense in the *Oxford Spanish Dictionary* whereas all the others separate them into two distinct senses. Presumably, in the Oxford dictionary the accompanying idea of “water” is not why the two senses are lumped together, because “water” is an integral part of the sense in this dictionary 2. The *Collins Spanish Dictionary* groups together several complements under the meaning indicator “jet, gush” which concentrates on the manner the substance moves, but because Spanish does not typically have a manner component entailed in the meaning of its verbs, the resulting entry has a wide range of unrelated complements mixing the literal and extended senses (*stream of light*,

*stream of abuse*, *stream of visitors*, and *thin stream of water* are all in sense 3, and *light*, *abuse*, *visitors*, and *water* do not belong to a single semantic class). The *Oxford Spanish Dictionary* also groups together unrelated complements under the meaning indicator “flow”, but unlike the *Collins Spanish Dictionary* does not identify any equivalent; rather, sense 2 is a listing of examples. Although the translations provided are appropriate, the fact that the classification under “flow” yields no equivalent indicates that the sense distinction based solely on the source language is not suitable for a bilingual dictionary. The Italian dictionary provides a large number of equivalents, many of which are not used in examples and thus are of less use to native English speakers, who may lack enough knowledge of Italian to use them properly in a language encoding task (which, of course, is why an English speaker would go to an English to Italian dictionary). It is difficult to see the differences between some of the examples given in sense 4 (e.g. *an endless stream of tourists*) and those in sense 5 (e.g. *streams of immigrants*).

By looking at the data from Sketch Engine, the complements of *stream* may be grouped into the following ten semantic types: LIQUID, AIR, SMOKE, LIGHT, PARTICLES, PEOPLE, SCHOOLS OF THOUGHT OR BELIEF, DOCUMENTS, INSULTS AND ABUSE, and REVENUE. None of the dictionaries include any examples similar to a phrase like *streams of Judaism*, which had a logDice association figure of 6.91 with 435 occurrences, and only one (the Italian dictionary) has an example with *income*, which is one of the most frequent collocates of *stream* (a logDice association of 8.17; another noun belonging to that semantic class, *revenue*, is not in any of the entries). If the dictionaries were to structure their senses around the semantic grouping of collocates, the resulting entries would be longer, and arguably more complex, because the semantic groupings correspond to a more fine-grained analysis than the current linguistic analysis behind the dictionary entries: note that none of the dictionaries indicate that *stream* combines primarily with ten types of entities. Today’s corpus tools give us the ability to see this distributional data, and it should be incorporated into dictionary entries. In addition, structuring entries around the notion of non-derived vs. derived senses and further centering these around groupings of collocates would aid in avoiding entries consisting of only lists of examples, with no equivalents.

We turn now to *sow*.

[seed] sembrar

to sow doubt in sb’s mind sembrar dudas en algn

to sow mines in a strait, sow a strait with mines sembrar un estrecho de minas ♦ colocar minas en un estrecho

Figure 7: Sow in the Collins Spanish Dictionary

1 (plant)

(seeds/barley/field)

sembrar

to sow a field with wheat

**sembrar un campo de trigo**

2 (mines)

plantar

poner

to sow a field with mines

**sembrar un campo de minas**

Figure 8: Sow in the Oxford Spanish Dictionary Online

transitive verb

1. [seed, field] **semer**

2. (= spread) [doubts, confusion, dissension] **semer**

Figure 9: *Sow* in the *Collins French Dictionary*

A vti

V(+D) seminare, fare la semina, spargere

◇ **this is the best time to sow** questo è il momento migliore per seminare

B vt

1 V+D+with/in/on+IN piantare, seminare, sementare

◇ **to sow the land** seminare la terra; **to sow seeds in pots/in the open ground** piantare semi in vasi/nel terreno aperto; **the land was sown with corn** la terra fu seminata a grano

2 V+D(+in+IN) (= *to stir up*) fig seminare, diffondere, suscitare, provocare, promuovere

◇ **to sow doubts in sb's mind** mettere dubbi in mente a qn; **to sow the seeds of** creare le premesse di/per; **to sow the seeds of discontent** gettare il seme della discordia; **to sow unrest** diffondere malcontento.

Figure 10: *Sow* in the *Grande Dizionario Hoepli Inglese*

In English, as mentioned above, *sow* can combine either with seeds or grains that can be planted, or with the area where such planting is to take place. The strong resonance of “one reaps what one sows” is linked to the Bible, and is most often used in contexts in which what was sown (figuratively) is not good, and as a result the consequences that must be faced are equally negative. This negative resonance is clearly seen in the corpus data, as *sow* combines with abstract nouns with negative resonance. The Spanish dictionaries do not show this clearly, yet surely this is information that belongs in such works. The Italian dictionary, in contrast, provides a meaning indicator in the form of a paraphrase in the source language (*to stir up*) which effectively conveys the idea, and the several examples provided all go to reinforce the verb’s negative resonance. Even though its entry is very short, the French dictionary does a reasonable job of explaining the extended sense of *sow* and the meaning indicators of what can be *sown* (*doubts, confusion, dissension*) suggest negative resonance to the reader. On the whole, the dictionaries’ representation of *sow* is more in line with the corpus data available than was the case for *stream*, although the entries could be improved by clearly stating the extended sense of *sow* takes complements with negative resonance.

## 5 Conclusion

Overall, the bilingual dictionaries consulted all represent the extended senses of the words studied to one degree or another. What they fail to do in some cases is provide a more consistent representation of the extended use: on the one hand, several equivalents are given, yet we found numerous examples of extended senses in the dictionaries studied that did not include the equivalent provided. This would not be difficult to improve. If *colonna* is the equivalent for *stream* in the context of cars, which is what the *Collins Italian Dictionary* states, it is a mystery why the dictionary gives an example referring to cars that does not use *colonna* (*un fiume ininterrotto di machine* is listed). The same practice may be seen in sense (2) of the *Oxford Spanish Dictionary*’s entry for *sow*: the dictionary lists two equivalents, *plantar* and *poner*, yet the example given is with *sembrar*.

If lexicographers ask users to plough through long entries, with senses, subsenses and numerous examples, then using meaning indicators to group together equivalents is a welcome strategy, because adding visual and semantic structure to the representation of equivalents should aid consultation. A recurring problem in representing sense extension and overlapping polysemy is the presence of long lists of translated phrases. It is not clear that the mere presence of translated examples is useful to many users, who may not be willing to read them. Those phrases could be grouped according to semantic criteria and a brief explanation could be provided so that the user is pointed in the right direction. In essence, this is what the *Grande Dizionario Hoepli Inglese* does by providing a paraphrase (as in *sow = to stir up*). Of the dictionaries consulted in this study this one has the best treatment of sense extension and overlapping polysemy. It employs a strategy of “splitting” as opposed to “lumping” more than the other dictionaries, and as a result the display of equivalents is generally clearer. It also tends to have the longest entries. That is not a coincidence, as there are no shortcuts in representing the complex interplay of word senses across languages.

## References

- Alsina, V. & DeCesaris, J. (2002). Bilingual lexicography, overlapping polysemy, and corpus use. In B. Altenberg, S. Granger (eds.) *Lexis in Contrast: Corpus-based Approaches*. Amsterdam/Philadelphia: John Benjamins, pp. 215–230.
- American Heritage Dictionary of the English Language*. (2012, Fifth edition). Boston: Houghton Mifflin Harcourt.
- Boas, H. C. (ed.) (2009). *Multilingual FrameNets in Computational Lexicography*. Berlin/New York: Mouton de Gruyter.
- Collins French Dictionary*. Accessed at: <https://www.collinsdictionary.com/dictionary/english-french> [15/02/2018].
- Collins Italian Dictionary*. Accessed at: <https://www.collinsdictionary.com/dictionary/english-italian> [15/02/2018].
- Collins Spanish Dictionary*. Accessed at: <https://www.collinsdictionary.com/dictionary/english-spanish> [15/02/2018].
- Diccionarios.com*. Accessed at: <https://www.diccionarios.com/> [18/02/2018].
- Falkum, I. L. & Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua* 157, pp. 1-16.
- Fontanelle, T. (2016). Bilingual Dictionaries. History and Development; Current Issues. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 44-61.
- Hanks, P. (2000). Do Word Meanings Exist? *Computers and the Humanities* 34, pp. 205-15.
- Kilgariff, A. (1997). I Don't Believe in Word Senses. *Computers and the Humanities* 31, pp. 91-113.
- Landau, S. (2004, Second edition). *Dictionaries: The Art and Craft of Lexicography*. New York: Cambridge University Press.
- Macmillan English Dictionary Online*. Accessed at: <http://www.macmillandictionary.com> [20/02/2018].
- Merriam-Webster Unabridged*. Accessed at: <https://unabridged.merriam-webster.com> [25/03/2018].
- Oxford Dictionary of English*. Third edition, online version 2015. Accessed at: [www.oxfordreference.com](http://www.oxfordreference.com) [24/03/2018].
- Oxford Spanish Dictionary*. Accessed at: <https://premium.oxforddictionaries.com/spanish/> [22/01/2018].
- Picchi, F. *Grande Dizionario Hoepli Inglese*. Accessed at: <http://dizionari.repubblica.it/inglese.php> [18/02/2018].
- Sennet, A. (2016). Polysemy. *Oxford Handbooks Online* (Subject: Philosophy, Philosophy of Language; DOI:10.1093/oxfordhb/9780199935314.013.32). Accessed at: <http://www.oxfordhandbooks.com> [27/02/2018].
- Sketch Engine Corpus Management. Accessed at: <https://www.sketchengine.co.uk/> [27/03/2018, last access].
- Taylor, J. R. (2003). Polysemy's Paradoxes. *Language Sciences* 25, pp. 637-655.

## Acknowledgements

Research for this paper was carried out as a part of the project *Significado léxico dependiente del context y polisemia, y las implicaciones para la representación lexicográfica* (Ref. FFI2015-70375-P (DeCesaris, Principal researcher)) funded by the Spanish Ministry of Economy and Competitiveness, whose support is acknowledged.





# Associative Experiments as a Tool to Construct Dictionary Entries

**Ksenia S. Kardanova-Biryukova**

*Moscow City University*

*E-mail: kardanova81@yandex.ru*

## Abstract

Associative experiments have been used in psychology and psycholinguistics for over a hundred years and have proven to be an efficient tool in identifying those components of a cognitive structure which are relevant for a contemporary language user and arranging them into a hierarchy. We argue that the findings obtained through such an experimental approach can serve as the basis for compiling a dictionary entry that reflects the language in use now and does not lag behind by failing to depict changes in the semantic structure of a word or modifications in its use. To demonstrate how it works we have considered a rather complex notion of *empire* and its representation in the linguistic consciousness of Russian and American English native speakers. Relying on the findings of the associative experiment held with 148 Americans and 434 Russians we suggest ways of drafting dictionary entries that would reflect those semantic components which are relevant for language users given the current political and economic environment.

**Keywords:** associative experiment, modelling cognitive structures, stimulus and reaction

## 1 Introduction

Among the most challenging tasks that a lexicographer faces is a requirement to compile dictionary entries that stand up to the current language use rather than reflect the language in use 20 or so years ago. One of the ways to achieve this is to build dictionary entries upon the study of the relevant semantic components of the word meaning embedded in the linguistic consciousness of native speakers. Relying on this analysis as the foundation for drafting a dictionary entry can be very helpful, as the findings of such research can reflect even minor changes in the semantic structure of a word, in its distribution and use, and in the most typical associations with a given word.

From this perspective an associative experiment can serve as an efficient tool to unveil how the word meaning manifests itself in the public and individual consciousness. It is a way to uncover and prioritize those components of cognitive structures that are embedded in the language and enjoy diverse ways of verbal representation. It can be argued that these findings prove very helpful when constructing conventional dictionary entries, as they can unveil those uses of a word that are engraved in the public consciousness yet are not featured as a word meaning component. Some prominent Russian psycholinguists (A.A. Zalevskaya (2011; 2014), V.A. Pishchalnikova (2002) and others) consider *word meaning as a cognitive structure* to denote these cognitive components outside the frame of conventional word meaning.

It is common practice to systemize the findings of associative experiments in associative dictionaries with entries that feature an assortment of reactions by a representative group of subjects, varying from the most frequent ones to unique ones (see a typical associative dictionary entry structure in Fig. 1).

Stimulus word	Associations
<b>Power</b>	absolute, agency 2, authority 4, awful, careful, company, control 3, cord 2, corporation, country, dark, discipline, dominate, drive, electricity 7, empire, energy, Federal Credit Union, frightening, gender, George Bush, girl, government, grip, help, hour, house, hungry 2, in charge, justice 2, king, knowledge 3, lift, light 2, man, me 3, mighty, money 4, motor, Nietzsche, on 2, oppression, outrage 3, people 2, plant 4, play 3, point 5, political, politics, President 3, rangers 2, red 2, speed, station, status, steadfast, strained, strength 18, strong 4, structure, struggle 8, Superman, super, supply, surge 4, to the people, tool, trip 2, truth, ultimate, unfair, violent, weakness 4, work, yes <b>149+75+49+1</b>
<p>Legend:</p> <ul style="list-style-type: none"> <li>- the number next to associations shows the number of subjects who have provided this response (if there is no number, only one respondent provided this association);</li> <li>- the numbers at the bottom feature (1) the number of responses received, (2) the number of different associations, (3) the number of single associations, (4) the number of void responses.</li> </ul>	

Figure 1: Typical structure of an associative dictionary entry  
(the findings of the associative experiment with American subjects conducted by the author).

There are numerous examples of associative experiments conducted by Russian researchers that can serve as the basis for lexicographic work (Petrova 2008; Iashin 2009; Stepykin 2011; Panarina 2017, etc.). However, at present there are few papers published outside Russia that focus on the findings of associative experiments, with examples being Deese (1965), Martinek (2009), and Ufimtseva (2014), among others.

When a lexicographer is challenged with a task of drafting a dictionary entry for a complex notion (such as *democracy*, *globalization*, *empire*, etc.), s/he is bound to consider the historic and the current geopolitical context. With that s/he risks being steered into an interpretation which is torn away from a user of the language by identifying and giving prominence to those semantic components that remain within the domain of political terminology. This results in a dictionary entry failing to mirror the representation of the notion in the linguistic consciousness of present day native speakers.

We argue that to compile a dictionary entry a researcher needs to consider those semantic components that make up the integral parts of the mental representation verbalized by this word. Such an approach is certain to help a definition and interpretation of any complex notion stand up to the criteria of relevance, and to reflect the current language use.

## 2 Research Objectives and Foundations

Our research focuses on the complex notion of *empire* which is engraved in the public consciousness of the population of any powerful, domineering state, including the USA, China, Russia and many more (these states are commonly dubbed superpowers). Given the current global political and economic environment, we considered those cognitive structures which are characteristic of the population of the USA and Russia.

Common sense suggests that any state that is characterized by a proactive position in global terms, that is involved in international affairs and to a large extent dictates domestic policies of other states, adopts a number of features of an empire. To ensure the best domestic and international performance

of such a superpower a certain effort is invested to promote these features through the mass media and political speeches (see examples 1-3 of American media texts, and examples 4-5 of Russian media texts).

- (1) Fighting back tears, Bush vows that America will “**lead the world to victory**” over terrorism in a struggle he termed the first war of the 21st century. (www.september11news.com)
- (2) At the center of the war’s vast changes was the military – transformed by the nation into a colossus and, in turn, transforming the nation into a superpower. (*Time*, March 9, 1998)
- (3) Whether it is a **fight** against fascism or communism, or even misconceived interventions like Vietnam, America’s mission is **to further** not only its interests but also its **values**. And that idealism streak is the source of **its global influence**, even more than its **battleships**. (*Time*, April 13, 1998)
- (4) Вести из стран СНГ как **вести с фронта. Потеряны** Грузия, Украина, Молдавия, Киргизия. Аджария **пала. Враги устраивают диверсии** в Узбекистане и Казахстане, **подбираются** к Белоруссии. Минск **держится**, но если он **падет**, страшно подумать, - открывается дорога на Москву. Что же за **война** идет на просторах СНГ? Кто, с кем и за что **воюет**? (*НГ*, 27.03.2006) [The news arriving from the CIS states is like the news from the combat zone. Georgia, Ukraine, Moldavia, Kirgizia are lost. Adzharia has fallen. The enemy is sabotaging Uzbekistan and Kazakhstan and is advancing on Belarus. Minsk is holding ground, yet if it surrenders, Moscow will remain unguarded. What war is sweeping the CIS states? Who is combating who? And why?]
- (5) Глава МИД РФ Сергей Лавров в интервью радиостанции «Эхо Москвы» **подверг** Тегеран **беспрецедентно жесткой критике**. Более того, российский министр дал понять, что Россия **поддержит предложение передать ядерное досье** Ирана в Совбез ООН – хотя до сих пор Москва никогда **не одобряла такого шага**. (*Коммерсантъ*, 13.01.2006) [In his interview to “Ekho Moscvy” Sergey Lavrov, head of the RF Ministry of Foreign Affairs, lashed Teheran. Moreover, the minister suggested that Russia would support handing over Iran’s nuclear development profile to the UN Security Council despite the fact that Russia’s government never before supported such measures.]

It is vital to convey the positive image of the empire and establish strong links between the components of the corresponding cognitive structure and the current policies of the superpower. Yet it takes additional effort, as it is at odds with the democratic values which have been adopted and promoted by these states. The detailed analysis of the media texts featured in a number of mainstream Russian and American newspapers and magazines over the past 20 years suggests that the mass media are establishing and maintaining these links by making the features of an empire explicit in contexts relating to the current domestic and international affairs (some of the examples are listed above). It results in the transformation of the cognitive structure of the notion of empire embedded in the public consciousness of the present-day population of both states. Without going any further into the details of this analysis (as it remains outside the scope of this paper) we would like to focus on the changes in the cognitive structure which become apparent when we compare the dictionary entry for empire which reflects the conventional interpretation and use of the notion, and the findings of the associative experiment with American English and Russian native speakers.

### 3 Componential Analysis Findings

At the initial stage of the research we studied different dictionary entries to compile a list of relevant semantic components of the word *empire*. To do this we relied on the *Dictionary of Contemporary*

*Russian Literary Language* (1956-65), *Dictionary of the Russian Language* edited by A.P. Evgenyeva (1999), *Oxford English Dictionary* (1989), *Webster's New World Dictionary* (1991), *Macmillan's English Dictionary* (2000), *Webster's New College Dictionary* (1995), *Political Encyclopedia* (1999), *Historical and Etymological Dictionary of Contemporary Russian Language* (Chernykh 2003), *Glossary.ru: Dictionaries for Social Sciences* ([www.glossary.ru](http://www.glossary.ru)), and *Dictionary of Foreign Words* (Krysin 2000). Listed below are some examples of dictionary entries:

- (1) Империя – исторически преходящая форма государства, характеризующаяся обширной, но не обязательно целостной территорией, многонациональным составом населения, централизованным (монархическим) управлением, стремлением к политическому и силовому господству в мировом масштабе (*Political Encyclopedia* 1999: I, 429) [Empire – a historically transient form of state characterized by a vast, still not necessary single territory, multiethnic population, centralized power (monarchy), aspiring to exercise political and military control globally].
- (2) Империя – 1. Монархическое государство, во главе которого стоит император (Империя Карла Великого). 2. Крупная империалистическая держава (Британская империя) (*Dictionary of the Russian Language* edited by A.P. Evgenyeva 1999: I, 662) [Empire – 1. A monarchical state reigned by an emperor (i.e. Empire of Charles the Great. 2. A large imperialistic state (the British Empire)].
- (3) Empire – 1. Imperial rule or dignity. 1). Supreme or extensive political dominion, esp. that exercised by an emperor or by a sovereign state over its dependencies. 2). Paramount influence, absolute sway, supreme command or control. 3). The dignity or position of an emperor, the reign of an emperor. 2. That which is subject to imperial rule. 1). An extensive territory (esp. an aggregate of many separate states) under the sway of an emperor or supreme ruler, also, an aggregate of subject territories ruled over by a sovereign state (*Oxford English Dictionary* 1989: V, 187).
- (4) Empire – 1. Supreme rule, absolute power or authority. 2a. Government by an emperor or empress. 2b. The period during which the government prevails. 3a. A group of states or territories under the sovereign power or an emperor or empress. 3b. A state uniting many territories and under a single sovereign power. 4. An extensive social or economic organization under the control of a single person, family or corporation (*Webster's New World Dictionary* 1991: 445).

As a result of the componential analysis we made up two lists of the relevant semantic components for the word *empire* (for the Russian and English languages). These lists included those content words (naturally the function words were not considered) which were repeatedly used by lexicographers in different dictionaries, hyponyms were then replaced with superordinate terms (such as 'emperor' / 'empress' / 'sovereign' with 'monarch'), words identical or nearly identical in meaning

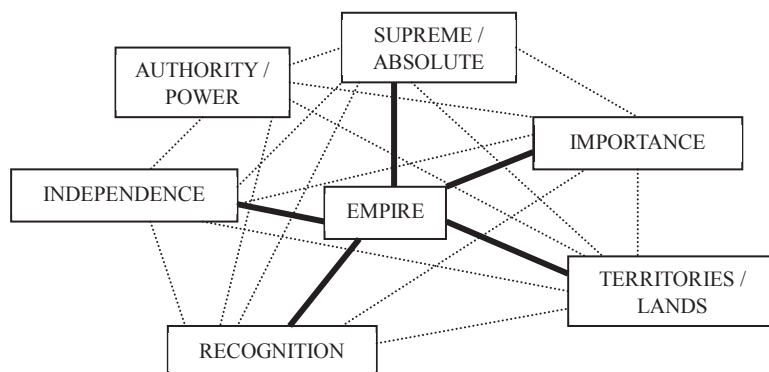


Figure 2: Relevant semantic components of the word *empire* (in the English language).

were separated by a slash (such as ‘supreme’ / ‘absolute’). We referred to these semantic components as *relevant* because statistically they were most commonly attributed to an empire. By employing many and diverse lexicographic sources we attempted to make the list as complete and comprehensive as possible. These lists were similar yet not identical for the Russian and English languages. The findings of this stage of our research are featured in Figures 2-3. These lists of the relevant semantic components served as the reference point for our further study.

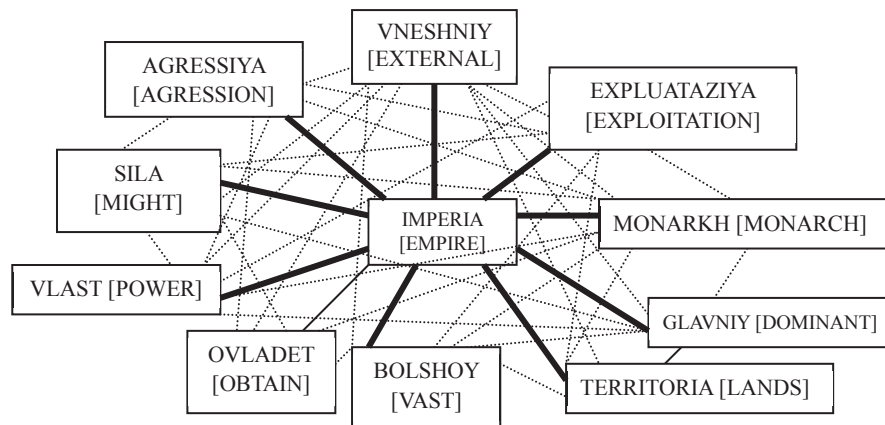


Figure 3: Relevant semantic components of the word *empire* (in the Russian language).

## 4 Associative Experiment Findings

We then proceeded to the associative experiment with 434 Russians and 148 Americans (both groups of subjects involved people in three age ranges (under 25, 25 to 45 and over 45) and were balanced in terms of gender to make them representative).

Table 1: Number of subjects in each cluster.

Citizenship	Russian			American		
Age span	<25	25-40	>40	<25	25-40	>40
Number of subjects	139	138	157	48	63	37

### 4.1 Procedure and Materials Used

The subjects were offered a questionnaire featuring a list of stimulus words (involving all the items identified at the initial stage of the research) and distracters (random words which prevent associative linking when a subject fulfills the task). The task said that the subject was expected to write down any word or collocation which flashes through his/her mind on reading the stimulus word (only the first association was to be put down). The subjects were asked to complete the task within a very tight timeframe which was calculated by allocating about seven seconds to put down an association for each stimulus word (the timing slightly differed for American and Russian subjects as the number of stimulus words was different, but in both cases it was under five minutes).

### 4.2 Verified Hypotheses

The associative experiment was designed to verify the following hypotheses (in propounding the hypotheses we relied on the findings of the componential analysis and the analysis of the media texts



featured in a number of Russian and American mainstream newspapers and magazines over the past 20 years).

1. As the democratic values supported and promoted internationally prevent the mass media from establishing direct links between a superpower and empire, the word *empire* is never used in describing the current domestic and international activity of the state. Yet media and political texts do verbalize an associative link with the relevant components of the corresponding cognitive structure. Thus the first hypothesis implies that the link between the cognitive structure and the term *empire* is deteriorating. This means that the number of associations between stimulus words denoting different components of the cognitive structure and the word *empire* is unlikely to be high.
2. It can be assumed that the hierarchy of the relevant components of this cognitive structure is being reconsidered, with new elements becoming central to the structure and replacing empire as the integral component. This implies that some stimulus words are likely to prompt more diverse and multiple associations than others (including the stimulus word *empire*).
3. As American newspapers and magazines promote the positive image of empire by drawing links between this cognitive structure and fundamentally religious concepts (such as messiah, apocalypse, antichrist, etc.), we can suggest that the overall sentiment with the US population is rather positive. The US mass media seem to have implanted a positive image of empire in the public consciousness, and are efficiently promoting it further. This means that we should expect more associations reflecting the positive appraisal of empire in the responses of the American subjects than that in the responses of the Russian subjects.

### 4.3 Analysis and Interpretation of the Experimental Data

Initially the reactions to the stimuli featured in the associative experiment were classified into clusters joining together words and collocations with similar or identical meanings. Step two related to calculating the frequency of each cluster of associations and on the basis of their frequency arranging them into a hierarchy to construct a model of the cognitive structure represented by the word *empire* as featured in the public consciousness of contemporary Russian and American English native speakers (the most frequent responses have been placed at the top of this hierarchy, and are referred to as relevant components of the cognitive structure in the text to follow).

The most frequent associations to the stimulus word *empire* included *State Building 15*, *state 11*, *strikes back 8*, *Rome 7* (the figures name the number of English-speaking subjects having provided these responses) among them superordinate terms (*state*), historic notions (*Rome*), realia (*[Empire] State Building*) and allusions (*strikes back* – an allusion to the *Star Wars* series of films). The relevant components of the cognitive structure for Russian native speakers were represented by the following associations *power 51*, *of evil 36*, *state 33*, *great 17*, *Roman 16*, *country 16*, *vast 14*, *Russia 14*, *Russian 12*, *emperor 11*, *Rome 11*, *evil 10*, *might 10*, *tsar 10*, *crown 9*, *powerful 9*, *of passion 9*, *sovereign state 7*, *very large 7*, *of feelings 6*. In terms of content these are superordinate terms (*state*, *country*), characterizing adjectives (*very large*, *powerful*), historic notions (*Rome*, *Roman*, *Russian*), and set collocations (*empire of evil*), to name just a few. Similarly, we worked through pools of associations to every stimulus word.

The findings of the experiment were contrasted with the initial model of the word meaning, and we discovered that some of the key semantic components tend to fade away in the public consciousness, giving way to some of the peripheral ones. For instance, when analyzing the responses of the Russian subjects we observed that the association *empire* turned out to be very rare: *monarch – empire* (eight instances), *of the empire* (three instances); *power – empire* (two instances). At the same time some peripheral components of the cognitive structure showed strong bilateral links: *power – might*,

*power – monarch, obtain – might, monarch – power, monarch – dominant, exploitation – might, lands – vast.* The association *might* appeared to have links with most relevant components of the cognitive structure, which implies that it is gaining momentum and taking over the dominant position in the hierarchy. These findings clearly demonstrate that the cognitive structure denoted by the word *empire* is undergoing a structural change, and suggest that the term *empire* is becoming torn away from its mental representation. These changes are featured in Figure 4 (two-way arrows demonstrate bilateral associative links, one-way arrows show that the link is unilateral and rather unstable).

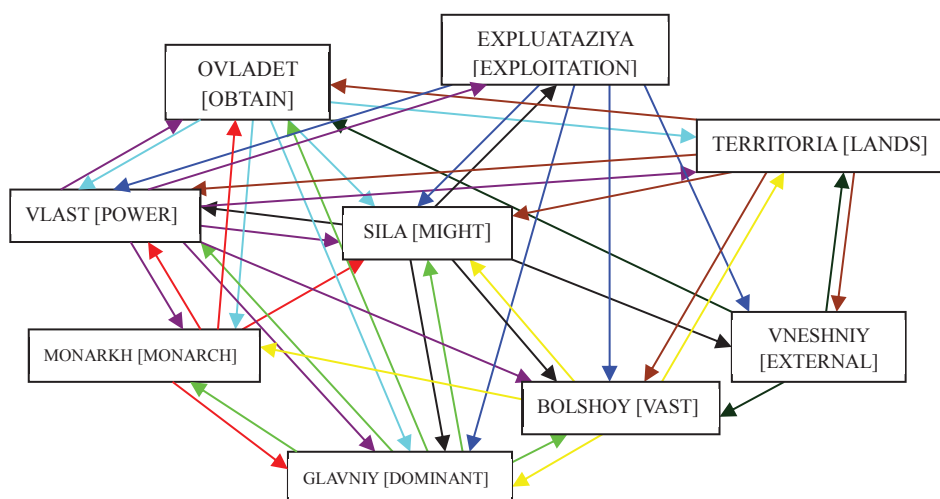


Figure 4. Changes in the hierarchy of the relevant components of the cognitive structure denoted by the word *empire* (in the public consciousness of Russian native speakers).

The data obtained from American English native speakers turned out to be less consistent. The most apparent link was established between power and might (stimulus: *power* – associations: *outrage, oppression, strength, strong, struggle, surge, violent* (26% of all responses), stimulus: *might* – associations: *power* (22.6% of all responses). The analysis of the experimental data suggests that the links between other components of the cognitive structure represented by the term *empire* are mediated by these two components (*might* and *power*) as they feature as associations for nearly every stimulus word: *recognition – power; supreme – power; importance – might; territory – power; independence – power + might*. Judging by these findings both *power* and *might* are pushing their way to the center of the cognitive structure, though it remains unclear whether the other components are adopting new roles within the cognitive structure or not (we refrained from drafting a figure in this case as the structure is not transparent).

Overall *might* turned out to be very frequent in the associations of the American subjects with its frequency in responses of the Russian speakers far behind (18% for American English native speakers vs. 8.4% for Russian native speakers). We thus consider this component to be the integral element of the cognitive structure represented by the word *empire* in the public consciousness of the US population. At the same time it is certainly gaining momentum in the corresponding cognitive structure of Russian native speakers, with the trend quite apparent.

Another finding related to the fact that there appears to be a very vague link between the initial definition of *empire* as a powerful economically and politically developed state that enjoys dominion over other territories and is commonly reigned by a sovereign, and the word *empire* which routinely pops up in collocations like *the evil empire, Empire State Building, business empire* and other examples. These associations signal that the word *empire* is losing its link with the historic concept.

In the responses of the Russian subjects there were 32% of associations that pertain to the definition of the term (including *power*, *vast*, *emperor (monarch)*, *might* and superordinate terms *state*, *country*). Accordingly, in the responses of the American subjects there were 20% of associations that reflect relevant semantic components (including *big*, *large*, *vast*, *great*, *dictator*, *emperor*, *king*, *royalty*, *crown*, *government*, *reign*, *rule*, *power*, *powerful*, *dominant*, *kingdom*, *realm*, *dominion*). The majority of associations, however, were not linked to the definition of the term as featured in conventional dictionaries.

We have also attempted to verify a hypothesis that, despite the fact that the mass media aim to support the positive image of the state and relate to the mental representation of empire on a day-to-day basis (though latently), it is nonetheless negatively appraised by the population of Russia, while the American mass media seem to be quite successful in promoting the positive appraisal of empire. Yet there were very few associations that fall into this category: positive appraisal in 0.2% of responses vs. negative appraisal in 12% of responses (for Russian native speakers), and positive appraisal in 2% of responses vs. negative appraisal in 2% of responses (for American subjects). Thus we cannot consider this hypothesis confirmed or disproved.

Yet another finding became apparent when we contrasted different age groups. It appeared that the Russian subjects aged between 25 and 40 frame considerably fewer associations that relate to the relevant semantic components of the word *empire* than other age groups (25% vs. 34% by the subjects aged under 25 and 35% by the subjects aged over 40). From our perspective, this can be accounted for by the environment in which these people were raised: they were children and young adults at the time of *perestroika* (mid-80s) and can adapt well to the new realia and corresponding changes in certain cognitive structures.

When contrasting the data obtained from the American and Russian subjects, we observed that Americans suggested fewer relevant semantic components than their Russian counterparts, with the most obvious discrepancies in the age group “under 25”. It is apparent that these discrepancies stem from evident differences in the social and political environments of these two countries.

Relying on the analysis and interpretation of the experimental data we have managed to confirm the first two hypotheses. However, the third hypothesis needs further research and cannot be either confirmed or disproved as of today, and perhaps it requires a different experimental method or larger groups or subjects.

## 5 Conclusions

This paper presents just some of the findings which could be employed in lexicographic work to extend and specify the dictionary entry of the word *empire*. Though associative meaning is sometimes seen as beyond the scope of dictionary-making practices, we argue that associative experiments can be an efficient tool in lexicography.

In the final analysis, the findings of the associative experiment carried out with Russian and American English native speakers helped reveal the relevant semantic components of the word meaning which are embedded in the linguistic consciousness of contemporary native speakers. It was revealed that empire is no longer viewed in its conventional meaning as a form of state. Instead, it is commonly seen as a policy implying a proactive position in global terms, involvement in various international affairs and exerting influence. This change in the hierarchy of the relevant components of the mental representation of empire opens loopholes for those states which exercise control and adopt domineering positions as it salves negative sentiment and helps promote such a policy.

Moreover, the associative experiment demonstrated that the number and type of the relevant components of this cognitive structure are culture-determined, with American and Russian subjects suggesting contrasting associations. The universal trend is for the term *empire* to become torn away from its mental representation, yet this process is more apparent with Americans.

These findings are a helpful resource for lexicographers, as they reflect those semantic and associative components which are relevant for language users given the current political and economic environment.

## References

- Chernykh, P.Ya. (2003). *Istoriko-etymologicheskyy slovar sovremennogo russkogo yazyka* [Historical and Etymological Dictionary of Contemporary Russian Language]. Vol. 1-2. Moscow: Russkiy yazyk. (in Russian).
- Deese, J. (1965). The structure of associations in language and thought. Baltimore: Johns Hopkins Press.
- Glossary.ru: *slovari po obschchestvennym naukam* [Glossary.ru: Dictionaries for Social Sciences]. Accessed at: [www.glossary.ru](http://www.glossary.ru) [26/03/2018]. (in Russian).
- Iashin, P.N. (2009). *Natsionalno-kulturnaya specifika obrazov "zhizn" i "smert" v yazykovom soznanii russkikh, nemtsev i anglichan* [National and Cultural Specifics of Notions of "Life" and "Death" in the Linguistic Consciousness of Russian, German and English Native Speakers]. Moscow. (in Russian).
- Krysin, L.P. (2000). *Tolkovy slovar inoyazychnykh slov* [Dictionary of Foreign Words]. Moscow: Russkiy yazyk. (in Russian).
- Macmillan's *English Dictionary* (2000). Oxford.
- Martinek, S. (2009). Everyday Experience in Word Meaning: How an Associative Experiment Reveals it. In: *Studies in Language and Cognition*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 142-159.
- Oxford English Dictionary* (1989). Vol 1-20. London.
- Panarina, N.S. (2017). *Psikholingvisticheskoye modelirovaniye mehanizma realizatsii pretsedentnosti* [Psycholinguistic Modeling of the Mechanism of Precedence]. Moscow. (in Russian).
- Petrova, E.Y. (2008). *Tsvetooboznacheniya v yazykovom soznanii vtorichnoy yazykovoy lichnosti: na material russkogo i frantsuzskogo yazykov* [Color Nominations in the Linguistic Consciousness of a Secondary Language Person: in Russian and French Languages]. Moscow. (in Russian).
- Pishchalnikova, V.A. (2002). Osnovaniya dinamicheskoy teorii znacheniya: kognitivnyy aspekt [Foundations of the Dynamic Theory of Word Meaning: Cognitive Aspects]. In: *Lukashevitch E.V. Kognitivnaya semantika: evolyutsionno-prognosticheskiy aspekt* [Cognitive Semantics: Evolutionary and Prognostic Aspects]. Moscow – Barnaul: Altay University. (in Russian).
- Politicheskaya Entsiklopedia* [Political Encyclopedia] (1999). Vol. 1-2. Moscow: Mysl. (in Russian).
- Slovar russkogo yazyka pod redaktsiye A.P. Evgenyevoy [Dictionary of the Russian Language edited by A.P. Evgenyeva] (1999). Vol. 1-4. Moscow: Russkiy yazyk. (in Russian).
- Slovar sovremennogo russkogo literaturnogo yazyka [Dictionary of Contemporary Russian Literary Language] (1956-65). Vol. 1-17. Moscow – St. Petersburg: Izvestiya akademii nauk. (in Russian).
- Stepykin, N.I. (2011). *Sposoby strukturno-soderzhatelnogo modelirovaniya lingvokulturnogo kontsepta* [Ways of Structural and Content-Related Modeling of a Lingvocultural Concept]. Moscow. (in Russian).
- Ufimtseva, N.V. (2014). The Associative Dictionary as a Model of the Linguistic Picture of the World In: *Procedia – Social and Behavioral Sciences*, 154, pp. 36-43.
- Webster's New College Dictionary* (1995). New York.
- Webster's New World Dictionary* (1991) New York.
- Zalevskaya, A.A. (2011). *Znachenkiye slova cherez prizmu experimenta* [Word Meaning through the Prism of an Experiment]. Tver: University of Tver. (in Russian).
- Zalevskaya, A.A. (2014). *Chto tam – za slovom? Voprosy interfeisnoy teorii znacheniya slova* [What is there beyond the Word? Issues Related to the Interface Theory of Word Meaning]. Moscow – Berlin: Direct-Media. (in Russian).





# Lexicographic Potential of the Syntactic Properties of Verbs: The Case of Reciprocity in Czech

Václava Kettnerová, Markéta Lopatková

Charles University

E-mail: kettnerova@ufal.mff.cuni.cz, lopatkova@ufal.mff.cuni.cz

## Abstract

Reciprocity has been the focus in much theoretical research in recent years. It has been primarily studied as a grammatical property, which is not of high relevance for the description of the lexical stock of a language. At the same time, however, it has been widely accepted that languages substantially differ with respect to the inventory of words allowing for reciprocity, and that the applicability of reciprocity is rarely derivable from the semantic and/or syntactic properties of these words. The integration of the information on reciprocity into lexicons would thus be highly beneficial for both human users (esp. for foreign speakers) and for natural language processing tasks. In this paper, we demonstrate how the reciprocity of Czech verbs can be represented in a lexicon in a comprehensive and systematic way. Czech represents a language where reciprocity is a highly productive phenomenon. We show which semantic and syntactic properties are relevant for the description of reciprocal verbs, and based on this a user (be it human or computer) can acquire their reciprocal constructions.

**Keywords:** reciprocity, reciprocal verbs, valency structure of verbs, lexicon, syntax, Czech

## 1 Introduction

In the last century, linguistic research had a strong tendency to disassociate meaning from form. Particularly in the tradition of transformational generative grammar, many linguistic studies adopted the unchallenged view that the form exhibited by a word is independent from its meaning. Under this view, a lexicon serves as an inventory of separate words bearing some meanings, while a grammar provides grammatically correct combinations of these words. However, the development of corpus linguistics has revealed that semantically similar words exhibit similar grammatical patterns, indicating that there are many interdependencies between the grammatical properties of words and their meaning (Sinclair 1991; Levin 1993). In this paper, we demonstrate how a primarily grammatical property of words, namely reciprocity, can contribute to a better description of the vocabulary of a language. We focus on Czech reciprocal verbs and their representation in a valency lexicon of Czech verbs, VALLEX (Lopatková et al. 2016).<sup>1</sup>

Reciprocity is generally understood as a complex of forms and patterns of mutuality and exchange. In line with König and Kokutani (1996) and Haspelmath (2007), among others, we distinguish between symmetry as a semantic property of a word and reciprocity as a grammatical or lexical coding of the given property. Let us repeat a notorious description of *symmetric predicates* as predicates that denote binary (or  $n$ -ary, where  $n \geq 2$ ) relations  $R$  among members of a set  $A$  of semantic participants with the following semantic property:

(i) “ $x, y \hat{I} A (x \neq y \rightarrow R(x, y))$ ” (König & Kokutani 2006);

as a consequence, for two particular  $a, b \hat{I} A$  it holds  $(R(a, b) \leftrightarrow R(b, a))$ .

<sup>1</sup> <http://ufal.mff.cuni.cz/vallex/3.0>

Reciprocal constructions are then grammatical means for the expression of symmetrical relations for an  $n$ -ary predicate and for set of participants  $A$  with a cardinality of at least 2 ( $|A| \geq 2$ ). For example, in (1a) the predicate *hádat se* ‘quarrel’ denoting relation  $R$  among semantic participants from the set  $A = \{\text{Petr, Pavel}\}$  is a symmetric predicate, as the participants from the set  $A$  (*Petr* ‘Peter’ and *Pavel* ‘Paul’) are distinct and related to each other by the relation  $R$ , as required by (i). A typical reciprocal construction is then instantiated in (1a).

- (1a) Petr se hádá s Pavlem a zároveň Pavel se hádá s Petrem.  
Peter REFL quarrels with Paul and at the same time Paul REFL quarrels with Peter  
‘Peter is quarreling with Paul and at the same time Paul is quarreling with Peter.’
- (1b) Petr a Pavel se hádají.  
Peter and Paul REFL quarrel  
‘Peter and Paul are quarreling.’
- (2) Petr a Pavel se na sebe dívají.  
Peter and Paul REFL at REFL look  
‘Peter and Paul are looking at each other.’

Let us stress, however, that reciprocals are not associated with a uniform meaning – on the contrary, their meaning varies, as discussed in detail by Dalrymple et al. (1998). The above attempt to formally describe symmetry and reciprocity is relevant for the so-called strong reciprocity when each member of the set  $A$  is related by the relation  $R$  to every other member (Langendoen 1978). Formula (i) holds for most reciprocal structures in which two participants are involved, as in (1b) and (2). Reciprocity can, however, be associated with different semantic facets; as these facets are not linguistically structured and they typically remain vague, we leave them aside here.<sup>2</sup>

*Reciprocity* represents the linguistic means for encoding symmetry. It can be characterized as an operation resulting in the fact that two (sometimes more)<sup>3</sup> valency complementations of a predicate stand in symmetry. In Czech, verbs (1a,b) and (2), nouns (3), adjectives (4) or even some adverbs (5) can be used as reciprocal predicates. In reciprocal constructions, one valency complementation of these predicates is typically occupied by the whole set  $A$ , while the second is either reduced on the surface (1b), (3), and (5), or filled by coreferential expressions (2) and (4). As a result, dual thematic roles (in an unreciprocal structure mapped onto two complementations separately) are then associated with both valency complementations involved in symmetry, see Figure 1, displaying double mapping of thematic roles Agent and Patient onto valency complementations ACT and PAT with the verb *políbit* ‘to kiss’.

2 For example, if more than two participants are involved, a weaker condition on symmetry may be applied:

$\$ x, y \hat{A} (x \neq y \rightarrow (R(x, y) \leftrightarrow R(y, x)))$

The weaker condition is a more probable interpretation for, for example, *Petr, Pavel a Hanka se na sebe dívají*. ‘Peter, Paul and Hannah are looking at each other.’, which can be interpreted as, for example, *Petr a Pavel se dívají na Hanku, Hanka se dívá jen na Petra*. ‘Both Peter and Paul are looking at Hannah, Hannah is looking at Peter only.’

Such semantic nuances are primarily associated with ways of how a particular action can be performed. For example, in *Pytle s pískem jsou naskládány na sebe*. ‘Sandbags are stacked on top of each other.’, if sandbags should effectively function as a flood barrier, they must be arranged on top of each other in an overlapping way. In other situations, as there are on building sites, such an arrangement is not necessary and the sandbags can be put in piles without overlapping. Although the truth conditions are different, both these situations can be described by the same reciprocal sentence.

3 Although reciprocity involving two participants prevails in a language, reciprocity of three participants is not excluded. See example of the Czech verb *představovat* ‘to introduce’ in the following reciprocal structure with the interpretation that each child introduced another child to every other children:

*Děti se představovaly (navzájem).*

children REFL introduced (mutually)

‘Children were introducing each other.’

- (3) vyostřená hádka mezi Petrem a Pavlem  
‘escalated quarrel between Peter and Paul’
- (4) Petr a Marie si byli věrní.  
Peter and Mary REFL were faithful  
‘Peter and Mary were faithful to each other.’
- (5) Domy jsou orientovány rovnoběžně.  
‘Houses are oriented in parallel.’

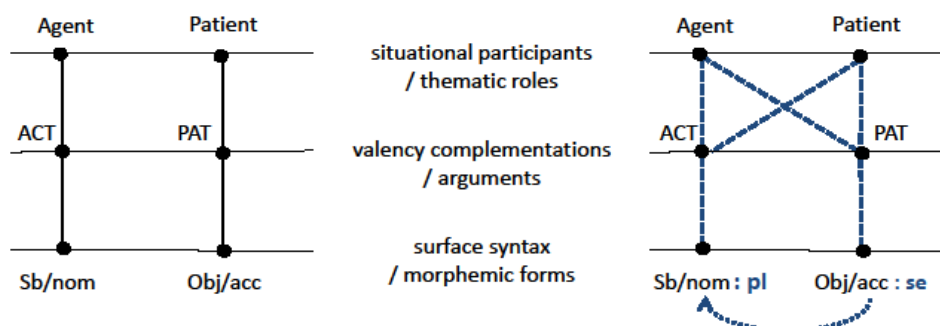


Figure 1: Schematic representation of the sentences *Jan políbil Marii*. ‘John kissed Mary.’ (left) and *Jan a Marie se políbili (navzájem)*. ‘John and Mary kissed each other.’ (right). For the reciprocal construction, the scheme models doubling the thematic roles of the situational participants, their plurality, and symmetrical correspondence to valency complementations, as well as their morphemic forms.

Reciprocity did not attract much attention in either theoretical or computational linguistics until the turn of the century. Since then, reciprocity has been gradually gaining interest among theoretical linguists, the focus being on both syntactic and semantic analyses and cross-linguistic typological studies (König & Kokutani 2006; Heim et al. 1991; Dalrymple et al. 1998; Siloni 2001; Frajzyngier & Curl 2000; Nedjalkov 2007; König & Gast 2008; Evans et al. 2011). As reciprocity is expressed mostly by regular grammatical means, it is predominantly treated as a grammatical phenomenon. However, as Siloni (2002) and Reinhart and Siloni (2005) argue, reciprocity is of high relevance to both grammar and lexicon in many languages. Czech is one of the languages where reciprocity is reflected in both lexicon and grammar, thus representing a prototypical phenomenon at the lexical-grammar interface. As such, in this paper we attempt to provide a comprehensive and systematic representation of reciprocal verbs, making use of both parts of the language description. As reciprocity is lexically conditioned and its applicability to verbs cross-linguistically varies, the theoretical results achieved here can be further made use of in practical lexicography in building both monolingual and bilingual dictionaries. In other words, reciprocity as a lexically determined characteristic of verbs can greatly assist in an adequate description of their meaning, and thus also in better word-sense disambiguation. The lexicographic description of reciprocity can be beneficial for exploring which strategies verbs adopt for encoding mutuality.

The rest of this paper is structured as follows. First, two basic types of Czech reciprocal verbs are distinguished according to whether they encode symmetry in their lexical meaning or not. Further, their representation in the valency lexicon of Czech verbs, VALLEX, is proposed (Section 2). Second, a theoretically adequate and economical representation of their reciprocal constructions is provided for the given valency lexicon (Section 3).

## 2 Czech Reciprocal Verbs in VALLEX

In Czech reciprocity is expressed by the verbs that either denote a mutual situation, or can potentially denote such a situation (Panevová & Mikulová 2007). The former group of verbs – the so-called inherent or lexical reciprocal verbs, as symmetry is an inherent part of their lexical meaning – is semantically restricted (see Section 2.1). The latter group of verbs is semantically very broad; their semantic interpretation is primarily asymmetrical (e.g., *podezírat* ‘to suspect’, *klamat* ‘to deceive’), but they denote events which can be on certain conditions – primarily lexical – conceived as mutual. This group is referred to as syntactic reciprocal verbs (see Section 2.2).

### 2.1 Inherent Reciprocal Verbs in Czech

Inherent reciprocal verbs are those that bear the semantic feature of symmetry in their lexical meaning (Evans, 2008), as discussed for Czech in Panevová and Mikulová (2007). Their various types are described below, with emphasis on different functions of the reflexive clitics *se* and *si*, which represent (besides their other functions) one of main grammatical ways of encoding reciprocity in Czech. The reflexive clitics in Czech, as in other European languages, are highly polysemous, marking some word formation processes, reflexivity, middle voice, and reciprocity, see e.g. (Medová 2009). In Czech reciprocal constructions, the reflexive clitics represent either a part of verb lemmas, or the inflected forms of the reflexive pronoun; the latter can be substituted – depending on word order and topic-focus articulation – by their long forms *sebe* and *sobě*, respectively. While the reflexive clitics of the first type are associated with verb lemmas (and not with any valency position of a verb), the clitics of the latter type fill one of their valency positions, just like nouns and other pronouns, as we show below.

#### 2.1.1 Types of Inherent Reciprocal Verbs

Inherent reciprocal verbs encompass symmetry in their lexical meanings: they especially express social actions or relations (e.g., *hádat se*, ‘to quarrel’, *spolupracovat* ‘to cooperate’, *vyjednávat* ‘to negotiate’), spatial relations (e.g., *sousedit* ‘to adjoin’, *oddělit* ‘to separate’), and relations of (non-) identity (e.g., *rozlišit* ‘to distinguish’), e.g. (Haspelmath 2007). From the point of view of syntactic properties, Czech inherent reciprocal verbs can be either intransitive (6) and (8), or ditransitive (7) and (9). One participant involved in symmetry is typically mapped either onto the subject position, or onto the direct object position, while the other is expressed in the indirect object position, typically in the form of the comitative prepositional group *s+Instr* ‘with+Instr’, see examples (6b) and (9b). A limited number of inherent reciprocal verbs are characterized by the form *od+Gen* ‘from+Gen’ (e.g., *oddělit* ‘to separate’, *izolovat* ‘to isolate’, *rozlišit* ‘to distinguish’, *rozpoznat* ‘to recognize’, etc.).

In reciprocal constructions of inherent reciprocal verbs, participants can be reciprocalized, i.e., expressed in a single syntactic position. Inherent reciprocal verbs predominantly exhibit subject-oriented reciprocity, where the reciprocalized participants of an event denoted by a verb are expressed in the subject position (6a), (7a), (8). Object-oriented reciprocity, where reciprocalized participants occupy the direct object position, is rather limited in its number (9a). For further information on the reciprocal constructions of these verbs, see Section 3.1.

(6a) Petr a Pavel (si) korespondovali.

Petr<sub>nom.sg.masc</sub> and Paul<sub>nom.sg.masc</sub> (REFL<sub>verblemma</sub>) corresponded  
 ‘Peter and Paul corresponded with each other.’

- (6b) Petr (si) korespondoval s Pavlem. ≈ Pavel (si) korespondoval s Petrem.  
 Peter<sub>nom.sg.masc</sub> (REFL<sub>verblemma</sub>) corresponded with Paul<sub>s+instr.sg.masc</sub>. ≈ Paul<sub>nom.sg.masc</sub> (REFL<sub>verblemma</sub>) cor-  
 responded with Peter<sub>s+instr.sg.masc</sub>  
 ‘Peter corresponded with Paul. ≈ Paul corresponded with Peter.’
- (7a) Kolegové (spolu) diskutovali všechna rozhodnutí.  
 colleagues<sub>nom.pl.masc</sub> (together) discussed all decisions  
 ‘Colleagues discussed all decisions with each other.’
- (7b) Kolega diskutoval všechna rozhodnutí s kolegou.  
 colleague<sub>nom.sg.masc</sub> discussed all decisions with colleague<sub>s+instr.sg.masc</sub>  
 ‘The colleague discussed all decisions with his colleague.’
- (8) Kamarádi se (spolu) sázeli o pivo, kdo bude rychlejší.  
 friends<sub>nom.pl.masc</sub> REFL<sub>verblemma</sub> (together) bet about beer who will be faster  
 ‘Friends were betting about beer who would be faster.’
- (9a) Kriminalisté porovnávali otisk prstu A a otisk B.  
 criminal investigators compared print<sub>acc.sg.masc</sub> of finger A and print<sub>acc.sg.masc</sub> B  
 ‘Criminal investigators were comparing fingerprint A and fingerprint B.’
- (9b) Kriminalisté porovnávali otisk prstu A s otiskem B. ≈ Kriminalisté porovnávali otisk prstu B s  
 otiskem A.  
 criminal investigators compared print<sub>acc.sg.masc</sub> of finger A with print<sub>s+instrsg.masc</sub> B. ≈ Criminal inves-  
 tigators compared print<sub>acc.sg.masc</sub> of finger B with print<sub>s+instr.sg.masc</sub> A  
 ‘Criminal investigators were comparing fingerprint A with fingerprint B. ≈ Criminal investiga-  
 tors were comparing fingerprint B with fingerprint A.’

In contrast to syntactically reciprocal verbs (see Section 2.2), inherent reciprocal verbs express symme-try even when the participants of the events denoted by these verbs are not reciprocalized. See examples (6b) and (9b), where the participants are expressed in separate syntactic positions provided by their valency complementation. Due to the symmetry as an inherent part of the meaning of these verbs, the participants expressed in separate syntactic positions can be switched without any change in meaning.<sup>4</sup>

(A) *Irreflexive inherent reciprocal verbs.* Some inherent reciprocal verbs are characterized by irre-flexive lemmas, like *diskutovat* ‘to discuss’ in (7) or *porovnávat* ‘to compare’ in (9). In reciprocal constructions with these verbs – regardless of whether the participants involved in symmetry are reciprocalized (7a) and (9a), or not (7b) and (9b) – no reflexive clitic *se* or *si* is present.

Many constructions with inherent reciprocal verbs, however, contain the reflexive clitic *se* or *si*. As this reflexive clitic is present in all instances of these verbs, it is usually classified as a part of their verb lemmas.<sup>5</sup> The reflexive inherent reciprocal verbs in Czech can be further subclassified

4 We disregard changes in topic-focus articulation here.

5 The classification of the clitic *se* or *si* as a part of verb lemmas with inherent reciprocal verbs is supported by the fact that their presence in such constructions is not associated with any valency position of these verbs, as the following constructions of the verb *sázet se* ‘to bet’ show: if the clitic *se* is replaced (i) by the long form of the reflexive pronoun *sebe*, (ii) by the pronoun *je* ‘them’, or (iii) by the noun *kolegové* ‘colleagues’, it necessarily adds an extra valency position of the verb *sázet se* ‘to bet’, which results in ungrammatical structures:

(i) \*Kamarádi sebe sázeli o pivo, kdo bude rychlejší. / \*Sebe kamarádi sázeli o pivo, kdo bude rychlejší.  
 friends REFL-long bet about beer who will be faster / REFL-long friends bet about beer who will be faster  
 (ii) \*Kamarádi je sázeli o pivo, kdo bude rychlejší. / \*Je kamarádi sázeli o pivo, kdo bude rychlejší.  
 friends them bet about beer who will be faster / them friends bet about beer who will be faster  
 (iii) \*Kamarádi kolegy sázeli o pivo, kdo bude rychlejší. / \*Kolegy kamarádi sázeli o pivo, kdo bude rychlejší.  
 friends colleagues bet about beer who will be faster / colleagues friends bet about beer who will be faster



into reflexive tantum verbs, decausative verbs, and the so-called derived inherent reciprocal verbs.

(B) *Reflexive tantum reciprocal verbs*. These verbs have no irreflexive counterparts (e.g., *poprat se* ‘to brawl’ and \**poprat*) or they have only a seeming counterpart, which has, however, a completely unrelated meaning (e.g., *sázet se* ‘to bet’ (8) and *sázet* ‘to plant’). The clitic with reflexive tantum reciprocal verbs has no overt semantic and/or syntactic function.

In rare cases, inherent reciprocal verbs can be either irreflexive or reflexive, without any substantial shift in their semantics or syntax, see the verb *korespondovat (si)* ‘to correspond’ in (6a,b), which can be used either with or without the reflexive clitic *si*, without any change in its meaning and/or syntactic behavior.

(C) *Decausative reciprocal verbs*. With some inherent reciprocal verbs, the clitic *se* can function as a verbal intransitivizing operator, as exemplified in (10a). These reciprocal verbs are systematically related to irreflexive inherent reciprocal verbs, representing their causative transitive counterparts, by the lexical operation of decausativization; this operation drops a causator of an event denoted by the irreflexive transitive verb. Consequently, while the causative irreflexive verbs represent object-oriented inherent reciprocal verbs (10c), decausative reflexive reciprocal verbs are subject-oriented (10b).

(10a) *Děšť se mísil se sněhem.*

rain<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> mixed with snow<sub>s+instr.sg.masc</sub>  
‘Rain mixed with snow.’

(10b) *Děšť a sníh se mísily (dohromady).*

rain<sub>nom.sg.masc</sub> and snow<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> mixed (together)  
‘Rain and snow mixed (together).’

(10c) *Maminka mísila vajíčka s cukrem.*

mother mixed eggs<sub>acc.pl.neutr</sub> with sugar<sub>s+instr.sg.masc</sub>  
‘Mother mixed eggs with sugar.’

(D) *Derived inherent reciprocal verbs*. These verbs represent a specific type of inherent reciprocal verbs, as discussed by Dimitriadis (2004), Siloni (2001) and Evans (2008) under the term *discontinuous reciprocal verbs*. In Czech they can be derived from both transitive and ditransitive verbs without the feature of symmetry in their lexical meaning by the lexical operation of reciprocalization; this operation consists in the use of the derivational morphemes *se* or *si*, which intransitivize the respective verbs. See, for example, the reciprocal verbs *políbit se* ‘to kiss’ in (11b) derived by the clitic *se* from the transitive verb *políbit* ‘to kiss’ (11a) and *vyprávět si* ‘to tell (something to each other)’ in (12b) derived by the clitic *si* from the ditransitive verb *vyprávět* ‘to tell (something to somebody)’ (12a).<sup>6</sup> As discussed in Dimitriadis (2004), derived inherent reciprocal verbs exhibit specific syntactic properties. The participants of these verbs – despite being involved in symmetry – always remain expressed in separate syntactic positions determined by the respective complementations: one participant is always mapped onto valency complementation expressed in the subject position, while the other corresponds to the complementation which has the comitative prepositional form *s+Instr* ‘with+Instr’, see (11b) and (12b). Dimitriadis (2004) argues that the meaning of derived inherent reciprocals is necessarily symmetrical.

6 Neither of the base verbs *políbit* ‘to kiss’ (11a) and *vyprávět* ‘to tell (something to somebody)’ (12a) express symmetry in their lexical meaning, but allow their participants to stand in symmetry as a result of the syntactic operation of reciprocalization, see Section 2.2.1.

(11a) Petr políbil Marii.

Petr<sub>nom.sg.masc</sub> kissed Mary<sub>acc.sg.fem</sub>  
 ‘Peter kissed Mary.’

(11b) Petr se políbil s Marií.

Peter<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> kissed with Mary<sub>s+instr.sg.fem</sub>  
 ‘Peter kissed with Mary.’

(12a) Jan vyprávěl Pavlovi strašidelné history.

John<sub>nom.sg.masc</sub> told Paul<sub>dat.sg.masc</sub> spooky stories  
 ‘John was telling Paul spooky stories.’

(12b) Jan si vyprávěl s Pavlem strašidelné historky.

John<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> told with Paul<sub>s+instr.sg.masc</sub> spooky stories  
 ‘John and Paul were telling spooky stories to each other.’

### 2.1.2 Representation of Inherent Reciprocal Verbs in the VALLEX Lexicon

Inherent reciprocal verbs are represented in the lexicon by lexical units contained in separate lexemes introduced by their respective lemmas; if a verb is cliticized, its lemma includes the respective clitic *se* or *si*. As discussed in Section 2.1.1, the cliticization of inherent reciprocal verbs is either characteristic of reflexive tantum verbs (B), or it can be a result of lexical operations deriving either decausative verbs (C), or inherent reciprocal verbs (D).

A specific attribute “reciprverb” attached to lexical units corresponding to reciprocal verbs provides information on their type. In case of inherent reciprocal verbs of all types (A)-(D), this attribute has the value “inherent”. To each inherent reciprocal verb where the clitic *se* or *si* functions as a derivational means, the specific attribute “derived” is attached, recording information as to whether the given verb is derived by the lexical operation of decausativization (the value “decaus”), or by lexical reciprocalization resulting in derived inherent reciprocal verbs (the value “lex-reciprocal”). In the VALLEX lexicon, 201 lexical units of verbs in total are annotated as inherent reciprocal verbs (109 out of them have irreflexive lemmas, 33 represent reflexive tantum verbs, 29 verbs with derived reflexive lemmas are decausative verbs, and 30 represent derived inherent reciprocal verbs), see Table 1.

Table 1. Reciprocity in VALLEX – basic statistics (LUs stands for lexical units).

Reciprocal verbs in VALLEX			joint LUs (subject/object-oriented)	distributed LUs (subject/object-oriented)
Inherent reciprocal verbs	201 LUs (281 verb lexemes)	all types	197 (137 / 60)	4 (4 / 0)
		(A) irreflexive verbs	103 (47 / 56)	0
		(B) reflexive tantum verbs	33 (32 / 1)	0
		verbs with ir/reflexive	6 (3 / 3)	0
		(C) refl. decausative verbs	29 (29 / 0)	0
		(D) derived inherent reciprocal verbs	26 (26 / 0)	4 / 0
Syntactic reciprocal verbs	1,923 LUs (2,017 verb lemmas)	all types	613 (31.9%)	1,310 (68.1%)
TOTAL	2,124 LUs		810 (38.1%)	1,314 (61.9%)

In addition, the attribute “recipvent” describes whether reciprocal verbs refer to a joint action in which the participants involved act symmetrically (the value “joint”, e.g. *vyjednávat* ‘to negotiate’, *mluvit (s někým)* ‘to talk (with somebody)’, *oddělit* ‘to separate’), or to a plurality of actions where each single action is asymmetrical (the value “distributed”, e.g. *udávat* ‘to report each other’). The annotation of inherent reciprocal verbs reveals that these predominantly express joint actions. Surprisingly, reciprocal events that are denoted by a small number of derived inherent reciprocal verbs in the annotated data can be interpreted as a series of asymmetrical actions (e.g., *navštěvovat se* ‘to visit’ and *vyprávět si* ‘to tell’), c.f. (Dimitriadis 2004).

Finally, each relevant lexical unit is assigned a specific attribute “recipr” providing pairs of those valency complementations that are involved in reciprocity and which can be thus reciprocalized, as is discussed in more detail in Section 3. See Figure 2, displaying the lexical entry of the derived inherent reciprocal verb *políbit se* ‘to kiss’ (right).

<p><b>políbit<sup>pf</sup></b></p> <p>① dát někomu polibek 'to kiss somebody'</p> <p>-frame: ACT<sub>1</sub> PAT<sub>4</sub> LOC BEN</p> <p>-example: Jan políbil Marii na tvář. Políbil ji na ruku. Políbil tvář krásné dívky. 'John kissed Mary on her cheek. He kissed her on her hand. He kissed a cheek of a nice girl.'</p> <p>-reflex: ACT-PAT Políbila se na ruku. 'She kissed herself on her hand.' ACT-BEN Políbila si ruku. 'She kissed a hand to herself.'</p> <p>-recipr: ACT-PAT Jan a Marie se políbili na tvář. John and Mary REFL kissed on cheek 'John and Mary kissed each other (on their cheeks)'</p> <p>-recipvent: distributed</p> <p>-recipverb: gram</p> <p>-class: contact</p> <p>-diat: deagent, passive-být</p>	<p><b>políbit se<sup>pf</sup></b></p> <p>① dávat si s někým polibek; líbat se s někým (navzájem) 'to kiss each other'</p> <p>-frame: ACT<sub>1</sub> PAT<sub>s+7</sub> LOC</p> <p>-example: Jan se políbil s Marií. John REFL kissed with Mary 'John kissed with Mary.'</p> <p>-recipr: ACT-PAT</p> <p>-recipvent: joint</p> <p>-recipverb: inherent</p> <p>-derived: lex-reciprocal</p> <p>-class: contact</p>
---	---

Figure 2. Two lexical entries of the verbs *políbit* ‘to kiss’ and *políbit se* ‘to kiss’, representing different types of reciprocal verbs, as they are described in the VALLEX lexicon.

## 2.2 Syntactic Reciprocal Verbs

### 2.2.1 Types of Syntactic Reciprocal Verbs

Syntactic reciprocal verbs are those for which the lexical meaning does not imply symmetry; however, they allow their participants to be put into symmetry (e.g., the verbs *podezírat* ‘to suspect’, *řadit* ‘to arrange’ and *políbit* ‘to kiss’) (Panevová & Mikulová 2007; Siloni 2008; Evans 2008). This symmetry is achieved by the syntactic operation of reciprocalization which, when applied to the given verbs, results in reciprocal constructions. In contrast to inherent reciprocal verbs (Section 2.1), Czech syntactic reciprocal verbs represent an open group of verbs with various semantic and syntactic properties, comprising intransitive (13), transitive (14) and ditransitive verbs (15). Similar to in the case of inherent reciprocal verbs, one participant which can be involved in symmetry is expressed in the subject position in the nominative, and the other participant occupies either the direct object position expressed in the accusative (14), or the indirect object position (13), which can have various forms. In rare cases, syntactic reciprocal verbs realize reciprocity between participants when one of them is mapped onto the direct object position in the accusative and the other corresponds to the indirect object position expressed in various forms (16).

(13) Petr se dívá na Marii.

Peter<sub>nom.sg.masc</sub> REFL<sub>verb.lemma</sub> looks at Mary<sub>na+acc.sg.fem</sub>  
'Peter is looking at Mary.'

- (14) Petr políbil Marii.  
 Petr<sub>nom.sg.masc</sub> kissed Mary<sub>acc.sg.fem</sub>  
 ‘Peter kissed Mary.’
- (15) Petr podezíral manželku z nevěry.  
 Petr<sub>nom.sg.masc</sub> suspected wife<sub>acc.sg.fem</sub> from infidelity.  
 ‘Peter is suspecting his wife of infidelity.’
- (16) Musíte konfrontovat sen s realitou.  
 (you) have to confront dream<sub>acc.sg.masc</sub> with reality<sub>s+instr.sg.fem</sub>  
 ‘You have to confront your dream with the reality.’

(A) Most syntactic reciprocal verbs are characterized by *irreflexive* lemmas.

(B) Those syntactic reciprocal verbs that have reflexive lemmas can first represent *reflexive tantum* verbs, verbs without irreflexive counterparts (e.g., *podívat se* ‘to look’, *postěžovat si* ‘to complain’, *hledět si* ‘mind’), or verbs with semantically unrelated irreflexive counterparts (e.g., *chovat se* ‘to behave’), with which the clitics *se* or *si* represent an obligatory part of their verb lemmas. Second, with a small number of verbs, the reflexive clitic *se* or *si* is an optional part of their verb lemmas, the use of which does not bring about any changes in meaning and/or syntactic behavior (e.g., *pamatovat (si)* ‘to remember’).

(C) Finally, the clitic *se* functions as a derivational means of *decausative* syntactic reciprocal verbs, which are derived by the lexical operation of decausativization from transitive or ditransitive syntactic reciprocal verbs (e.g., *nakazit se* ‘be infected’ ← *nakazit* ‘to infect’, *opřít se* ‘lean’ ← *opřít* ‘lean’, *stáhnout se* ‘to retreat’ ← *stáhnout* ‘to withdraw’). With decausative syntactic reciprocal verbs, the given clitic has the same function as with inherent reciprocal verbs (as discussed in Section 2.1.1.)

### 2.2.2 Representation of Syntactic Reciprocal Verbs in the VALLEX Lexicon

Syntactic reciprocal verbs are represented in the lexicon by the respective lexical units of verbs in lexemes headed by their respective (irreflexive or reflexive) lemmas. Syntactic reciprocal verbs of all types (A)-(C) are identified by the value “gram” of the attribute “reciprverb”.

The attribute “recipvent” describes whether a syntactic reciprocal verb – when its participants are reciprocalized – refers to a joint action (the value “joint”, e.g. *cítit spolu* ‘to sympathize (with each other)’, *bojovat* ‘to fight’, *skoncovat spolu* ‘to finish (with each other)’), or to a series of actions where each single action is asymmetrical (the value “distributed”, e.g. *kritizovat* ‘to criticize’, *nazývat* ‘to call’, *pamatovat* ‘to remember’).

In the VALLEX lexicon, out of 2,124 lexical units corresponding to reciprocal verbs, 1,923 (90.54%) represent syntactic reciprocal verbs. As the annotation revealed, syntactic reciprocal verbs predominantly express distributed reciprocal events (almost 70%), while joint reciprocal events with syntactic reciprocal verbs are rather rare. See Table 1, above.

The valency frames of syntactic reciprocal verbs stored in the VALLEX lexicon describe the usage of these verbs in unreciprocal constructions. Their reciprocal constructions can be obtained by application of rules, as described in Section 3.2 – these rules make use of further subclassification of syntactic reciprocal verbs, based on information about which pair (or triplet in rare cases) of valency complementations are involved in reciprocity. Similar as for inherent reciprocal verbs, this information is provided by the attribute “recipr”.

See the left lexical entry displaying the syntactic reciprocal verb *políbit* ‘to kiss’ in Figure 2, above.

### 3 Reciprocal Constructions in VALLEX

This section thoroughly describes the operation of syntactic reciprocalization. This operation is systematic enough to be captured by formal rules operating over the information stored in the lexicon. On the basis of these rules, all possible morpho-syntactic manifestations of both inherent and syntactic reciprocal verbs can be obtained. As these rules represent an economic and systematic way of language description, they are included in the lexicon.

#### 3.1 Reciprocal Constructions of Inherent Reciprocal Verbs

Inherent reciprocal verbs express symmetry in each of their instances (see Section 2.1), whether syntactic reciprocalization is applied to them or not. In case this syntactic operation is used, the participants involved in symmetry are not expressed in separate syntactic positions, but instead fill a single syntactic position of either subject (with subject-oriented inherent reciprocal verbs), or direct object (with object-oriented verbs of the given type) (derived inherent reciprocal verbs being the only exception, see Section 2.1.1, type (D)). As a result, the given syntactic position is plural, expressed either by coordinating (17b) or subordinating coordination (17c), or by morphological (18b), or semantic plural (17d).

The less prominent syntactic position determined by the other valency complementation involved in reciprocity, typically the position of indirect object, is either deleted from the surface, or is filled with the reflexive pronoun, depending on the form of the given complementation. Two forms are typical of this complementation with inherent reciprocal verbs: the comitative form *s+Instr* ‘with+Instr’ and the prepositional form *od+Gen* ‘from+Gen’. In the first case, the indirect position is removed, compare (17a) with (17b), while in the latter case the indirect object is occupied by the respective form of the reflexive pronoun *sebe*, compare (18a) and (18b). Alternatively, in both cases the less prominent position can be filled with the quantifier-like bipartite expression *jeden druhý* ‘each other, lit. one other’, see e.g. (Evans 2008) – here the first part *jeden* ‘each’, usually referred to as “range argument” (Heim et al. 1991), has the form of the nominative (with subject-oriented verbs) (17e), or the accusative (with object-oriented verbs) (18c), while the second part *druhý*, referred to as “contrast argument”, is inflected for the case (prepositionless or prepositional) as the given complementation prescribes; both parts have the singular form and exhibit the agreement in gender with the reciprocalized participants, compare (17e) with (17a) on the one hand and (18c) with (18a) on the other.

Further, reciprocity can be optionally emphasized by the adverbial modifiers *navzájem*, *vzájemně*, ‘mutually’. Moreover, with inherent reciprocal verbs with the indirect object in the form *s+Instr* ‘with+Instr’, the modifiers *spolu* ‘together’ or *mezi sebou* ‘between each other’ can be used as well.

See Figure 3, exemplifying the rules capturing morpho-syntactic properties of reciprocal constructions.

(17a) Petr se hádá s Pavlem.

Peter<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> quarrels with Paul<sub>s+instr.sg.masc</sub>  
 ‘Petr is quarreling with Paul.’

(17b) Petr a Pavel se (spolu) hádají.

Peter<sub>nom.sg.masc</sub> and Paul<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> (together) quarrel  
 ‘Petr and Paul are quarreling (together).’

(17c) Petr s Pavlem se (spolu) hádají.

Peter<sub>nom.sg.masc</sub> with Paul<sub>s+instr.sg.masc</sub> REFL<sub>verblemma</sub> (together) quarrel  
 ‘Petr with Paul is quarreling (together).’



- (17d) Družstvo se hádá.  
 Team<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> quarrels  
 ‘The team is quarreling.’
- (17e) Petr a Pavel se hádají jeden s druhým.  
 Peter<sub>nom.sg.masc</sub> and Paul<sub>nom.sg.masc</sub> REFL<sub>verblemma</sub> quarrel one<sub>nom.sg.masc</sub> with other<sub>s+instr.sg.masc</sub>  
 ‘Peter and Paul are quarreling.’
- (18a) Lékaři oddělili dvojče od dvojčete.  
 surgeons separated twin<sub>acc.sg.neutr</sub> from twin<sub>od+gen.sg.neutr</sub>  
 ‘Surgeons separated a twin from the other twin.’
- (18b) Lékaři oddělili dvojčata od sebe.  
 surgeons separated twins<sub>acc.pl.neutr</sub> from REFL-long<sub>od+gen</sub>  
 ‘Surgeons separated twins from each other.’
- (18c) Lékaři oddělili dvojčata jedno od druhého.  
 surgeons separated twins one<sub>acc.sg.neutr</sub> from other<sub>od+gen.sg.neutr</sub>  
 ‘Surgeons separated twins from each other.’

Basic rule for subject-oriented reciprocity (change of verb form, agreement, form for ACT)		
conditions:	recipr: <b>ACT-X</b> form: ACT(nom)	
actions:	agreement: number+gender+person, ACT change form of ACT: * → nom : plural	

Basic rule for object-oriented reciprocity (change of form for PAT)		
conditions:	recipr: <b>PAT-X</b> or <b>X-PAT</b> form: PAT(acc)	
actions:	change form of PAT: 4 → acc : plural	

Additional rules for all types of reciprocity (form of the reflexive pronoun)			
original form		reciprocal form(s)	reflexive pronoun
dat	→	si / sobě sobě / Ø	for irreflexive verbs: clitic / long form, dative for reflexive verbs: long form / unexpressed, dative
acc		se / sebe	clitic / long form, accusative
gen		sebe / Ø	long form, genitive / unexpressed
s+instr 'with'		Ø	unexpressed
preposition + case		long form of the reflexive pronoun in the respective prepositional case	

Figure 3. Examples of the simplified rules capturing the morpho-syntactic properties of reciprocal constructions for both inherent and syntactic reciprocal verbs in the VALLEX lexicon (two rules are successively applied to the relevant valency frames).

### 3.2 Reciprocal Constructions of Syntactic Reciprocal Verbs

Syntactic reciprocalization with syntactic reciprocal verbs is a productive process. In Czech, there are only few restrictions for its application concerning the semantic homogeneity of participants and their status with respect to topic-focus articulation (Panevová 1999).

In reciprocal constructions of syntactic reciprocal verbs, the participants involved in symmetry are obligatorily reciprocalized. The reciprocal structure is thus characterized by a plural subject (with

subject-oriented syntactic reciprocal verbs) (19b), or plural object (with object-oriented syntactic verbs) (20b). The less prominent syntactic position of the complementation involved in reciprocity is filled with the reflexive pronoun, which can have either the clitic or long form, the morphemic case of which is determined by the given complementation, compare (19a) with (19b). In rare cases, if this complementation has the comitative form *s+Instr* ‘with+Instr’, it is not expressed on the surface. Moreover, the less prominent position can also be filled with the quantifier-like bipartite expression *jeden druhý* ‘each other’, and the same morphological marking as in reciprocal structures with inherent reciprocal verbs applies (see Section 3.1).

In reciprocal constructions marked by the reflexive pronoun, the adverbial modifiers *vzájemně*, *navzájem* ‘mutually’, or in a limited cases also the modifiers *spolu* and *dohromady* ‘together’, can further emphasize the reciprocal meaning. In case of ambiguity with reflexive constructions, these modifiers have a disambiguating function. See, for example, the construction with the verb *obviňovat* ‘to accuse’ (21a), which can have either reciprocal interpretation (21b), or reflexive interpretation (21c).

See Figure 3 above, exemplifying the rules capturing morpho-syntactic properties of reciprocal constructions.

(19a) Marie políbila Janu.

Mary<sub>nom.sg.fem</sub> kissed Jane<sub>acc.sg.fem</sub>  
‘Mary kissed Jane.’

(19b) Marie a Jana se políbily. ≈ Sebe Marie a Jana políbily.

Mary<sub>nom.sg.fem</sub> and Jane<sub>nom.sg.fem</sub> REFL-clitic<sub>acc</sub> kissed ≈ REFL-long<sub>acc</sub> Mary<sub>nom.sg.fem</sub> and Jane<sub>nom.sg.fem</sub>  
kissed  
‘Mary and Jane kissed each other.’

(20a) Dítě řadí obrázek k obrázku.

child arranges picture<sub>acc.sg.masc</sub> to picture<sub>k+dat.sg.masc</sub>  
‘The child arranges a picture with another picture.’

(20b) Dítě řadí obrázky k sobě.

child arranges pictures<sub>acc.pl.masc</sub> to REFL-long<sub>k+Dat</sub>  
‘The child arranges pictures with each other.’

(21a) Hráči se obviňují.

players REFL-clitic<sub>acc</sub> accuse  
‘The players accuse each other/themselves.’

(21b) Hráči se obviňují navzájem.

players<sub>nom.pl.masc</sub> REFL-clitic<sub>acc</sub> accuse mutually  
‘The players accuse each other.’

(21c) Hráči obviňují sami sebe.

players<sub>nom.pl.masc</sub> accuse alone<sub>acc.pl.masc</sub> REFL-long<sub>acc</sub>  
‘The players accuse themselves.’

*Haplology with Czech reciprocal verbs.* In reciprocal constructions formed by syntactic reciprocal verbs with reflexive lemmas, both the clitics *se* and *si* are subject to haplology in cases when reciprocity is marked by the clitic form of the reflexive pronoun (Petkevič 2013; Rosen 2014). In the case of haplology, the single occurrence of the reflexive clitic *se* or *si* is associated with both the reflexive pronoun and the reflexive morpheme representing a part of a verb lemma. For example, in the reciprocal construction of the reflexive tantum verb *stěžovat si* ‘to complain’ in (22), a single occurrence

of the clitic *si* represents both the verb lemma and the reflexive pronoun. In the case of haplology, reciprocity is obligatorily marked by the adverbs.

- (22) Otec a matka si navzájem stěžují na synovo chování.  
 father and mother REFL-clitic<sub>verb/lemma/reflpron</sub> mutually complain about son's behavior  
 'Father and mother are complaining to each other about their son's behavior.'

## 4 Conclusion

In this paper, we have proposed a theoretically adequate and economical description of Czech reciprocal verbs in the valency lexicon of Czech verbs, VALLEX. We have demonstrated that, for this purpose, three-fold information on the type of reciprocal verbs, on the type of reciprocal events they denote, and on valency complementations that are involved in reciprocity, is sufficient for their adequate description. Such a formalized representation of reciprocity allows the user (being it a human or computer) to generate well-formed reciprocal structures of the relevant lexical units of Czech verbs.

## 5 References

- Dalrymple, M., Kanazawa, M., Kim, Y., Mchombo, S. & Peters, S. (1998). Reciprocal Expressions and the Concept of Reciprocity. In *Linguistics and Philosophy*, 21, pp. 159-210.
- Dimitriadis, A. (2004). *Discontinuous Reciprocals*. Ms., Utrecht Institute of Linguistics OTS.
- Evans, N. (2008). Reciprocal constructions: Towards a structural typology. In E. König, V. Gast (eds.) *Reciprocals and Reflexives: Theoretical and Typological Explorations*. Berlin: Mouton de Gruyter, pp. 33-103.
- Evans, N., Gaby, A., Levinson, C. S. & Majid, A. (eds.) (2011). *Reciprocals and Semantic Typology*. Amsterdam / Philadelphia: John Benjamins.
- Frajzyngier, Z., Curl, T. S. (eds.) (2000). *Reciprocals: Forms and Function*. Amsterdam/Philadelphia: John Benjamins.
- Haspelmath, M. (2007). Further Remarks on Reciprocal Constructions. In V. P. Nedjalkov (ed.) *Reciprocal Constructions*. Amsterdam/Philadelphia: John Benjamins, pp. 2087-2115.
- Heim, I., Lasnik, H. & May R. (1991). Reciprocity and Plurality. In *Linguistic Inquiry*, 22, pp. 63-101.
- König, E., Gast, V. (eds.) (2008). *Reciprocals and Reflexives: Theoretical and Typological Explorations*. Berlin: Mouton de Gruyter.
- König, E., Kokutani, S. (2006). Towards a typology of reciprocal constructions: focus on German and Japanese. In *Linguistics*, 44(2), pp. 271-302.
- König, E., Kuhle, A. (2007). Concepts of Reciprocity in Linguistics and other Fields. Talk at the Reciprocals Cross-linguistically Conference, Freie Universität Berlin & Utrecht Institute of Linguistics.
- Langendoen, D. T. (1978). The logic of reciprocity. In *Linguistic Inquiry*, 9(2), pp. 177-197.
- Levin, B. (1993). *English Verb Classes and Alternations*. Chicago: Chicago University Press.
- Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A. & Žabokrtský, Z. (2016). *Valenční slovník českých sloves, VALLEX*. Praha: Karolinum.
- Medová, L. (2009). Reflexive Clitics in the Slavic and Romance Languages. A Comparative View from an Antipassive Perspective. PhD thesis. Princeton University, Princeton, NJ, USA.
- Nedjalkov, V. P. (ed.) (2007). *Reciprocal Constructions*. Amsterdam/Philadelphia: John Benjamins.
- Panevová, J. (1999). Česká reciproční zájmena a slovesná valence [Czech reciprocal pronouns and valency of verbs]. In *Slovo a slovesnost*, 60, pp. 269-275.
- Panevová, J., Mikulová, M. (2007). On Reciprocity. In *The Prague Bulletin of Mathematical Linguistics*, 87, pp. 27-40.
- Petkevič, V. (2013). Formal (Morpho)Syntax Properties of Reflexive Particles *se*, *si* as Free Morphemes in Contemporary Czech. In K. Gajdošová, A. Žáková (eds.) *Natural Language Processing, Corpus Linguistics, E-learning, Proceedings of Slovko 2013, 13-15 November 2013*. Bratislava: RAM-Verlag, pp. 206-2016.

- Rosen, A. (2014). Haplogy of Reflexive Clitics in Czech. In E. Kaczmarska, M. Nomachi (eds.) *Slavic and German in Contact: Studies from Areal and Contrastive Linguistics*. Slavic Research Center, Hokkaido University, pp. 97-116.
- Reinhart T., Siloni, T. (2005). The lexicon-syntax parameter: Reflexivization and other arity operations. In *Linguistic Inquiry*, 36, pp. 389–436.
- Siloni, T. (2001). Reciprocal Verbs. In Y. N. Falk (ed.) *Proceedings of the Israel Association of Theoretical Linguistics 17, 11-12 June 2001*. Jerusalem: Hebrew University of Jerusalem, pp. 1-17.
- Siloni, T. (2002). Active lexicon. In *Theoretical Linguistics*, 28, pp. 383-400.
- Siloni, T. (2008). The Syntax of Reciprocal Verbs: An Overview. In E. König, E., V. Gast (eds.) *Reciprocals and Reflexives: Theoretical and Typological Explorations*. Berlin: Mouton de Gruyter, pp. 451-498.
- Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

## Acknowledgements

The research reported in this paper has been supported by the Czech Science Foundation (GAČR), grant No. 18-03984S *Between Reciprocity and Reflexivity: The Case of Czech Reciprocal Constructions*.

This work has been using language resources developed and/or stored and/or distributed by the *LINDAT-Clarin project* of the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2015071.

# LexBib: A Corpus and Bibliography of Metalexicographical Publications

**David Lindemann, Fritz Kliche, Ulrich Heid**

*Universität Hildesheim*

*E-mail: david.lindemann@uni-hildesheim.de, fritz.kliche@uni-hildesheim.de, heid@uni-hildesheim.de*

## Abstract

This paper presents preliminary considerations regarding objectives and workflow of LexBib, a project which is currently being developed at the University of Hildesheim. We briefly describe the state of the art in electronic bibliographies in general, and bibliographies of lexicography and dictionary research in particular. The LexBib project is intended to provide a collection of full texts and metadata of publications on metalexicography, as an online resource and research infrastructure; at the same time, LexBib has a strong experimental component: computational linguistic methods for automated keyword indexing, topic clustering and citation extraction will be tested and evaluated. The goal is to enrich the bibliography with the results of the text analytics in the form of additional metadata.

**Keywords:** bibliography, metalexicography, full text collection, e-science corpus, text analytics

## 1 Introduction

Domain-specific bibliographies are important tools for scientific research. We believe that much of their usefulness depends on the metadata they provide for (collections of) publications, and on advanced search functionalities. What is more, bibliographies for a limited domain may offer hand-validated publication metadata. As for lexicography and dictionary research, several bibliographies with different scopes and formats exist independently from each other; none of them covers the field completely, and most of them do not support advanced search functionalities, so that usability is dramatically reduced. Searches for bibliographical data and for the corresponding full texts are therefore most often performed using general search engines and domain-independent bibliography portals. However, big domain-independent repositories have two major shortcomings: They often contain noisy or incomplete publication metadata which have to be hand-validated by the users when copying them into their personal bibliographies, e. g. for citations. Closely related to that, the search functions of leading bibliography portals still focus on query-based information retrieval, since a combination of cascaded filter options using keywords and metadata such as persons, places, events, and relations to other items, only yields good results if the metadata meet certain requirements on precision and completeness.

Our goal is a domain-specific online bibliography of lexicography and dictionary research (i.e. metalexicography) which offers hand-validated publication metadata as they are needed for citations, and which in addition is complemented with the output of an NLP toolchain.

Several methods from computational linguistics produce useful results for seeking and retrieving scientific publications. For example, topic clustering has become very popular in the Digital Humanities. We suggest that assigning topics to publications provides valuable metadata for finding related work. Methods for term extraction have a similar objective. They detect text patterns (thus: terms) that are more significant in a (more specific) domain corpus than in a (more general) reference corpus.



Scientific publications usually contain a reference section. The analysis of citations is useful for the retrieval process in different dimensions. The number of citations a paper receives is an indicator of its scientific impact. Next, a citation network discloses clusters of collaborating researchers and of related work. Third, metadata on citations can be combined with other metadata in different ways; this is useful, for instance, when citation clusters are not strongly interconnected, but the corresponding authors still work on similar topics. Tools for parsing the reference sections of scientific publications (e. g. GROBID, Romary & Lopez 2015) use NLP methods because the high number of different citation styles makes the use of machine learning on text data desirable.

Section 2 discusses existing resources for lexicography and metalexicography. Section 3 details the goals of the LexBib project. Section 4 describes the NLP methods we use for providing the bibliographical items with additional metadata. In Section 5, we present some results of a study on overlaps between Lexicography and Digital Humanities, for which we have compiled the actual LexBib publication metadata and full text corpus, together with a similar collection of Digital Humanities publications.

## 2 The State of the Art

In the following subsections, we describe existing collections of full texts and/or metadata from the fields of lexicography and metalexicography in terms of scope and qualitative features, as well as the state of the art regarding online presentations of bibliographical databases.

### 2.1 Full Text and Metadata Collections of Metalexicographical Publications

For lexicography and research on dictionaries, some collections exist as printed publications or are accessible online. Table 1 contains a selective list of recently published bibliographies of (meta-) lexicography, and the list of publication metadata for the LexBib test set (see Table 3 in Section 5); for collecting publication metadata for LexBib, we focus on these resources in the first place, and also collect the corresponding full texts. Later we shall include the contents of further bibliographical data collections we might have access to, and search by ourselves for full texts and publication metadata. For the retrieval of relevant publications that have not been included in any of the existing bibliographies, we might use keywords and citation metadata extracted from our lexicography e-science corpus (*cf.* workflow description in Section 4.)

In addition to the metadata listed in Table 1, the resources differ in terms of the item types of the publications they contain. Only three resources are dedicated to dictionaries (domain: “Lex”). Regarding metalexicography (domain: “Metalex”), all resources cover scientific publications of any type (monographs, journal articles, articles in conference proceedings, book chapters, dissertations) as well as references to their containers (collective volumes such as handbooks, conference proceedings, etc.); some resources also contain references to other bibliographies. Córdoba Rodríguez’ collection is the only resource which includes relevant newspaper articles. While our LexBib test set only contains contributions in English, all other resources list articles in multiple languages; Ahumada focuses on Hispanic metalexicography which is represented in publications written mainly in Spanish.

Another feature to look at is whether the bibliographies present their contents as alphabetically ordered list or, additionally, in a thematic order. Córdoba Rodríguez groups the bibliographical data in hierarchically organized thematic blocks; EURALEX, in turn, presents its references according to approximately 125 different keywords. The items of Obelex Meta are manually keyword-indexed; these approximately 70 keywords function as filter option in the extended search interface.

Table 1: Some existing bibliographies of metalexicography.

<i>Title</i>	<i>Scope (years)</i>	<i>Scope (domains)</i>	<i>Scope (languages)</i>	<i># Items</i>	<i>Format</i>
LexBib Testset	2000-2017	Metalex	English	2,056	Structured database
EURALEX Bibliography <sup>1</sup>	1600-2010	Lex/Metalex	Multiple	1,325	Unstructured list (pub. as Wiki)
Obelex Meta <sup>2</sup>	1982-2017	Metalex	Multiple	ca. 2,000	Structured database
WLWF <sup>3</sup>	1420-2016	Lex/Metalex	Multiple	2,370	Unstructured list (pub. as PDF)
Wiegand <sup>4</sup>	1850-2014	Lex/Metalex	Multiple	33,339	Unstructured list (pub. as PDF)
Hartmann <sup>5</sup>	1930-2007	Metalex	Multiple	ca. 570	Unstructured list (pub. as PDF)
Córdoba Rodríguez <sup>6</sup>	1940-2003	Metalex	Multiple	10,192	Structured database
Ahumada <sup>7</sup>	1535-2010	Metalex	Mainly Spanish	6,560	Structured database (in progress)

Obelex Meta and the LexBib test set are available to us as structured data collections stored in relational databases. All metadata are stored as attribute-value pairs, which is a necessary condition for their processing, e.g. by algorithms for duplicate merging, or for its representation in machine-readable formats such as BibTeX or TEI-XML.

Concerning the application of computational text analysis to metalexicographical full text collections, the lexicography community can already count on several studies that show the usefulness of this kind of methodology, including term extraction and bibliometrics, for depicting trends in our discipline (De Schryver 2009, 2012; Lew & De Schryver 2014).

## 2.2 Features of Bibliographical Databases as Online Resources

As an example of a state-of-the-art bibliographical database we may cite DBLP, an online bibliography of computer science<sup>8</sup> maintained at Trier University (Ley 2002; Weber et al. 2006). Features of DBLP relevant as a guiding reference for a resource like LexBib are its data model which includes indices for journals and conferences, TOC (table of contents) pages for single volumes, the disambiguated person index and individual author pages, and the data presentation, that, in addition to query-based access, allows multi-layered browsing and faceted search. Third, all DBLP bibliographic records are accessible in multiple formats via an API, so that personal reference managers (e. g. *Zotero*) can take advantage of downloading metadata sets individually as well as in bulk. As a fourth point we may add that advanced search and visualization tools exist that use data retrieved from the DBLP API (Burch et al. 2015), and that could be fed with bibliographical data compliant to that format. A fifth guiding feature of DBLP that also matches to LexBib is its limited and well-defined scope as a specialized bibliography for one discipline. This is a condition for a resource that should stay small enough to be maintained noise-free by manual validation, and it limits the problem of irrelevant results in information retrieval, two problems that doubtlessly reduce the usability of global academic search engines.

1 The EURALEX bibliography is accessible at <http://euralex.pbworks.com>.

2 See Möhrs (2016). Accessible at <http://www.owid.de/obelex/meta>.

3 Bibliography accessible to the editors of WLWF (Wiegand et al. 2010; 2017).

4 Wiegand (2006–2015): 'Internationale Bibliographie'.

5 Accessible at <http://euralex.pbworks.com/f/Hartmann+Bibliography+of+Lexicography.pdf>.

6 Accessible at <http://www.udc.es/grupos/lexicografia/bibliografia/index.html>.

7 I. Ahumada (ed.) (2006–2014): *Diccionario bibliográfico de la metalexicografía del español*. Starting with Volume III (2006–2010) and backwards, this work is being transformed into a structured database (cf. Porta Zamorano 2016).

8 DBLP is accessible at <http://dblp.dagstuhl.de>.

In addition to unique identifiers and to standard publication metadata, i.e. the bibliographic data necessary for citing or referencing, some online repositories have started to perform citation extraction and semantic annotation of items using computational text analysis, and to provide the results as metadata for display and advanced search options (*cf.* e.g. the discussion in Zeni et al. 2007). In a future stage of the LexBib project, we aim at generating metadata of this kind, using an e-science corpus consisting of abstracts and full texts of publications in the domain of metalexicography as showcase.

### 3 Goals

The goals of the LexBib project can thus be described in an infrastructural and in a research dimension. On the one hand, it is our aim to provide an online bibliography of (meta-)lexicography that meets with the state of the art as described in Section 2.2. On the other hand, we will set up, test and evaluate a pipeline of NLP tools for citation extraction, and automatic keyword indexing, and it is our intention to include the results in the published version of the LexBib collection, marked as automatically generated publication metadata.

In general terms, the tasks which a user of an online bibliography might want to perform, and that the metadata-based searches we want to offer in LexBib shall consider, may be the following, among others (list adapted from Buch et al. 2015: 163):

- Papers with certain words or substrings in their titles, abstracts, and/or text bodies;
- Papers published in certain time frames, by certain persons, published in certain journals or presented at certain events;
- Keywords relevant for a specific time interval, list of authors, and subset of the bibliography (e.g. a conference series or an event);
- Keywords co-occurring with other words (multiword term candidates);
- Frequency distributions of correlated keywords presented in their distribution over time;
- Keyword correlations and citation relations between several authors;
- Author correlations and their change over time.

To this end, interactive (browsable) visualizations of keyword and author relations shall be created and made accessible as part of the LexBib online resource.

As for publication metadata to be included in LexBib, the intended minimal coverage for each item includes all metadata necessary for citing (*cf.* Section 4.1), as well as unique identifiers of publications (ISBN, DOI) and persons (ORCID), and item relations such as “is review of” and “is reviewed in” for reviews, “is part of” and “contains” for volumes, and “citing” and “is cited by”, regarding citations.

We are aware that the intended manual validation of publication metadata is a labor-intensive task, and we foresee a considerable amount of manual editing work, which we will track in detail in order to draw conclusions on how much manual work is necessary for a noise-free collection of bibliographical data. This kind of process metadata evaluation on the relatively limited LexBib e-science corpus may yield valuable hints for possible applications of the proposed workflow to larger e-science corpora. In a first phase, we propose to consider only items in English published between 2000 and 2017, and to move on towards other languages after an intermediate evaluation of the workflow, and later back to the past.

## 4 Methods

LexBib documents are provided with additional metadata on two dimensions: Publication metadata and metadata on the contents. Publication metadata are collected together with the full texts by semi-automatic means using the Zotero tool,<sup>9</sup> and they are manually validated (see Section 4.1). For retrieving contents metadata, PDF or HTML full texts are processed with the NLP pipeline described in Section 4.2.

### 4.1 Publication Metadata

As publication metadata, a predefined minimal metadata set is collected and hand-validated for every publication, including author, title, publishing year, name of the publication (e.g. the journal), place, DOI/ISBN, etc. Publication metadata include authors, their affiliations, the title of the publication, as well as document metadata like the source or the publishing year, which can be easily retrieved. The metadata of the LexBib test set collection that exists since 2018 will be merged with data from the resources listed in Table 1, as soon as they are available in or have been converted into a structured format (e.g. TEI-XML or BibTeX), in order (1) to obtain the intersecting set, i.e. duplicate items, for semi-automated metadata validation, (2) to enrich LexBib, and (3) to allow cross-resource referencing, i.e. to be able to point exactly to where an item appears in another bibliography.<sup>10</sup>

Regarding the use of some of the resources to be merged, practical and licensing issues will have to be addressed. In case an enrichment of LexBib will not be possible because of licensing issues, only cross-resource referencing is planned; nevertheless, also for that purpose, the publication metadata items to be referenced have to be accessible in a structured format.

### 4.2 Full Text Processing and Content Metadata

Full texts are cleaned and processed in the following way (see pipeline schema in Figure 1): Both PDF and HTML files are converted into plain text. The full text bodies are isolated, processed with the TreeTagger (Schmid, 1994) for part-of-speech tagging and lemmatization, which makes them accessible to topic clustering on lemmas and to term extraction. For citation extraction, the list of bibliographic references at the end of each full text is isolated and parsed (see Section 4.2.2).

#### 4.2.1 Term Extraction and Topic Clustering

The full text content is converted into a lemmatized variant and processed with Mallet (McCallum 2002) for topic clustering. For each publication, Mallet provides a measure of how it is related to the different topics. Our goal is to use these assignments as LexBib metadata. The idea is to provide an access structure towards the items in the bibliography by browsing topics.

For term extraction, we use a tool suite developed at IMS (University of Stuttgart, cf. Rösiger et al. 2015; 2016). It extracts the instances of part-of-speech patterns, e. g. (1) NN (single common nouns), (2) NN-NN (two common nouns), or (3) NN-NN-NN (three adjacent common nouns). Then, it ranks the extracted instances according to their *termhood* or *keyness* which is measured by dividing the relative frequency of the instance in a text by the relative frequency of the instance in a reference corpus (*weirdness ratio*, cf. Ahmad et al. 1992). We run this method twice for each document; once with the British National Corpus (BNC) as a reference corpus in general language in order to retrieve

<sup>9</sup> See <http://zotero.org>.

<sup>10</sup> This can be useful, for example, if an item carries further information in the other resource, e.g. a short review, as in Wiegand (2006-2015).

domain specific terms; and once with the whole LexBib corpus as a reference corpus in order to identify text specific keywords. As an example of this procedure, the terms extracted from a Euralex 2014 keynote speech (Heid 2014) are given in Table 2. In the left column, they are ordered according to their keyness relative to the BNC, in the right column according to their keyness in comparison to the LexBib full text test set. It can be observed that terms relevant to Lexicography in general are ranked lower in the right column.

Table 2: Term candidates extracted from an example publication.

<i>Top 20 Terms (Ref. BNC)</i>	<i>Top 20 Terms (Ref. LexBib)</i>
text reception	information-on-demand
text production	on-demand
dictionary function	data repository
multiword	user friendliness
word formation	user orientation
production dictionary	text reception
user orientation	production dictionary
information-on-demand	text production
data repository	dictionary function
dictionary entry	repository
user friendliness	valency
internet	guidance
markup	orientation
on-demand	concord
valency	word formation
concord	scenario
dictionary	markup
corpus	production
collocation	classification
language processing	advance

In addition to manual revision and assessment of term candidates, we plan the evaluation of the term extraction results using the manually defined keywords given in Obelex Meta (Möhre 2016) as a silver standard, in order to obtain data for adapting the performance of the automatic keyword indexing methods. A further possible application of the latter is a revision of the set of keywords used for indexing Obelex Meta items, and a grounding of term variants, i.e. an association of different variants of a term to a single keyword regarded as the canonical form or base variant (see discussion and methodology in Theofilidis 2018).



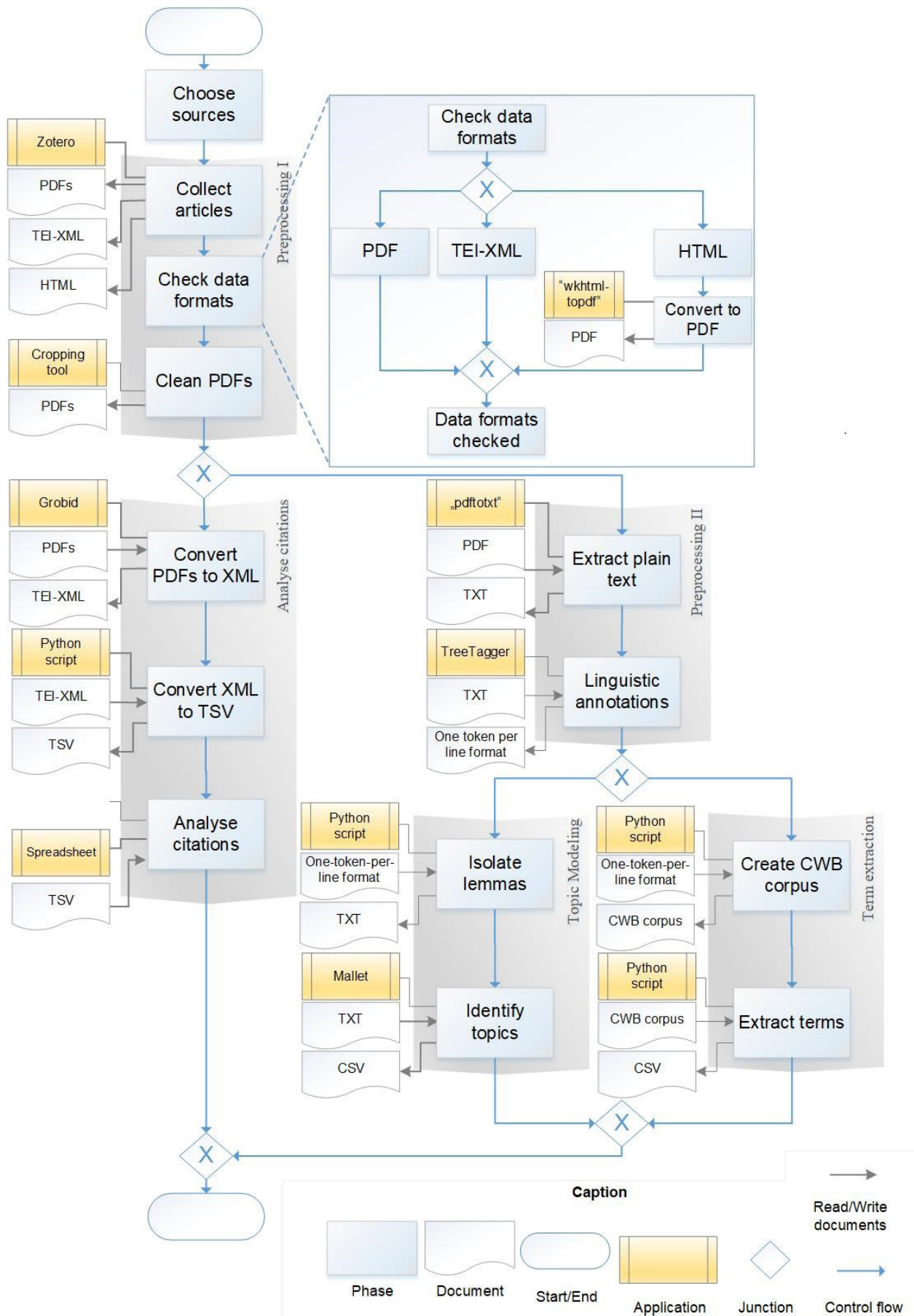


Figure 1: Workflow for LexBib corpus building and nlp processing.

#### 4.2.2 Citation Network

The item relations obtained from the analysis of the reference sections in the full texts include (1) the publications cited in a publication, (2) the publications citing a publication, and (3) the membership of a publication in a cluster of a citation network. The GROBID tool (Lopez 2009) extracts a plain text version of the full text content and isolates the block of bibliographic references, the entries of which are then parsed and converted into a structured format compliant to the TEI guidelines (element `<listBibl>`). GROBID uses conditional random fields, a supervised machine learning method which learns a model based on annotated training data. Problematic citation styles, i.e. formats that are not properly parsed by the tool, will require further annotated training data. As a by-product, GROBID's recognition performance will be enhanced.

Based on the extracted references and the publication metadata sets, a citation network is modeled and publication clusters are identified. The analysis requires a mapping from a citation given in a publication to the metadata of the cited publication. We are aware that the metadata given in citations differ significantly. Letters with diacritics may be replaced with those without diacritics. The titles can also differ, e.g. subtitles can be left out. In our preliminary study (see Section 5), we even found a considerable amount of instances where different publishing years were given for the same publication. The deviations can be due to mistakes in the references of a publication, but they can also be caused by an erroneous output of the GROBID tool or by errors in our programming scripts.

For validating the mapping, we generate a triplet representing a citation, consisting of the last name of the first author, the publication year and the title. Authors and titles are normalized by a conversion to lower case, the reduction to the letters [a-z] (thus deleting non-alphabetic characters and whitespace) and a limitation of the normalized title to a maximum length of 40 characters. For example, the triplet representing our present paper is "lindemann\_2018\_lexbibacorporisandbibliographyofmetalexico". The mapping is considered valid if one of several validity categories are fulfilled; if, for example, three triplets are found where the (non-normalized) Levenshtein distance of the titles is  $\leq 8$ ; the Levenshtein distance of the authors is  $\leq 2$ , and the publishing years may differ by one year. Note that this restriction implies that documents with less than three citations are filtered out from the citation analysis.

## 5 Preliminary Experiments

For a study on the overlap of topics and citations between Digital Humanities (DH) and lexicography, an e-science corpus has been built and processed applying the methodology described in section 4. Table 3 shows the composition of the lexicography subcorpus, which is identical to the LexBib test set mentioned in section 2.1. The DH publications stem from four major DH journals and a DH handbook (see Lindemann, Kliche & Kutzner 2018 for the complete reference).

The results of the computational text analysis performed on that corpus confirm an initial hypothesis that despite a very small overlap of the citations (i.e. in spite of the fact that authors from DH and lexicography hardly cite each other), quite a wide range of overlapping topics and terms is found. Topic clustering disclosed a very significant amount of topics where publications from both disciplines are found among the publications with the highest weight for a topic; in other words: a list of topics that can be regarded as important for both DH and lexicography. We visualized the results of the topic modeling in the table-like model in Figure 2. Columns represent the topics and contain the top 100 most relevant publications for a topic. Publications from lexicography are highlighted in green; publications from DH in purple. The figure shows that for some topics the top 100 relevant publications belong (nearly) exclusively to one of the two domains, while in many other cases publications from both domains appear.

Table 3: Composition of the LexBib full text collection (test set).

<b>Journals</b>	<b>915</b>
Lexikos	376
International Journal of Lexicography	282
Dictionaries (Journal of the DSNA)	257
<b>Conference Proceedings</b>	<b>984</b>
Euralex	782
eLex	202
<b>Handbooks</b>	<b>157</b>
HSK 5/4 (Gouws et al. 2013)	110
Routledge Handbook (Fuentes Olivera 2018)	47
<b>Total</b>	<b>2,056</b>

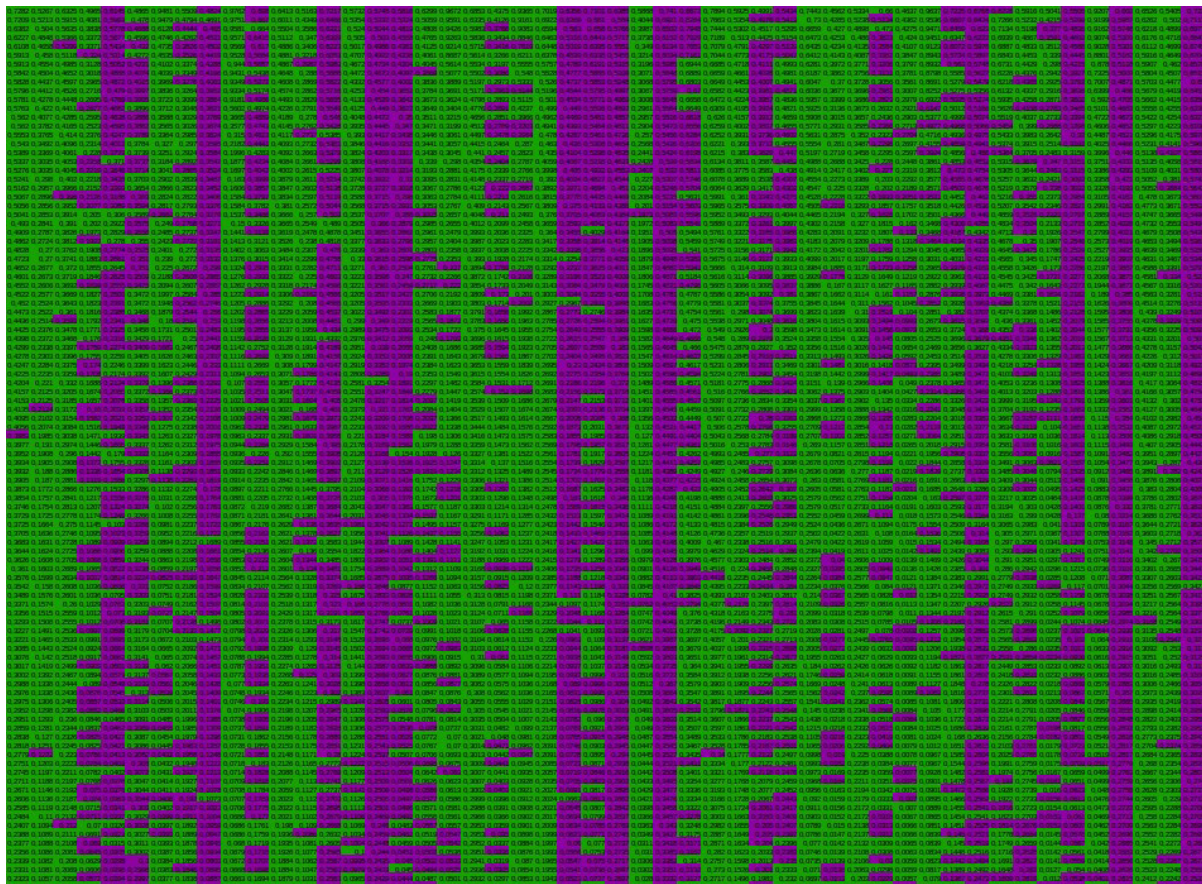


Figure 2: Visualization of topic clusters and their relevance in the DH/lexicography subcorpora.

In the following, we will focus on some details of the term extraction results. The term extraction tool produced a list of term candidates for each of the two subcorpora, Digital Humanities (DH), and Lexicography (Lexicog), ranked by their *termhood* in relation to the frequency in the reference corpus, the BNC. Table 4 lists the top 25 terms for DH, Lexicog, and their overlap, i.e. lexicography terms (out of the top 1,000) also found in the DH top 500, sorted by their termhood ratio in comparison to their frequency in the BNC.



Table 4: Term candidates extracted from the DH and Lexicography subcorpora.

<i>Top DH Terms</i>	<i>Top Lexicog Terms</i>	<i>Top LexTerms found in DH</i>
website	dictionary article	website
pdf	access structure	lemmatization
xml	dictionary user	wordnet
stemma	lemma sign	reference corpus
text mining	text reception	corpus query
authorship attribution	multiword	search engine
blog	website	internet
text classification	dictionary consultation	web site
cyberinfrastructure	word formation	web page
search engine	lexicography	text box
feature selection	corpus evidence	print version
url	text production	subcorpus
classification accuracy	lemmatization	frequency list
web page	dictionary research	crowdsourcing
web site	function theory	web interface
php	article stretch	corpus research
crowdsourcing	word sketch	language documentation
open-source	wordnet	word alignment
base text	reference corpus	hyperlink
internet	translation equivalent	Wikipedia
text categorization	dictionary information	search interface
test text	pdf	blog
text reuse	lemma list	source word
book history	dictionary making	text genre
metadata	definition	word sense disambiguation

After manually validating the 500 most salient DH terms, we performed the same term extraction procedure for every year of publication and measured the intersection of the Lexicog terms and the top 500 DH terms. As Figure 3 shows, the amount of DH terms (top 500) in the diachronically indexed Lexicog subcorpora (top 1,000 candidates for each year) shows an upward tendency. Three years appear to be the most salient ones in that respect, and these happen to be years when a conference of the eLex series has taken place.

In order to verify this observation, we had a closer look at the publications contained in the eLex conference proceedings: and indeed, the term extraction results show a higher representation of DH-relevant terms in the eLex subcorpus than in the Lexicog corpus in general (see Figure 4).

These trend analyses are only two examples of applications that imply a re-use of text analysis outcomes in the first place meant as additional publication metadata for an online bibliography; two examples of insights driven by quantitative text analysis that require a minimal effort once the lexicography e-science corpus is built and processed in the described way.

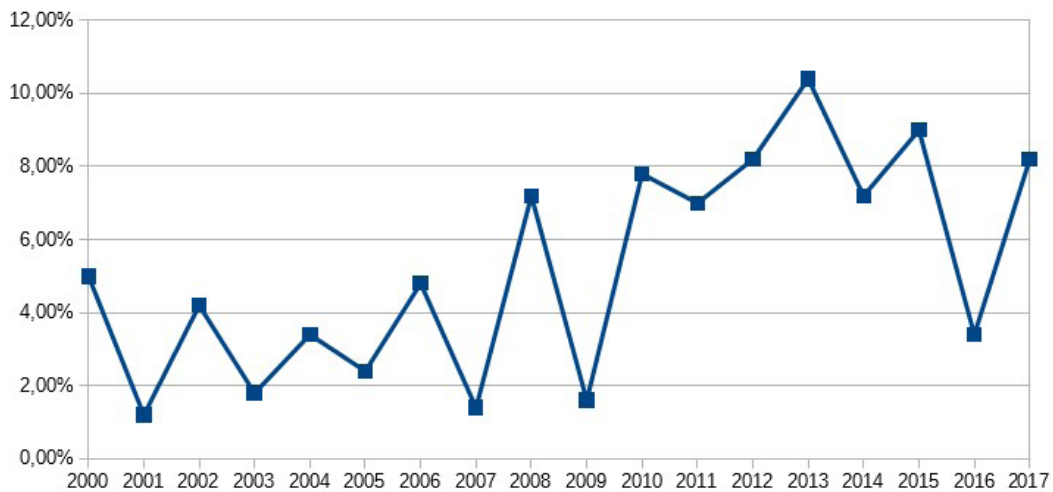


Figure 3: Overlap of term candidates from the DH and Lexicog subcorpora.

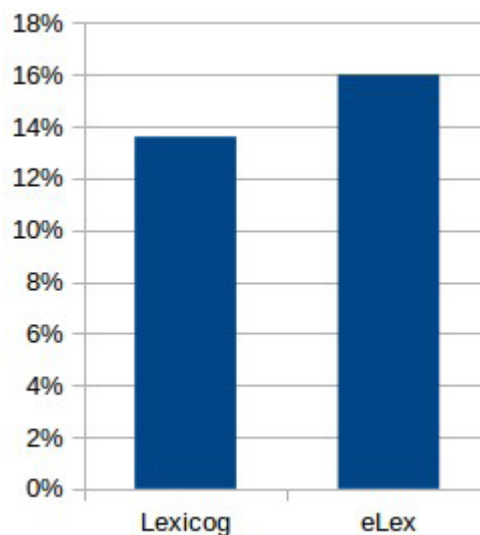


Figure 4: Overlap of DH and lexicography term candidates in the Lexicog vs. the eLex subcorpora.

## 6 Outlook

We think that lexicography is a discipline important enough as to deserve a well-structured and well-maintained bibliography as research infrastructure, and, at the same time, that it is a discipline small enough as to allow a collective reflection and a continuous evaluation of a project of this kind. The main idea is that LexBib should become a collaboratively run and widely used resource. We will call to the community for collaboration, e.g. regarding author grounding, i.e. ORCID indexing, and completion of author pages, and the evaluation of automatic keyword indexation and automatic summaries. In case the LexBib user community reaches a critical mass for introducing user generated content, we will also study the possibility of enabling user comments or discussion threads on LexBib items.



## References

- Ahmad, K., Davies, A., Fulford, H & Rogers, M. (1992). What is a term? The semi-automatic extraction of terms from text. In *Translation Studies - An Interdiscipline. Selected papers from the Translation Studies Congress, Vienna*, 267 – 278.
- Ahumada, I. (2006). *Diccionario bibliográfico de la metalexicografía del español. Vol. I: orígenes-año 2000*. Jaén: Universidad de Jaén.
- Ahumada, I. (2009). *Diccionario bibliográfico de la metalexicografía del español. Vol II: años 2001-2005*. Jaén: Universidad de Jaén.
- Ahumada, I. (2017). *Diccionario bibliográfico de la metalexicografía del español. Vol III: años 2006-2010*. Jaén: Universidad de Jaén.
- Ahumada, I. (2016). Metalexicografía del español: clasificación orgánica y tipología de los diccionarios en el Diccionario Bibliográfico de la Metalexicografía del Español (DBME). In *Anuario de estudios filológicos*, (39), 5–24.
- Burch, M., Pompe, D., & Weiskopf, D. (2015). An analysis and visualization tool for DBLP data. In *Proceedings of the 19th International Conference on Information Visualisation (iV)*, IEEE, 163–170.
- De Schryver, G.-M. & R. Lew (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography*, (27,4), 341-359.
- De Schryver, G.-M. (2012). Trends in Twenty-Five Years of Academic Lexicography. *International Journal of Lexicography*, (25,4), 464–506.
- De Schryver, G.-M. (2009). Bibliometrics in Lexicography, *International Journal of Lexicography*, (22,4), 423–465.
- Fuertes-Olivera, P. A. (Ed.). (2018). *The Routledge Handbook of Lexicography*. Routledge Handbooks in Linguistics. London: Routledge.
- Gouws, R. H., Heid, U., Schweickard, W., & Wiegand, H. E. (Eds.). (2013). *Dictionaries. An International Encyclopedia of Lexicography*. HSK 5/4. Berlin, Boston: De Gruyter Mouton.
- Heid, U. (2014). Natural Language Processing techniques for improved user-friendliness of electronic dictionaries. In A. Abel, C. Vettori, N. Ralli (Eds.), *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, 47-61
- Jacinto García, E. J. (2016). El Diccionario Bibliográfico de la Metalexicografía del Español como obra de consulta: estructura, fuentes y funciones. *Anuario de estudios filológicos*, (39), 147–169.
- Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *String Processing and Information Retrieval*, Lecture Notes in Computer Science. Presented at the International Symposium on String Processing and Information Retrieval, Berlin, Heidelberg: Springer, 1-10.
- Lindemann, D., Kliche, F. & Kutzner, K. (2018). Lexikographie: Explizite und implizite Verortung in den Digital Humanities. In G. Vogeler (Ed.), *5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. DHd 2018 - Kritik der Digitalen Vernunft, Konferenzabstracts*. Köln: Universität Köln, 257-261.
- McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Amherst, MA: University of Massachusetts. Retrieved from <http://mallet.cs.umass.edu/>
- Möhrs, C. (2016). Online Bibliography of Electronic Lexicography. The Project OBELEXmeta. In T. Margalitadze & G. Meladze (Eds.), In *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity*. Presented at the XVII International Euralex Conference, Tbilisi: Tbilisi State University, 906-909.
- Porta Zamorano, J. (2016). DBME\_3: Adquisición de datos, composición y base de datos Nebrija-Valdés. *Anuario de estudios filológicos*, (39), 349–355.
- Rösiger, I., Bettinger, J., Schäfer, J., Dorna, M., & Heid, U. (2016). Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, 41-51
- Rösiger, I., Schäfer, J., George, T., Tannert, S., Heid, U., & Dorna, M. (2015). Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries. In I. Kosem, M. Jakubiček, J. Kallas, & S. Krek (Eds.), *Proceedings of the eLex 2015 conference*. Herstmonceux Castle, United Kingdom, Ljubljana; Brighton: Trojina, Institute for Applied Slovene Studies; Lexical Computing Ltd.
- Romary, L. & Lopez, P. (2015). GROBID – Information Extraction from Scientific Publications. ERCIM News, Scientific Data Sharing and Re-use, (100).
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester.

- Theofilidis, A. (2018). Methoden der inhaltlichen Erschließung neuer Fachdomänen auf der Grundlage von Termextraktionsverfahren. In P. Drewer, F. Mayer, K.-D. Schmitz (Eds.), *Terminologie und Texte. Akten des DTT-Symposion 2018, Mannheim*. München, Karlsruhe, Köln: Deutscher Terminologie-Tag e.V., 87-97
- Weber, A., Reuther, P., Walter, B., Ley, M., & Klink, S. (2006). Multi-Layered Browsing and Visualisation for Digital Libraries. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 520-523
- Wiegand, H. E. (2006a). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 1: A-H*. Berlin, Boston: De Gruyter.
- Wiegand, H. E. (2006b). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 2: I-R*. Berlin, Boston: De Gruyter.
- Wiegand, H. E. (2007). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 3: S-Z*. Berlin, Boston: De Gruyter.
- Wiegand, H. E. (2014). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 4: Nachträge*. Berlin, Boston: De Gruyter.
- Wiegand, H. E. (2015). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 5: Register*. Berlin, Boston: De Gruyter.
- Wiegand, H. E., Beißwenger, M., Gouws, R. H., Kammerer, M., Mann, M., Storrer, A., & Wolski, W. (Eds.). (2017). *Wörterbuch zur Lexikographie und Wörterbuchforschung, Band 2*. Boston: Walter de Gruyter.
- Wiegand, H. E., Beißwenger, M., Gouws, R. H., Kammerer, M., Storrer, A., & Wolski, W. (Eds.). (2010). *Wörterbuch zur Lexikographie und Wörterbuchforschung, Band 1*. Berlin: De Gruyter.
- Zeni, N., Kiyavitskaya, N., Mich, L., Mylopoulos, J., & Cordy, J. R. (2007). A Lightweight Approach to Semantic Annotation of Research Papers. In *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science.. Berlin, Heidelberg: Springer, 61-72.



# Process Nouns in Dictionaries: A Comparison of Slovak and Dutch

**Renáta Panocová<sup>1</sup>, Pius ten Hacken<sup>2</sup>**

<sup>1</sup>Pavol Jozef Šafárik University in Košice, <sup>2</sup>Leopold-Franzens-Universität Innsbruck

E-mail: [renata.panocova@upjs.sk](mailto:renata.panocova@upjs.sk), [pius.ten-hacken@uibk.ac.at](mailto:pius.ten-hacken@uibk.ac.at)

## Abstract

Process nouns are deverbal nouns that designate the process indicated by the corresponding verb. Often, they have additional readings, such as a result reading. We present the productive mechanisms for the formation of process nouns in Slovak and Dutch. The two rules in Slovak and three rules in Dutch differ in the degree of regularity and the tendency to have additional senses.

In their process readings, process nouns are prototypical examples of what can be covered in a run-on entry, i.e. a sub-entry under the headword it is related to without a separate definition. The felicity of run-on entries depends on the regularity and predictability of the word. Some of the rules for process nouns are so regular that there is no reason to specify their output. Other rules are better suited to a representation of their output as a run-on entry, but only if the meaning is constrained to the process reading.

**Keywords:** nominalization, process nouns, process-result alternation, run-on entries, Slovak, Dutch

## 1 Introduction

Process nouns are deverbal nouns that designate the process indicated by the corresponding verb, e.g. *activation* corresponding to *activate*. The formation of process nouns is a typical example of what Dokulil (1962) calls *transposition*. Following ten Hacken (2015: 196), we define *transposition* as a process that changes the syntactic category of a word without changing its semantic category or modifying any of its semantic features. In general, transpositions are good candidates for run-on entries. Run-on entries are sub-entries with a much-reduced information content. They appear at the end of the main headword and are typically used for derivationally related words. Atkins and Rundell (2008: 236-238) give a set of conditions for the productive use of this lexicographic device, which can be summarized as the generalization that information about the words covered in run-on entries is predictable on the basis of general rules. This makes unmarked transpositions good candidates for run-on entries. In the case of process nouns, a complicating factor is that they often also have a result meaning. Thus, *translation* can refer to the activity of a translator, but also to the target text they produce. A further point to be kept in mind is the usefulness of dictionary entries. Atkins and Rundell (2008: 397-398) warn against the inclusion of words, even as run-on entries, which are of no practical value to the target user.

In this paper, we discuss the representation of process nouns in monolingual dictionaries of Slovak and Dutch. As a reference point, we use KSSJ (2003) for Slovak and van Dale (2015) for Dutch. Run-on entries are not common in these dictionaries, which is not entirely unexpected, as Svensén (2009: 132) presents them as a typical feature of the Anglo-Saxon lexicographic tradition. However, we will consider in which cases such run-on entries would be useful.

We start in section 2 with a presentation of the relevant formation rules in Slovak and Dutch. Section 3 gives examples of the coverage of process nouns in dictionaries and discusses in which contexts it would be useful to include them in run-on entries. Section 4 summarizes our recommendations.

## 2 Word formation rules for process nouns

In order to understand the representation of process nouns in Slovak and Dutch, it is first necessary to consider the different formation processes available in the two languages.

### 2.1 Process noun formation in Slovak

In Slovak, the main word formation processes forming process nouns are suffixation by *-nie/-tie* and by *-ácia/-izácia/-fikácia*. Other word formation processes are semantically marked and do not lead to the kind of equivalent meanings that we are interested in here. In line with Dokulilean tradition, the formation of deverbal nouns in Slovak is commonly referred to as *transposition*. Horecký et al. (1989: 116-117) make it explicit that further in-depth research into the transposition of verbs to their nominalized counterparts is needed to shed more light on the process of how such deverbal nouns gradually become independent lexical items in the lexicon. Transposition of deverbal nouns takes place in oral and written speech. When the resulting nominalizations are used frequently, they have the potential to become independent naming units in the lexicon. This dynamic process is determined by the need of a speech community to express nominalized actions and processes. Some examples of process nouns with *-nie/-tie* are given in (1).

- (1) a. čakanie ('waiting<sub>N</sub>') from čakať ('wait<sub>V</sub>')  
 b. stretnutie ('meeting<sub>N</sub>') from stretnúť ('meet<sub>V</sub>')

Nominalizations with *-nie/-tie* in (1a) and (1b) are available for nearly all verbs, and are often considered as part of the inflectional paradigm in Slovak grammars. The difference in the form of the suffix in (1a) and (1b) is correlated with the thematic vowel used in the infinitive form of the verb. In (1a) the thematic vowel is *-a-* whereas in (1b) it is *-ú-*. The thematic vowel determines the conjugation pattern. This dependence on the inflection class is a property that makes nominalizations of this type similar to inflection. A frequent use of nominalization of action and processes can be described as making communication more intellectual. Process nominalizations are frequent in professional communication, where they name methods, technological processes and procedures, e.g. *plastovanie karosérie* 'plasticating<sub>N</sub> of [a] car body'. It is interesting to observe that a relatively large number of process nouns are lexicalized, especially with specific collocational combinations, for instance *udelovanie cien* 'awarding<sub>N</sub> of prizes, prize award', *poskytovanie služieb* 'providing<sub>N</sub> of services, service provision'. According to Horecký et. al. (1989: 117) such lexicalization tendencies can be explained by the need of a speech community to name official and sometimes ceremonial procedures.

In (2), the variants of *-ácia* are illustrated. The suffix *-ácia* in (2a) with its variants *-izácia* in (2b) and *-fikácia* in (2c) is restricted to the [+learned] or [+international] part of the vocabulary.

- (2) a. reprezentácia ('representation') < reprezentovať ('represent')  
 b. modernizácia ('modernization') < modernizovať ('modernize')  
 c. identifikácia ('identification') < identifikovať ('identify')

As opposed to the forms with the suffix in (1), it cannot be predicted whether a form in *-ácia* exists. In the Slavic linguistic tradition the formations in (2) are labelled as *internationalisms* (e.g. Mistrík, 1976; Horecký et al., 1989; Buzássyová, 2010). It is understood that such international words can be found in a number of different languages, and they are relatively easily recognizable. The bases of such internationalisms are generally of Latin or Greek origin. The suffix *-ácia* in (2a) and its variants *-izácia* in (2b) and *-fikácia* in (2c) are non-native suffixes which attach mostly to non-native or international bases. In this domain, they compete with the native suffixes



*-nie/-tie* in (1). Occasionally *-ácia* combines with native bases resulting in formations such as *mečiarizácia*, where the first component is based on the name of Vladimír Mečiar (b. 1942, Slovak Prime Minister between 1990 and 1998). The variation in the form of the suffix is also found in the corresponding verb, as illustrated in (2b-c).

With both suffixes and their variants it is possible that the output word has other readings than a pure transposition, in particular a result reading. The result meaning of process nouns with native suffixes is illustrated in (3).

- (3) a. *premostenie* ('bridging<sub>N</sub>')  
 Myšlienka na premostenie Dunaja sa prvýkrát zrodila v roku 1402. (SNC)  
 ('The idea of bridging [the] Danube first appeared in 1402')
- b. *zateplenie* ('heat cladding<sub>N</sub>')  
 zateplenie panelových domov postavených po roku [yyyy] (SNC)  
 ('heat cladding of panel/prefabricated houses built after the year [yyyy]')
- c. *prevzdušnenie* ('aerating<sub>N</sub>')  
 Naplánovali si prevzdušnenie futbalového ihriska. (SNC)  
 ('They planned [the] aerating of [a] football playground')

In line with Horecký et al. (1989: 124), we distinguish three semantic classes of result reading, illustrated in (3). In (3a), the result reading designates an output or product of the process denoted by a verb. For instance, when it is necessary to connect two parts of the city by a bridge, the noun in (3a) is used to refer to the final product. In (3b), the noun describes the result of a procedure applied, i.e. a house with an external thermal insulation having an outer layer which prevents heat loss. In (3c), the noun implies the meaning of a resulting state in the context of soil aerating. Similarly, process nouns with international or non-native suffixes may have a result reading as given in (4).

- (4) a. *informatizácia* ('informatization')  
 b. *popularizácia* ('popularization')  
 c. *modernizácia* ('modernization')

The examples in (4) demonstrate the result meaning of particular processes. In (4a) the meaning of the noun is 'adoption of information technology', or in other words, computerization. In (4b) the noun may refer to the action or process of being popularized, and also to the result or product of this process. Similarly, (4c) can mean 'adopting or introducing modern styles, technologies, design, etc.' or the output of these processes.

In some cases, competing forms exist with the same input verb. In such a case, the form with *-nie* tends to have a process reading and the form with *-ácia* a result reading. The examples in (5) illustrate this difference.

- (5) a. *špecifikovanie* ('specifying<sub>N</sub>') - *špecifikácia* ('specification')  
 b. *preferovanie* ('preferring<sub>N</sub>') - *preferencia* ('preference')

In (5a) the noun with *-nie* emphasizes the progress of specifying something without any implication of completeness of the process denoted. Horecký et al. (1989: 280) found out that if the noun in *-ácia* is lexicalized with a result reading and a speaker intends to express process, they will prefer the formation with the native suffix *-nie*. The two nouns in (5b) differ in their stylistic use. The noun formed with *-nie* is less frequent in formal written style. It should be noted that a number of similar pairs do not display semantic differences and both forms can be used interchangeably, for instance, *popularizácia* ('popularization') and *popularizovanie* ('popularizing<sub>N</sub>').

## 2.2 Process noun formation in Dutch

The most neutral word formation rule in Dutch for the formation of process nouns is the rule that forms a neuter noun from an infinitive. As stated by de Haas and Trommelen (1993: 240), every infinitive has a nominalized counterpart. Dutch infinitives are formed by the suffix *-en*, which is reduced to *-n* after monosyllabic stems ending in a vowel, e.g. *gaan* ('go') from *ga-*. In many respects, the relation between the infinitive and its corresponding noun can be compared to the position of English *-ing* forms, as illustrated in (6).

- (6) a. Het is moeilijk deze tekst te vertalen. ('It is difficult this text to translate',  
i.e. ... to translate this text)  
b. Het vertalen van deze tekst is moeilijk. ('The translating of this text is difficult')  
c. Vertalen is moeilijk. ('Translating is difficult')

It is not always straightforward to distinguish verbal and nominal uses of the infinitive. In (6a), *vertalen* is clearly a verbal form because of infinitival *te* ('to'). In (6b), the article *het* ('the<sub>Neut</sub>') marks *vertalen* as a noun. In (6c), there are no unambiguous markers indicating the syntactic category of *vertalen* and its categorization depends on further theoretical assumptions. The meaning of the noun is restricted to the process as an abstract entity and it is not possible to pluralize it. There are very few nominalized infinitives with specialized meanings. The most striking case is *eten* ('food, meal') from *eten* ('eat<sub>v</sub>'). Here the noun does not designate a process and does not block the formation of the regular process noun, as in (7).

- (7) Het eten van bedorven fruit is gevaarlijk. ('The eating of rotten fruit is dangerous')

There are two regular suffixation processes that form process nouns as transpositions of verbs. The first has the suffix *-ing*. Some examples with the same verb as in (6) are given in (8).

- (8) a. De vertaling duurde langer dan verwacht. ('The translation took longer than expected')  
b. De vertaling van deze tekst is moeilijk. ('The translation of this text is difficult')  
c. Er staat een fout in de vertaling. ('There stands [i.e. is] an error in the translation')

According to de Haas and Trommelen (1993: 241), nominalizations with *-ing* have almost always ('vrijwel altijd') a second meaning designating the result of the action. In (8a), (*de*) *vertaling* has a process reading only and could be replaced by (*het*) *vertalen* without a change in meaning. However, (8b) is ambiguous in a sense (6b) is not. Although it requires more context to foreground this reading, (8b) can also mean that the target text is difficult to read. In (8c), the result reading is the only one available, and it is not possible to express this meaning by means of a nominalized infinitive.

As a restriction on the use of *-ing*, de Haas and Trommelen (1993: 243) mention that the formation rule tends to prefer transitive verbs. An example of a nominalization of an intransitive verb is given in (9).

- (9) a. De zitting van het parlement was succesvol. ('The session of the parliament was successful')  
b. Zij heeft zitting in deze commissie. ('She has [a] seat in this committee',  
i.e. is a member)  
c. De zitting van de stoel is kapot. ('The seat of the chair is broken')  
d. \*De zitting in de auto was lang en saai. ('The sitting in the car was long and boring')

The intransitive verb *zitten* ('sit') has a nominalization *zitting*, of which the three examples in (9a-c) show the most common meanings. Compared to the verb *zitten*, which has a very general meaning, all examples are semantically marked. For (9a), there is no paraphrase involving the verb. In (9b), *zitting hebben* is equivalent to *zitten*, but it is not possible to use *zitting* in the relevant sense without *hebben* or an alternative support verb. In (9c), the meaning is entirely idiosyncratic. It is not possible to use

*zitting* as a process nominalization corresponding to the most common sense of *zitten*, as illustrated by the ungrammaticality of (9d). Here, only the noun *zitten* can be used. Also for some transitive verbs, a process nominalization with *-ing* is impossible, e.g. *\*eting* from *eten* ('eat'), *\*schrijving* from *schrijven* ('write').

The second suffixation process that produces process nouns involves the suffix *-atie*. Some examples are given in (10).

- (10) a. De provocatie van een misdrijf door de politie is illegaal.  
       ('the provocation of a criminal offence by the police is illegal')  
       b. De lancering van een raket was een provocatie aan het adres van Amerika.  
       ('the launch of a rocket was a provocation to the address of America',  
       i.e. ... directed at America)

The noun *provocatie* in (10) is related to the verb *provoceren* ('provoke'), which in Dutch usually has only a direct object ('someone'). When the goal of the provocation is expressed, as in (10a), the verb *uitlokken* ('provoke') is a more common synonym. In (10a), *provocatie* is used as a process noun and can be replaced by a nominalized infinitive. In (10b), the result of the process is designated.

De Haas and Trommelen (1993: 258-259) do not accept a suffix *-atie*, but analyse it into *-at-* and *-ie*, an analysis also adopted by Booij (2002: 128). For the Latin *provocatio* ('appeal'), such an analysis is certainly valid. Aronoff (1994: 31-59) shows that what he calls the "third stem" is used as the base in a variety of inflectional and derivational forms. In Latin, *-t-* is the marker of the third stem and *-a-* is the thematic vowel of the first conjugation. In *repetitio* ('repetition'), the thematic vowel is *-i-* and in *productio* ('lengthening'), there is no thematic vowel. While it is possible to point out a certain number of correlations between such forms in Latin and words of Dutch, it seems unlikely that *-at-* has a role in Dutch speakers' language competence. It is only of etymological relevance.

The distribution of *-atie* is much more limited than of *-ing*. It only attaches productively to verbs with a stem in *-eer*, as in the case of *provoceren*. There is no constraint as to transitivity, cf. *variatie* ('variation') corresponding to *variëren* ('vary'). However, there are several idiosyncratic gaps. In (11) some examples of verbs in *-eren* with nominalizations are given.

- |  |              |               |
|--|--------------|---------------|
| (11) a. isoleren ('isolate')             | isolatie     | isolering     |
| b. alarmeren ('alarm <sub>v</sub> ')     | *alarmatie   | alarmering    |
| c. arresteren ('arrest <sub>v</sub> ')   | arrestatie   | *arrestering  |
| d. riskeren ('risk <sub>v</sub> ')       | *riskatie    | *riskering    |
| e. protesteren ('protest <sub>v</sub> ') | *protestatie | *protestering |

In (11a), the two nouns can both be used as process and result nouns. Van der Wouden (2018) notes that in such cases the noun in *-atie* is often more common in Belgium, and the noun in *-ering* more common in the Netherlands. In (11b), the impossibility of *\*alarmatie* can be explained by the etymology of the word in French, which precludes the formation of a French noun in *-ation*. In French, *alarme* is derived from *à l'arme*, originally an interjection 'to the arm(s)!'. The reverse situation is found in (11c), where French *arrestation* supports Dutch *arrestatie*. In (11d), only the nominalized infinitive is available, and for the result reading a syntactic description would have to be used. (11e) represents a class of cases where the verb is denominal. Whereas (11d) is correlated to *risico* ('risk'), which is not a process, in (11e), the underlying noun *protest* ('protest<sub>N</sub>') denotes the relevant process and thus blocks nominalizations with *-atie* and *-ing*. In all cases in (11), the nominalized infinitive is also available. In view of data such as these, we do not follow van der Wouden's (2018) suggestion that in cases where the *-ing* form does not exist, it may be blocked by the nominalized infinitive.

Booij (2002: 125) lists eight suffixation processes that can be used for nominalization. However, apart from *-ing* and *-atie*, they are either non-productive (e.g. *winst* from *winnen*, ‘win, gain’) or they are not used for transpositions in the sense we adopt this term here, e.g. *-erij*, which forms nouns designating an (often negatively judged) activity (e.g. *filmerij* from *filmen*, ‘film<sub>v</sub>’) or a business (e.g. *drukkerij*, ‘printing office’, from *drukken*, ‘print<sub>v</sub>’).

This leaves us with three productive word formation processes producing process nouns in Dutch: nominalized infinitives as in (6b-c) and (7), suffixation with *-ing* as in (8) and suffixation with *-atie* as in (10).

### 2.3 Slovak and Dutch in comparison

It is interesting to compare the three competing processes in Dutch presented in section 2.2 with the two competing processes in Slovak presented in section 2.1. In Dutch, the nominalized infinitive only expresses the process reading. It is available for all verbs. Slovak *-nie* and its variant *-tie*, illustrated in (1), share with the Dutch nominalized infinitive their general availability for all verbs, but the resulting noun can also have a result reading, as illustrated in (3). In Dutch, the result reading can be expressed by *-ing* and *-atie*. Of these, *-ing* is widely available, but not for all verbs, as illustrated in (11c-e). The blocking of *\*protesting* by *protest* is directly comparable to Aronoff’s (1976: 43-45) example of *\*gloriosity* blocked by *glory*. This means that Slovak *-nie* does not quite correspond to either the nominalized infinitive or the suffixation with *-ing* in Dutch. Both Dutch *-atie* in (10) and Slovak *-ácia* and its variants in (2) are more restricted in their scope in the sense that fewer verbs can serve as their base. In Dutch, the regular use is restricted phonologically, in Slovak the central criterion is that we are dealing with internationalisms. The status of these generalizations is not the same. The Dutch rule that refers to verbs ending in *-eren* is a necessary, non-sufficient condition for the applicability of *-atie*, as illustrated in (11). The reference to internationalisms in Slovak represents rather a tendency than a condition for the application of *-ácia*, as illustrated by the use of *-ácia* with Slovak proper names. These considerations will play a role in the lexicographic representation, to which we will turn now.

## 3 The lexicographic representation of process nouns

As mentioned in section 1, KSSJ (2003) and van Dale (2015) do not use run-on entries as a lexicographic device. In the absence of run-on entries, the options for representing process nouns are reduced to including them as full entries with a separate headword or not treating them at all. Here, we will consider how the properties of each of the processes in Slovak and Dutch make these options more or less suitable.

The most regular process is the nominalized infinitive in Dutch. It is not only fully productive in applying to all verbs, but also fully predictable in its meaning, because only the process reading is available. In rare exceptions such as *eten<sub>N</sub>* (‘food, meal’), it is obvious that a full entry is needed. In all regular cases, not only a full entry, but also a run-on entry, would be redundant.

The Slovak formations with *-nie* are similar in the predictability of their existence, but cases where an additional reading exists are by no means exceptional. In (12), some examples of the dictionary definitions in KSSJ (2003) for such senses are given.

- (12) a. poistenie (‘insurance’)  
       zmluva poistenca s poisťovňou o úhrade škody (KSSJ)  
       (‘contract of [an] insured person with [an] insurance-company about [the] payment [in case] of damage’)

- b. vyšívanie ('embroidering<sub>N</sub>')  
rozpracovaná vyšívaná ručná práca (KSSJ)  
(‘unfinished embroidered hand work’)
- c. oznámenie ('announcement')  
písomná správa, tlačivo, na ktorom je uverejnená (KSSJ)  
(‘written message, document on which [it] is published’)

The senses recorded in (12) are no pure transpositions and are conventionalized. Therefore, they need to be recorded in a dictionary. In the case of (12a), the ‘contract’ reading is so dominant that it is hard to refer to the pure process reading. As no other form is available in Slovak, this will have to be done by means of a syntactic description. In (12b), the process reading of the noun refers to the ongoing action, whereas the result reading designates the piece of needlework being produced. As the definition in (12b) expresses, the needlework can only be referred to as *vyšívanie* as long as it is not finished, i.e. as long as the process is still ongoing. In (12c), the derived sense takes the form of the document in which the message being announced is expressed.

In the entries of (12), the pure process reading is not referred to, but at least for (12b-c) this reading is equally available. One may doubt whether this is the best solution. As mentioned, in (12a) the pure process reading is not readily available. The contrast between these two situations is lost when the regular meaning is not recorded for (12b-c).

In the case of Slovak *-nie* and *-tie*, run-on entries are in general not useful. The existence of the form is generally predictable. All verbs have one and the shape can be predicted on the basis of the inflection class, which is also visible in the form of the infinitive. Where special meanings exist, a separate headword is a better solution. When no special meanings exist, a run-on entry would be redundant. KSSJ (2003) and other Slovak dictionaries generally do not include them.

For Dutch *-ing*, we have seen in (11) that the existence of a nominalization cannot be predicted easily. Van Dale (1992) gives entries for *distantiëring* (‘distancing<sub>N</sub> oneself’) and *stranding* (‘running aground’) as in (13).

- (13) a. **distantiëring** (v.), het zich distantiëren.  
b. **stranding** (v.), het stranden: [...]; – keer dat een schip strandt.

The underline in the headwords indicates the stress. In (13a), the information in the entry is minimal. The gender information can be predicted from the suffix and the definition is no more than the nominalized infinitive of the corresponding verb as a synonym. In (13b), the additional information consists of an example (omitted here) and a second reading. *Keer* means ‘time’ in the countable sense. As we noted above, the nominalized infinitive does not have a plural. Here, attention is drawn to the fact that a plural can be used for multiple events of *stranding*. However, this is also a property of *-ing* in general. The only information added by the entries in (13) is that *distantiëring* and *stranding* exist. By contrast, no entries are given for *overheveling* (‘transfer’) and *verweving* (‘interweave’). This can hardly be due to the low frequency. CHN (2013) gives 990 occurrences of *overheveling* and 76 of *verweving* as against 13 for *distantiëring* and 65 for *stranding*. A systematic representation of the noun with *-ing* as a run-on entry if it only has a pure process reading would increase the consistency of the dictionary. Of course, for cases such as *vertaling* in (8) a separate headword would still be necessary.

In the case of *distantiëring*, the run-on entry would arguably have another advantage. As the verb stem ends in *-eer*, it is also in the scope of the application of the *-atie* suffixation rule. However, *\*distantiatie* does not exist. In this sense, the situation is similar to (11b). Giving only *distantiëring* as a run-on entry describes this situation concisely. Such information is not expressed very well in the case of *neutraliseren* (‘neutralize’), as seen in the entries in (14).



- (14) a. **neutralisatie** (v.).  
 b. **neutraliseren** [...]  
 c. **neutralisering** (v.), het neutraliseren of geneutraliseerd-worden in de bet. 1, 2, 3 en 4: [...]

The verb in (14b) is described grammatically, etymologically and semantically. It has five senses. The process noun in *-ing* in (14c) is described as applying to the first four senses only (“bet[ekenis]” is ‘sense’). Two examples at the end are omitted in (14c). The sense description gives the active and passive nominalized infinitive. The alternative process noun in (14a) is given without any further information. Although it can be argued that (14c) gives information that would be lost in a run-on entry, the close proximity to (14a) means that it does more to raise questions than give answers. Is (14a) only used in the active voice? Is (14a) the only nominalization for sense 5? Arguably, replacing (14a) and (14b) by run-on entries for the two nominalizations at the end of (14b) would give a clearer representation of the situation.

In the description of Slovak process nouns in *-ácia* and its variants in KSSJ (2003), similar problems can be found. In (15), we give the entries for *modernizácia* (‘modernization’) and *identifikácia* (‘identification’).

- (15) a. **modernizácia** modernizovanie  
 b. **identifikácia** identifikovanie

The definitions KSSJ gives consist exclusively of the alternative nominalization with the native suffix *-nie*. As for Dutch *-ing*, this strategy is not used consistently in van Dale (1992). Examples of words attested in SNC (2015) but not listed in KSSJ (2003) include *sebalikvidácia* (‘selfliquidating<sub>N</sub>’), *intimizácia* (‘intimizing<sub>N</sub>’), *efektivizácia* (‘effectivizing<sub>N</sub>’), *cyklizácia* (‘cycling<sub>N</sub>’), *prioritizácia* (‘prioritising<sub>N</sub>’). The nouns in (15) have both a process reading and a result reading. If the contrast between the two needs to be expressed, the noun in *-nie* denotes the process and the one in *-ácia* the result, but in general both nouns have both readings. This information should thus not so much be expressed in dictionary entries of the type illustrated in (15), as in grammars (as far as they cover word formation) or in the dictionary entries for *-nie* and *-ácia*.

For Slovak nouns in *-ácia*, there is an additional reason to combine their description with that of the corresponding verb. As argued by Panocová (2017), there are good reasons to assume that process nouns in *-ácia* are not derived from the corresponding verbs. Instead, the nouns were borrowed and the verbs should be seen as back formations based on the nouns. She gives evidence for this hypothesis based on frequency and meaning. Whereas in general, process nouns in *-nie* are (much) less frequent than their verbal counterparts, for nouns in *-ácia* the corresponding verb is generally (much) less frequent. Moreover, the meaning of the verb is often described on the basis of the noun. If this hypothesis is correct, a representation of the noun in *-ácia* in the same entry as the verb is also justified in terms of word formation. It might even be envisaged to make the noun in *-ácia* the headword and the verb in *-ovat’* a run-on entry.

## 4 Conclusion

The use of run-on entries is not common in dictionaries of Dutch and Slovak. Process nouns are a good example of the type of word for which run-on entries can be a suitable lexicographic device. The formation of process nouns is based on a number of different, language-dependent rules. Here we compared the rules in Slovak and Dutch. In Slovak, there are two main regular processes, whereas in Dutch there are three. This constellation gave us the opportunity to show how in a number of different

circumstances the use of run-on entries can be an improvement. The main recommendations of this work can be summarized as follows.

- Where the existence and shape of a process noun is entirely predictable, no run-on entries are used. If there is an additional sense, a separate entry is warranted. Otherwise, the process noun need not be mentioned.
- Where the existence of a process noun or the rule by which it is formed are not predictable, a run-on entry is used, unless the meaning is more complex than can be predicted.

For Slovak, this means that only nouns in *-ácia* are candidates for run-on entries. Combining the process noun and the corresponding verb in one entry has the added advantage of improving the representation of the relationship between the two, because in many cases it can be argued that the verb is a back formation based on the noun. For Dutch, the nominalized infinitive is generally not represented. Run-on entries are useful for process nouns in *-ing* and in *-atie*, in particular for verbs that could have process nouns with either suffix. Because there are two competing formation processes in Dutch, there is a stronger need to record which one(s) are in common use. In both languages, the practice of describing the meaning of a process noun by giving another process noun should be abolished.

## References

- Aronoff, Mark H. (1976), *Word Formation in Generative Grammar*, Cambridge (Mass.): MIT Press.
- Aronoff, Mark H. (1994), *Morphology by Itself: Stems and Inflectional Classes*, Cambridge (Mass.): MIT Press.
- Atkins, B.T. Sue & Rundell, Michael (2008), *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.
- Booij, Geert (2002), *The Morphology of Dutch*, Oxford: Oxford University Press.
- Buzássyová, Klára (2010), 'Vzťah internacionálnych a domácich slov v premenách času' [The relationship between international and native words in changing times], *Jazykovedný časopis* 61(2): 113-130.
- CHN (2013), *Corpus Hedendaags Nederlands*, Leiden: Instituut voor Nederlandse Lexicologie (since 2016 Instituut voor de Nederlandse taal), <https://portal.clarin.inl.nl/search/page/search>
- van Dale (1992), *Van Dale Groot Woordenboek der Nederlandse Taal*, 12th ed., Geerts, G. & Heestermans, H. (eds.), Utrecht/Antwerpen: Van Dale Lexicografie.
- van Dale (2015), *Van Dale Groot Woordenboek van de Nederlandse Taal*, 15th ed., den Boon, C.A. & Hendrickx, Ruud (eds.), Utrecht/Antwerpen: Van Dale Lexicografie.
- Dokulil, Miloš (1962), *Tvoření slov v češtině I: Teorie odvozování slov* [Word formation in Czech I: Theory of deriving words], Praha: Nakladatelství ČSAV.
- de Haas, Wim & Trommelen, Mieke (1993), *Morfologisch Handboek van het Nederlands: Een overzicht van de woordvorming*, 's-Gravenhage: SDU.
- ten Hacken, Pius (2015), 'Transposition and the Limits of Word Formation', in Bauer, Laurie; Körtvélyessy, Livia & Štekauer, Pavol (eds.), *Semantics of Complex Words*, Cham: Springer, pp. 187-216.
- Horecký, Ján, Buzássyová, Klára, Bosák, Ján. a kol. (1989), *Dynamika slovnej zásoby súčasnej slovenčiny*. [Dynamics of contemporary Slovak lexis.] Bratislava: Vydavateľstvo Slovenskej akadémie vied.
- KSSJ (2003), *Krátky slovník slovenského jazyka* [Concise dictionary of Slovak], 4th ed., Kačala, J., Pisárčiková, M. & Považaj, M. (eds.), Bratislava: Veda.
- Mistrík, Jozef (1976), *Retrográdny slovník slovenčiny*. [Retrograde dictionary of Slovak.] Bratislava: Univerzita Komenského.
- Panocová, Renáta (2017), 'Internationalisms with the Suffix *-ácia* and their Adaptation in Slovak', in Litta, Eleonora & Passarotti, Marco (eds.), *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, 5-6 October 2017, Milano, Italy, Milano: EDUCatt, pp. 61-72.
- SNC (2015), Slovenský národný korpus – prim-7.0-public-all. [Slovak national corpus – prim-7.0-public-all.] Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2015. Available at WWW: <http://korpus.juls.savba.sk>.

- Svensén, Bo (2009), *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*, Cambridge: Cambridge University Press.
- van der Wouden, Ton (2018), '-ing', Taalportaal, <http://taalportaal.org/taalportaal/topic/pid/topic-14125866172746534> [downloaded 7 May 2018].

## **Acknowledgements**

This work was supported by the Slovak Research and Development Agency under the Contract No APVV-16-0035.

# Definitions of Words in Everyday Communication: Associative Meaning from the Pragmatic Point of View

**Svitlana Pereplotchykova**

*Taras Shevchenko National University of Kyiv*

*E-mail: s.pereplotchykova@knu.ua*

## Abstract

In their interactions with others, people try to be cooperative if they want to be understood by their interlocutor(s). Spontaneous speech presupposes a good command of the language in use, but the overall meaning of certain words, their recognizability (visibility), is closely connected with the associations that the real-world objects they denote or the words themselves have acquired and are true for particular social groups. In the word-based games of the original TV show *Hollywood Game Night* and its Greek version *Celebrity Game Night*, the players provide their associations with the objects of reality in question or the words denoting them in order to convey the information effectively. The results of this research comprise an overview of how speakers of different languages process and understand words, and how different associations and verbalizations can be not only among languages but also among those who speak the same language but belong to different social groups. The analysis of these associations allows us to explain which elements of the information they provide can help us to understand each other correctly, which are the factors that make interlocutors choose this or that association, and why sometimes they fail to elicit the information they ask for, to explain what they mean.

**Keywords:** salience, association, associative meaning, definition of words, communication, English, Modern Greek, semantic

## 1 Introduction

Since 2013 the television show *Hollywood Game Night* has run on NBC (USA) so successfully that some channels from other countries decided to buy the rights to broadcast it. The Greek Channel Mega in 2014-2018 has produced 57 episodes (as of May 19, 2018) of the localized Greek version of the US show, called *Celebrity Game Night*.

In each episode two teams consisting of one player and three celebrities compete for a money prize by playing various games with words, music, pantomime and so on. In some of the games the host or participants are to elicit a pre-determined target word/expression from others by giving them clues. The target words/expressions in the TV show are embedded in some semantic or formal frame. Specifically, they either belong to a thematic topic, which the host gives, or they are words beginning with a particular letter, also given by the host.

Within the present research the following word-based games are analyzed. *Clue-boom* is a game with a specified topic in which clue-givers choose a card with a word/expression from a bowl and provide the players on their team with clues which they think can help elicit the word. In the game *Take the hint* (called in the Greek version *Πες μου μία λέξη*, 'Tell me a word') the captain of the team has to guess the word/expression presented on the screen behind him/her based on clues the team members provide. One player each time can give a clue consisting of one word only. In the game with a specified topic. *Off the top of my head* (called in the Greek version *Εδώ τόχω*, 'Here I have it') the players wear special glasses where a card with a target word/expression written on it is placed, and

then provide each other with clues to elicit the target words/expressions. In the game *Where are you going?* (called in the Greek version *Πού πας*, meaning the same as the English title) one player of the team becomes the driver of an imaginary car. The host shows team members the card with the target word on it. He/she pretends to get in the car and tries to explain to the driver where they are going.

Sometimes a player-guesser names the predetermined word/expression at once, sometimes two, three and more clues are needed, and there are also times when the target word/expression remains unrecognized, i.e. not elicited. This clearly means that whether or not participants name the target word/expression successfully depends crucially on particular verbalizations. The explanations have to be short, and present the most recognizable features of the object in question so that the clues the clue-givers provide, hypothetically, are the most salient features of the real-world object in question. Therefore they are supposed to be those most closely associated with them, logically or, more often, subconsciously.

Thus the concern of the present research is to trace the reasons for both positive and negative outcomes of these word-based games by means of semantic analysis of the clues provided and the words/expressions elicited, and through analysis of participants' mini-dialogues, taking into consideration the communicative and social characteristics of the interlocutors. The paper is therefore intended as a specific contribution to the development of social and cognitive semiotics and pragmatics. The material analyzed can also be considered fieldwork, as those words-clues are spontaneous but stimulus-based elicitations.

## 2 Associations in the focus of scholarly research

The term “association” was coined in 1690 by John Locke, who offered the concept of the association of ideas, namely that ideas are interconnected, sequential and descriptive of experience due to “the psychological tendency to associate ideas through experience” (“John Locke” 2017). And in the 18<sup>th</sup> century an English doctor, David Hartley, in his book *Observations on Man*, represented the associations among ideas as a universal mechanism of mental life, thus marking the start of association psychology. Hartley sought to explain how the most complex mental processes – imagining, remembering, reasoning – might be analyzed into clusters or sequences of elementary sense impressions, and that ultimately all psychological acts might be explained by a single law of association. (“David Hartley” 2014). Since then psychology has dealt with associations, which are considered as links among objects based on our personal subjective experience. Associations have been a matter of concern for many well-known scholars, including A. Potebnia, Baudouin de Courtenay, M. Kruszewski, R. G. Kent, G. Miller, and Ch. Cofer, among others, who dealt with associations from both a psychological and linguistic perspective.

Neuroscientists have proved that “object categorization is an adaptive function of the brain, allowing organisms to sort information from the external world into behaviorally relevant classes” (Bao, Raguet, Cole, Howard, & Gottfried 2016). Sensory systems “generalize across different objects sharing similar features, but at the same time maintain the specificity of individual objects and categories” (Bao, Raguet, Cole, Howard, & Gottfried 2016). Sensory inputs are associated with “internal templates that are established through a lifetime of experience and encoded into memory” (Bao, Raguet, Cole, Howard, & Gottfried 2016). This experience can coincide with the experience of the culture we belong to (i.e. the experience is shared by all the people in touch with a particular culture), though the experience always remains personal, lived by the person himself/herself.

So, associations are a way to keep information in the memory of its speakers. There is a belief that the quality of memory increases as associative thinking develops, as verbal memory plays an important



role in remembering words and images whose analogues, namely abstract notions, do not have access to. When a person wants to say something, he/she searches for connections between objective reality and its perception. Verbal memory is to a certain extent opposed to image memory. Image memory is a depository of unprocessed raw images which keep correlations of their spatial characteristics, while verbal memory is based on the encoding of notions. This is why words and expressions are stored in the memory not as separate elements, but in groups or networks. Phenomena of reality are thus defined by their relationships to other concepts rather than by some internal essence. And the strengths of the connections correspond to a stable pattern of activity distributed over the whole network.

### 3 Salience as a reliable criterion for choosing appropriate associations

When the clue-givers in this show are faced with having to choose a word or expression to provide as a clue, they have to activate the available choices, clues which they believe will best help the guesser to find the word in question, and if there are many choices, their ranking. The word or phrase is then selected for utterance on the basis of maximum salience, and thus the most recognizable characteristics of the real-world object in question.

The clue-giver has to take into consideration and evaluate to what extent the person he/she addresses is familiar with the association field of the target word, and with the ranking within the set of associates. The clue-giver thus has to decide whether this person knows the inherently salient feature of the object in question, and then evaluate the distance between meanings and decide which of the salient features are more relevant for a particular guesser, what is his/her probable background knowledge, and the extent to which the guesser is aware of a collective perception and representation of the object in question. In other words, s/he must assess the accessibility of the target word. Words with stronger connections between a clue and the target or stronger shared associate links are more likely to be recovered:

Similarly, words that appear infrequently in the language, that have more connections among their associations, and that have relatively smaller sets of associations are more likely to be recovered regardless of the retrieval intention of the speaker (Nelson & Goodmon 2002: 393).

The analysis of the clues provided and words elicited in the show reveals that the successful eliciting of a pre-determined word/expression mainly depends on successful assumptions of likely mutual knowledge and reference to the most salient feature of the object of the reality in question. According to I. Kecskes:

As a semiotic notion salience refers to the relative importance or prominence of signs, in pragmatics when we speak about salient information we mean given information that the speaker assumes to be in central place in the hearer's consciousness when the speaker produces the utterance. It is the most probable out of all possible. (2014: 1)

The most salient feature of the object of reality is the one we personally encounter most often, while the social group we belong to has agreed on a particular conceptualization of this object. The salience of different aspects of a word's meaning is determined by its frequency in participants' mental lexicon, and by the strength of the shared associations engendered by its real-world familiarity, conventionality and prototypicality (Giora 2003: 10).

When choosing a clue, the clue-giver has to take into consideration the current "reality" of the interlocutor. Both clue-giver and guesser have to share the same cultural space, to be closely connected with the players' current "reality", or their cultural space. The more frequent, familiar, conventional, and prototypical the meaning, the quicker it is to retrieve (Giora 2003: 16-17). As such, "salient

information includes several factors that might make it ‘feel right’: it springs to mind first, it is familiar, it is likeable and it resists change (it is hard to eradicate or attenuate)” (Giora 2003: 199).

## 4 Classification of associations

Grouping of the clues found in the show turned out to be a difficult task. There are various classifications of verbal associations based on different criteria. Aristotle distinguished three types of associations, namely associations of contiguity, similarity and contrast. The Russian scholar G. Martynovich, based on the formal, functional and semantic characteristics of responses retrieved in an association experiment, distinguishes only two types of associations, namely those of contiguity (in time or space) and those of similarity, claiming that they embrace all the other classifications of associations. For him the associations of contiguity are associations of metonymic character, pairs of words which are connected thematically. Such pairs of words comprise semantic groups of hyponyms-hyperonyms, meronyms and holonyms, and the types of figurative speech – metonymy and synecdoche. Associations of similarity are associations of metaphoric character, as they appear between objects which have one or several common essential features. Similarity of content of verbal associations, i.e. similarity of lexical meanings, implies that in the meaning of associations there are common semes (semata). They can be synonyms or antonyms, but most often they are pairs of words, where the semantic content of one comprises a part of the content of another: usually here belong words which act as modifiers often used with the other word in a pair. Alternatively, these may be pairs of words, which have in their content at least one common essential feature. But some pairs of words can be simultaneously associations of contiguity and similarity (Martynovich 1997: 5-6). Just like in our case, as most of our clues are associations of contiguity. We thus have to choose some other criteria.

In this respect the classification by C. Jung and F. Riklin, who carried out the series of experiments on subjects, turns out to be more relevant for our material. They suggested distinguishing the following types of associations: internal and external associations, sound reactions, miscellaneous, ego-centric reactions, perseveration, repetition of the reaction and linguistic connection (Jung & Riklin 2014).

## 5 The material of the present research

Our material of 300 target words/expressions in English and 300 in Greek with clues consisting of one to 25 words each, is received on the basis of transcriptions of videos of verbal interactions between players of the show in the games described earlier.

According to the rules of the show there are some restrictions: neither translation equivalents, nor cognate words can be provided as clues. There is also a time limit of 90 seconds (for each game for each team). This factor by no means diminishes the value of our research, as the time is not so crucial for perception and production of information as it may seem. According to the latest research carried out by neuroscientists from the University of Berkeley (Haller et al. 2018), the human brain needs only several seconds to respond to visual or aural stimuli. It has even been proved experimentally that the brain already begins to prepare for a response at the stage of perceiving stimuli, that is it is ready to answer before even knowing what to respond. But acting under time pressure the clue-givers more often provide those clues which appear in their minds first, and only behave otherwise if they take more seconds or in the course of the interaction choose more suitable clues. Besides, in such a stressful situation some people turn out not to be ready psychologically to quickly react in general, due to their personal characteristics.

Unlike association experiments, when participants are asked to give word responses to stimulus words/expressions for the sake of an experiment, in the TV show the clues-responses are provided with the aim of successful communication leading to earning points and money. Furthermore, the elicitations in the show are the result of the guesser's analysis of the clues given, and this allows us to examine whether associations are common among different people/groups of people/languages and to characterize associations judging by the success of communicative interaction. The clues comprise semantic constituents of the target words/expressions, either single words or utterances describing the object the target word/expression denotes, and chosen among other possible variants due to pragmatic scope.

## 6 Types of association used for defining target words from a pragmatic point of view

Taking this into consideration, Giora's classification of types of salience allows us to group all the clues into inherent associations (features comprising an unchanging characteristic of the object in question, as denoted by the target word/expression), acquired/external associations (based on information stored in the brain of the members of a particular social group), and emergent situational associations (unstored information) (Giora 2003: 7). It should be mentioned, however, that those three types can coexist, as clue-givers can provide all those associations simultaneously in the games, with communicative interaction in the form of a dialogue between participants.

### 6.1 Inherent associations

#### 6.1.1 Constant characteristics or attributes

The analysis of the clues provided reveals that the most recognizable features, those comprising inherent associations, are constant characteristics of the objects in question, indivisible, integral, and inseparable due to their origin, whether natural or as a result of the activity of people. This can be a part of the body of a living being or of an inanimate object, accompanied by some more specific characteristics (e.g. *peacock – tail, colorful*; *Pinocchio – nose, long*; *beaver – teeth, dam*; *ghost – scary, sheet*; *zebra – stripes, looks like horse*; *message – bird, twitter*; *δέντρα – φύλλα* “trees – leaves”; *ποδήλατο – πετάλια* “bicycle – pedals”; *ελέφαντας – προβοσκίδα* “elephant – trunk”; *φίδι – δαγκώνει, μία γλώσσα μεγάλη* “snake – bites, a long tongue”; *θρόνος – βασιλιάς* “throne – king”, etc.) or it can be just a recognizable attribute of the object in question, as it is its symbol, or an integral, inherent part of it, like *pumpkin* for *Halloween*, *kangaroo* for *Australia*, *the maple leaf* for *Canada*, or *the pyramids* for *Egypt*, *turkey* and *stuffing* for *Thanksgiving Day*, *το ηφαίστειο – η Σαντορίνη* “volcano – Santorini”, *το Κολοσσαίο στην Ιταλία* “the Colosseum in Italy”, *η Βηθλεέμ – ο Χριστός* “Bethlehem – Christ”, *τράπεζα – λεφτά* “bank – money”, *δωμάτιο – έχει μέσα το σπίτι* “room – a house has inside”, and so on. In this respect one interesting association should be mentioned. In the Greek show for the target word *μανιτάρι* “mushroom” the successful clue *Στρουμφάκια* “the Smurfs” was given, probably because those fictional creatures known from comics live in houses in the form of mushrooms, and the particular clue-giver made an assumption that the guesser was familiar with these characters. To be sure the target word would be named successfully, the clue-giver also specified what exactly was meant by providing the hyperonym *φαγητό* “food”. The inherent characteristic of the object the target word denotes can be a color, a shape, an origin, and so on, expressed with adjectives, or via other descriptions. Like *Kool-aid*, which is a brand of flavored drink mix and can be easily distinguished among other drinks by its colors *Kool-aid blue/red*, or a taxi, which is yellow in the USA. In the American show the hypo-hyperonymic relations are more often presented as clues, which are rare in the Greek version, like for example, *fox – animal in woods, usually scares people, red, you make*

*fur*. Some objects which do not necessarily comprise an integral part of American or Greek culture but are widespread among the citizens of these countries are even presented in the same way. Compare *donut – is a round thing you eat in pastry, very delicious round sprinkles, they have a hole in the middle* and *κρουασάν – έχει το σχήμα του φεγγαριού, είναι γλυκό, το τρώμε με τον καφέ* “croissant has the shape of the moon, it is sweet, we eat it with coffee”.

### 6.1.2 A function as an integral characteristics of the object

In both American and Greek versions clue-givers present the target object by means of verbs, describing how the object functions/works, what it is used for, or which action makes it possible for the object in question to exist: *egg – a chicken makes*; *zoo – all the available animals live there*; *napkin – you use it to wipe yourself when you eat*; *rice – Chinese people grow it*; *ο θερμοσίφωνας – το ανάβουμε για να έχουμε ζεστό νερό* “boiler – we switch it on to have warm water”; *οδοντίατρος – εκεί πας να κάνεις σφράγισμα* “dentist – you go there to fill a tooth”; *τρακτέρ – το οδηγούν οι αγρότες* “tractor – farmers drive this”; *όσπρια – τα τρώμε, φακές* “legumes – we eat them, lentils”; *εκκλησία – πάμε εκεί να προσευχηθούμε* “church – we go there to pray”; *δράκος – βγάζει φωτιές, μυθικό πλάσμα* “dragon – belches flames, mythical being”; *ρολόι – στο χέρι μας φοράμε και βλέπουμε την ώρα* “watch – we wear it on our wrist and see the time”; *ψηφοδέλτιο – το πετάω μέσα στην κάλπη* “ballot paper – I put it in a ballot box”, and so on. In this respect one experiment is worth mentioning, carried out on Russian language speakers, which showed that such syntagmatic associations are mostly provided by children (“chair – I sit”), while elderly people use paradigmatic associations (“chair – table”) (Nikolayeva 2008).

The fact of providing these sort of clues, which are easier for production and comprehension, may be explained on the one hand by the findings of recent research in psycholinguistics, while on the other hand may actually turn out to support such findings. According to the data collected by means of on-site fieldwork on English, Dutch and other languages, speakers tend to slow down before saying nouns compared with verbs. The researchers “attribute this slowdown effect to the increased amount of planning that nouns require compared with verbs”, and suggest that “there are robust universals of language processing that are intimately tied to how speakers manage referential information when they communicate with one another” (Seifart et al. 2018).

### 6.1.3 Association by contrast

The most salient feature of the object in question may also be identified by means of the name of the object (or its inherent feature), usually opposed to the object the target word/expression denotes: for example, the compass point of “North Pole” is represented as “not South” in both American and Greek versions of the show. Other examples: *nurse – not a doctor*; *dogs – not cats*; *CNN – not Fox News, but...*; *sorority – not a fraternity, but...*; *αριθμοί – όχι τα γράμματα, τα άλλα που μαθαίνουν* “numbers – not letters, the other things they (children) learn”; *μαμάδες – όχι μπαμπάδες* “mothers – not fathers”; *υψος – όχι βάθος* “height – not depth”; *ελικόπτερο – όχι αεροπλάνο* “helicopter – not airplane”, etc.

Sometimes providing the most salient feature, which seems to be true for all people acquainted with a particular object in their reality, does not secure a successful outcome, as languages name these objects in a different way. So from the conceptual, notional point of view, such clues correspond to reality, but from a linguistic point of view, taking into consideration the rules of the game, they cannot be provided. For example, in the Greek version of the show the clues given for the word *πεπόνι* “melon” were *κίτρινο, καρπούζι, αντίθετο* “yellow”, “watermelon” (a non-cognate word in Greek) and “opposite”. But in the American show the attempt to provide a clue “melon” for the target word



“watermelon” was not successful, as in English it is a cognate word, which is not allowed according to the rules.

#### 6.1.4 *Multiplicity of associations*

In most cases inherent associations provided as clues work well, and players name the target words/expressions correctly. But there are also cases when those associations are inherent for several objects, and a clue-giver fails to make it clear for a guesser which exact word/expression has to be named. For example, for the target word *Mozart*, the following clues have been provided: *Αυστρία* “Austria”, *μουσική* “music”, *Μπαχ* “Bach” and the guesser infers *Βιέννη* “Vienna”. Or the target word *Βρυξέλλες* “Brussels”, the city being an embodiment of the Greeks’ sufferings, as all the demands of the EU come from its headquarters situated here. So the clues provided for this target word are *Βέλγιο* “Belgium”, *ευρωβουλευτές* “Euro-MP”, *συνθηκη* “treaty”, *κονέ* “connection, nepotism”. The response of the guesser based on these clues was *κοινοβούλιο* “Parliament”. The clues provided for the target word *cactus* in the American version of the game were *plant* and *desert*, while in the Greek version *τεκίλα* “tequila”, *αγκάθια* “thorns”, *έρημος* “desert”. In the American version the word was successfully elicited, while in the Greek version the player gave the answer *Mexico* due to the abundant and inappropriate clues provided – a parallel situation to that with the opposites mentioned above. Some objects can be associated simultaneously with more than one object, thus there can be more than one pair of opposites accentuating various aspects. In this case, clues have to be more specific to single out the target object, and features inherent to it, thus distinguishing it from other possible responses. For example, the reaction of the guesser to the clue *όχι γάτες* “not cats” was *σκόλοι* “dogs”, while the guesser wanted to elicit *ποντίκια* “mice”.

## 6.2 *Acquired associations*

Acquired or external associations represent the particular picture of the “world” of a particular people. To this group belong either target words naming the objects of American or Greek reality, and so clues are also connected with it, or clue-givers may choose clues specific to their particular reality to successfully elicit the target word/expression. As such, we next distinguish extralinguistic associations and linguistic associations.

### 6.2.1 *Extralinguistic associations*

When words denote some objects or actions closely connected with the traditions, (social) customs and manners, history and culture of the particular ethnos (including those which became known for other ethne), they acquire some additional meanings and consequently other associations. For example, the target word *flowers* being an obligatory attribute of a marriage proposal can be associated with the particular social situation, and so we find it presented as *I propose to you with*. For the target word *δαμάσκηνο* “plum” the clue-giver provided the clue connected with the belief that this fruit helps with digestion, *το τρώμε όταν είμαστε δυσκοίλιοι* “we eat it when we have constipation”. One of the destinations in the game “Where are you going?” was *Συρία* “Syria”, and the clue-giver presented it as *εκτός Ελλάδος, κάπου μακριά, από κει έρχονται οι άνθρωποι με τα σωσίβια* “outside Greece, somewhere far, people in life-jackets are coming from there”. In the game with the given topic “Economy” the target word *Τρόϊκα* “European troika”, denoting the commission inspecting Greece, with the word having negative connotations for the Greek speakers, is presented as *δε μας αρέσει όταν έρχεται αυτή και μας κάνει ελέγχους* “we don’t like it coming and checking”. The guesser thought of *εφορία* “tax office”, so specification was needed, *απ’έξω* “from abroad”, and the word was successfully elicited. The target word *μέτρο* “underground” in the Greek show was presented as *μέσο* “means”



and Συγγρού-Φιξ “Sygrou-Fix” which is a metro station in Athens. The target word *ρίγανη* “oregano”, denoting a herb widely used in Greece, was presented through its connection with the most Greek salad, which is known as village salad, *στη χωριάτικη βάζουμε ντομάτα, πράσινο από πάνω, ψιλό* “we add to the village (Greek) salad tomatoes, the green above, thin”. In the American show for the target word *beaver*, the clue-giver provided words which describe inherent characteristics of this animal, namely *teeth*, *wood*, and *dam*, which beavers build to create ponds, but for some reason the guesser named *Abraham Lincoln*. The thing is that the clues “teeth” and “wood” in fact present the animal beaver. But the word “beaver” used with the word “dam” comprises a part of the name of the Battle of Beaver Dam Creek, which took place in 1862, shortly after Abraham Lincoln became US President.

### 6.2.2 Linguistic associations

The clue-givers provide not only clues presenting the most recognizable feature of the target object in reality, as some are connected with specific verbalizations characteristic for the particular language. It should be mentioned here that salience as a phenomenon, according to our analysis, is connected not only with the object itself – and the characteristics provided by the players are not always characteristics of the object as of physical substance, which have tangible presence like color or shape. Here we probably have to talk about linguistic salience, when the object of reality is recognized through the words which are usually used with the word denoting it, because of its lexical collocations, its compatibility. The Russian scientist Y. Karaulov referred to this “associative grammar”, stating that the human mind keeps words and grammar stored together, in the form of lexical and grammatical compatibility (1999). We memorize not simply isolated, single words, but their relations to other words through grammatical coordination.

These linguistically inherent and simultaneously acquired associations are found in both versions, but are used more often in Greek, which can also be explained by the fact that the Greek language is more spoken than written due to its history, and the fact that Greeks are more expressive in narration. For example, the Greek word *δίσκος* “disk” comprises a part of a set expression *ο ιπτάμενος δίσκος* “flying saucer”, so by providing the clue “flying” the clue-giver successfully elicits the target word “disk”, the word *γλέντι* “revelry” in Greek has a constant modifier *τρικούβερτο* “high jinks”, the clue for the word *παραμύθια* happens to be *του Αισώπου* which is a set expression from *Aesop's Fables*..

In the following dialogue, in order to elicit the word *μάρτυρες* “witnesses” within the topic “Court” the clue-giver decided to recall the image of court hearings and presents the address in the Greek court when a judge gives the floor to the participants in a hearing:

Clue-giver: *Πείτε μας κύριε...*

Guesser: *Συνήγορε.*

Clue-giver: *Όχι.*

Guesser: *Κύριε δικαστά...*

Clue-giver: *Όχι*

Guesser: *Κύριε εισαγγελέα...*

Clue-giver: *Αυτός που έχει το τυχαίο σου...*

Clue-giver: Tell us Mr....

Guesser: The council for the defense.

Clue-giver: No

Guesser: Your Honor Judge...

Clue-giver: No.

Guesser: Mr. District Attorney...

Clue-giver: The one thereon hangs your fate.

Proverbs can also be provided with the target words they include omitted. For example, we find the target word *πίτα* “pie” in the proverb *και αυτή ολόκληρη και ο σκύλος χορτάτος*, literally “it (a pie) remains whole, and a dog well-fed”, or “eat one’s cake and have it”. So here the Greek pie is elicited by means of the proverb-clue “the wolves are full and the sheep are whole”. The word *χρυσός* “gold” is elicited by means of the saying *ό,τι γυαλίζει, δεν είναι ο χρυσός* “all that glitters is not gold”.

It is worth mentioning here that one word can be used in various contexts, and it depends on the clue-giver which one to choose. In the following example the target word *παραβάν* “polling booth” can be used with the verb *κλείνω* “close”, as we have to do this when entering inside. But the overall context allows other interpretations, and the guesser thinks of a mobile phone which should be switched off, as in Greek the verb “close” is used to denote this action. This is why the clue-giver is forced to provide more specific details, and says that the object in question has to be closed for no one to see us inside.

So the acquired association can be interpreted correctly if the clue-giver and guesser share the same knowledge from the history and culture of their country. The target expression *Mary Poppins* from the American show is more than efficiently represented with the clue *supercalifragilisticexpialidocious*, but chiefly when addressing English-speaking people of a certain age. This nonsense word was coined by the writers of one of the best known songs in cinema, and was for a long time one of the most famous long words in English. For younger speakers, and in languages other than English, however, the value of the clue will vary depending on whether the story of *Mary Poppins* is still popular and on whether the film is still shown, dubbed or subtitled. We may compare this case with a very famous Soviet film made in 1983, where the heroine is surely described, and instantly recognized, by Russian-speakers born in the 1970s, as “Samo sovershenstvo” (perfection itself, pluperfection).

We also have to mention one more grammatical means widely exploited by Greek clue-givers, which does not exist in English or Ukrainian. In Greek there is a definite/indefinite article which has different presentations for each gender: “ο” for masculine, “η” [i] for female and “το” [to] for neuter gender. So when the player gives a clue, the word denoting the object of reality is often accompanied by its article, showing not biological but grammatical gender (as in Greek, for example, “a small girl” is of neuter gender), thus suggesting the target word. The assignment of gender in all the languages having it is voluntary, and not necessarily coincides with biological gender. The following interaction between the clue-giver and guesser did not end up with the successful eliciting of the target word *η μπάλα* “a ball” (female in Greek), probably because the reference to the article distracted the guesser. The first clue provided was a definite article for female gender, then the clue-giver added the words *βλέπω την* “I see [it] [definite female article in accusative case]”. This made the guesser think of something connected with “football” (the topic of the game) being female in gender, and the response given was thus *την καρτέλα* “the card” (of female gender in Greek). The following specification was *τη στρογγυλή* “the round one”, which elicited *κάμερα* “camera”. The clue-giver then decided to change tactics, and presented different kinds of the target object in reality *του μπάσκετ, του ποδοσφαίρου, βλέπω ...* “foot-ball, basket-ball, I see [it]”.

### 6.2.3 Emergent situational associations

Experiencing a word can prime its accessibility and associative connections to related words (Nelson & Goodmon 2002: 380). That is why in this show, which is a team game providing emergent situational associations, the participants typically try to offer clues connected with the appearance or occupation of their co-players, or those activating their experience of the contact with the object of reality in question. “Prior and actual situational experience is privatized / subjectivized and prioritized in the mind of interlocutors” (Kecskes 2017). As a result, there may be no single point in the recovery process at

which a clue-giver's utterance fully matches a guesser's interpretation. This is because in both clues and the guesser's interpretation the "analysis of clues is "contaminated" by individualized pragmatic elements" (Kecskes 2014: 192). This is what C. Jung calls "egocentric reactions". Such clues are not always enough, and other salient features are referred to verbally and non-verbally. For example, when the word *glasses* had to be elicited in the show from a person wearing glasses, the most secure/salient feature in that situation was *you are wearing them right now*. Or the target word *drought* within the topic "Los Angeles" in the American show was presented as *right now we have not much water* with reference to the actual weather conditions. Similarly, the color in the target expression *Red Square* was presented as *it's not black, ... I'm wearing it now* (the player was in a red shirt). The following text presents the dialogue to elicit the target word *song* within the topic "Nursery school":

Clue-giver: *Εγώ γράφω...*

Guesser: *Στο θρανίο;*

Clue-giver: *Η δουλειά που κάνω τι είναι;*

Guesser: *Στιχογράφος.*

Clue-giver: *I'm writing...*

Guesser: *On the school-desk?*

Clue-giver: *My job... what I am doing*

Guesser: *Lyrics writer.*

Even if clue-givers provide surely recognizable features, characteristics that are absolutely salient, some objects and the words denoting them remain unrecognized. The most probable explanation why this happens is that the clue-givers are providing the guessers with features which are salient for them personally, because of their education, experience, interests, social environment, and so on, but the guessers actually have "different prior experiences, varying evaluations of the actual situational context, individual degrees of salience which result in a subjectivized process of production and comprehension" (Kecskes 2014: 192). In the following dialogue from the Greek show the target destination in the game "Where are you going?" is Myrto Beach on the island Kefalonia, known for its huge waves.

Guesser: *Πού πας;*

Clue-giver: *Κεφαλονιά, θάλασσα, πολύ μεγάλα κύματα...*

Guesser: *Ωραία...*

Clue-giver: *Κεφαλονιά. Έχεις πάει Κεφαλονιά;*

Guesser: *Όχι.*

Clue-giver: *Τότε κατέβασέ με να φύγω.*

Guesser: *Where are you going?*

Clue-giver: *Kefalonia, sea, huge waves...*

Guesser: *Good...*

Clue-giver: *Kefalonia. Have you been there?*

Guesser: *No.*

Clue-giver: *Then stop and let me get out of the car.*

To prevent such failures and secure successful communication, some players provided clues they considered more relevant for a particular guesser. For example, when eliciting the word *ψαλίδι* "scissors" within the topic "Nursery school", one participant referred to the well-known children's game *πέτρα, ψαλίδι, χαρτί* "rock, scissors, paper". But within the topic "Football", where *ψαλίδι* "scissors kick" is a term, one male clue-giver presented it to the female guesser as *...κόβουμε τη κοτσίδα με αυτό* "we cut a plait with this".

## 7 Conclusion

TV shows like *Hollywood Game Night* are created to entertain viewers and to avoid insulting them. As it is not an intellectual game, the makers of the show choose as the target words to be elicited in well-known terms denoting objects used in everyday life, or which are otherwise widely-known. But the dialogues we find in the show also happen in everyday life rather often when we sometimes forget the name of an object we want to refer to. And just like the participants in the show in such cases we call not on dictionary definitions to get the word we want, but instead address the background knowledge we assume is common for our interlocutors.

The participants of the show thus search for the most salient characteristics of the objects in question, which are their inherent characteristics or attributes, and their functions. They may well fall back on linguistic associations which the names of the objects acquire due to compatibility with the target words. Moreover, all this is due to the fact that they rely on the commonality of the communicative situation, and evaluate the salience of this or that association according to Grice's maxims of quantity, quality, relation and manner. The clue-givers provide as much information as they judge to be enough for guessers to understand them. They provide their interlocutors with the associations that they believe to be true and familiar to the guessers, presumably because they have experienced the situations involving the objects in question personally. The clue-givers try to be clear and avoid obscurity and ambiguity.

We believe the further study of associations could provide lexicographers with a great deal of useful data. The criteria of salience that ordinary people, native speakers of a language, apply for choosing associations may help in writing definitions which will be more precise and comprehensible. In this respect, associative experiments should be held with subjects (obtained via crowdsourcing or otherwise encouraging different types of speakers to be involved) being asked to elicit definitions of target words, and not associations with real-world objects, as is usually done.

## References

- Bao, X., Raguette, L. L., Cole, S. M., Howard, J. D. & Gottfried, J. A. (2016). The role of piriform associative connections in odor categorization. In *eLife*, 5, e13732. Accessed at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4884078/#bib60> [01/04/2018].
- David Hartley. (2014). In *Encyclopædia Britannica*. Accessed at: <https://www.britannica.com/biography/David-Hartley> [01/04/2018].
- Giora, R. (2003). *On Our Mind: Salience, Context, and Figurative Language*. Oxford: Oxford University Press.
- Haller, M., Case, J., Crone, N.E., Chang, E.F., King-Stephens, D., Laxer, K.D., Weber, P., Parvizi, J., Knight, R.T. & Shestyuk, A. (2018). Persistent neuronal activity in human prefrontal cortex links perception and action. In *Nature Human Behaviour*, 2, pp. 80-91.
- John Locke. (2017). In *Encyclopædia Britannica*. Accessed at: <https://www.britannica.com/biography/John-Locke> [01/04/2018].
- Jung, C., Riklin, F. (2014). The associations of normal subjects. In Sir H. Read, M. Fordham & G. Adler (eds.) *C. Jung The collected works*. Vol. 2. Experimental Researches. New York and London: Routledge. Accessed at: <https://books.google.com.ua> [01/04/2018].
- Karaulov, Y. (1999). *Aktivnaya grammatika i asociativno-verbalnaya set*. Moscow: Institut russkogo yazyka im. V.V. Vinogradova RAN.
- Kecskes, I. (2014). *Intercultural pragmatics*. Oxford: Oxford University Press.
- Kecskes, I. (2017). From pragmatics to dialogue. In E. Weigand (ed.) *The Routledge Handbook of Language and Dialogue*. New York and London: Routledge, pp.78-92.
- Martynovich, G. (1997). *Verbalnye associacii v asociativnom eksperimente*. Saint-Petersburg. Accessed at: [http://lit.lib.ru/img/m/martinovich\\_g\\_a/01verbass/werb\\_associacii.pdf](http://lit.lib.ru/img/m/martinovich_g_a/01verbass/werb_associacii.pdf) [01/04/2018].

- Nelson, D., Goodman, L. (2002). Experiencing a word can prime its accessibility and its associative connections to related words. In *Memory & Cognition*, 30 (3), 380-398.
- Nikolayeva, E.I. (2008). *Psihofiziologiya. Psihologicheskaya fiziologiya s osnovami fiziologicheskoy psihologii*. Moscow: PER SE.
- Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., Witzlack-Makarevich, A., de Jong, N., & Bickel, B. (2018). *Nouns slow down speech across structurally and culturally diverse languages*. In *Proceedings of the National Academy of Sciences of the United States of America*. Accessed at: <http://www.pnas.org/content/early/2018/05/09/1800708115> [20/05/2018]



# Verifying the General Academic Status of Academic Verbs: An Analysis of co-occurrence and Recurrence in Business, Linguistics and Medical Research Articles

**Natassia Schutz**

*Centre for English Corpus Linguistics, Université catholique de Louvain*

*E-mail: natassia.schutz@uclouvain.be*

## Abstract

General academic vocabulary lists have been the subject of much debate. Because they focus on single words, they have been criticized for not considering “the importance of contextual environments which reflect different disciplinary practices” (Hyland & Tse 2007: 251). This study aims to provide insight into the reliability of such vocabulary lists by analyzing cross-disciplinary phraseological variation. To do so, I analyze the collocations and lexical bundles used with c. 30 academic verbs found in a 3-million-word corpus containing research articles in business, linguistics and medicine. The results seem to suggest that there are sufficient commonalities, both in terms of use and meaning, to justify the creation and use of general academic vocabulary lists. In addition to their discipline-specific uses, many of the verbs under focus also have general academic uses that relate to the core business of research, irrespective of the academic discipline (e.g. *provide + information/insight* and *as can be seen in*). The results of this study also demonstrate the benefit derived from adopting a bottom-up approach to phraseology, as it identified a considerable number of verb-based patterns that are not found in existing corpus-driven academic phraseology lists.

**Keywords:** academic vocabulary, verbs, collocations, lexical bundles

## 1 Introduction

While research in English for Academic Purposes (EAP) discovered quite early on that academic disciplines can differ in the way they use language to construct knowledge, the increase in the number of general EAP courses across the world acted as a catalyst for the search of a teachable common-core (de Chazal 2013). This led to a growing interest in the linguistic devices found to be common to various academic disciplines, and thus useful for mixed groups of EAP learners. The linguistic device that has undoubtedly attracted the most attention is academic vocabulary, i.e. the vocabulary that is “neither highly technical and specific to a certain field of knowledge, nor obviously general in the sense of being everyday words which are not used in a distinctive way in specialized texts” (Baker 1988: 91). One of the reasons for this is that academic vocabulary is said to be the most difficult type of vocabulary for EAP learners, as it is “not central to the topics of the texts in which they occur” (Coxhead 2000: 214) and “tend[s] to pass unnoticed” (Granger 2017: 9) – as opposed to technical vocabulary.

To meet this need, quite a number of general academic vocabulary lists have been created. Before the advent of corpus linguistics, academic vocabulary lists were based on manual frequency analyses of small corpora (Campion & Elley 1971; Praninskas 1972), the annotations found in students’ textbooks (Ghadessy 1979; Lynn 1973), or a combination of both (the University Word List, Xue & Nation 1984). Twenty years later, the need for a more representative and up-to-date vocabulary list was felt, and EAP scholars set out to propose academic vocabulary lists based on the analysis of EAP corpora. The very first corpus-based academic vocabulary list, viz. the ‘Academic Word List’

(AWL, Coxhead 2000), quickly met with great success. The AWL is based on a 3.5-million-word corpus of academic texts in various disciplines and contains 570 word families (i.e. a headword and its inflectional and derivational affixed forms, e.g. *authority*, *authorities* and *authoritative*) sorted in decreasing order of word family frequency. Notwithstanding its popularity, this list also met with some criticism. For example, it excludes high-frequency words on the grounds that learners should already master these vocabulary items. Research has however demonstrated that such vocabulary items can have academic uses as well, e.g. *gain weight* vs. *gain insight* (Martínez et al. 2009; Paquot 2007; Schutz 2013). Another issue is its organization according to word family. It has been shown, for instance, that some members of word families are not always very frequent in academic English (e.g. *establish* vs. *disestablish*) and do not always share the same core meaning as their headword (e.g. *react* vs. *reactivation*) (Gardner & Davies 2013). EAP scholars have recently attempted to address these weaknesses by using more empirical vocabulary extraction methods based on statistical analyses and by analyzing word lemmas only. This resulted in the creation of the ‘Academic Keyword List’ (AKL, Paquot 2010) and the ‘Academic Vocabulary List’ (AVL, Gardner & Davies 2013). The former is based on a two-million-word corpus of academic texts and contains 930 words. The latter is based on a 120-million-word corpus and contains over 3,000 words.

While such lists have attracted considerable attention and have been extensively used to write EAP teaching materials, the possible existence of general academic vocabulary has however also been questioned. Hyland and Tse (2007: 238), for example, question “the assumption that a single inventory can represent the vocabulary of academic discourse and be valuable to all students irrespective of their field of study”. They support their argument by showing that a number of vocabulary items found in the AWL (1) are not evenly distributed, and (2) show semantic variation across academic disciplines (e.g. the noun *volume* is mostly used to refer to a book in applied linguistics and sociology whereas, in the hard sciences, it refers to a type of quantity; *ibid.*: 246). Those scholars defending the idea of general academic vocabulary, on the other hand, argue that potential cross-disciplinary variation is not a reason to “throw out generalized word lists altogether” (Gardner & Davies 2013: 6). Gardner and Davies (*ibid.*: 2), for instance, underline the importance of such lists in helping EAP practitioners establish learning goals and design learning materials and tools. Granger and Paquot (2009) and Ming-Tzu and Nation (2004) further argue that it is possible to teach many vocabulary items by focusing on the central concept found behind the variety of uses: e.g. the verb *measure* should be described as referring to the ‘action of determining the size, amount, level, etc. of something’ despite the fact that disciplines use different methods, data and criteria (Granger & Paquot 2009). When analyzing the weight of general academic verbs compared to that of discipline-specific verbs, Schutz (2013) demonstrated the importance of general academic verbs as they represent over half of the verb tokens occurring in a corpus of research articles in business, linguistics and medicine; the verbs that were considered as discipline-specific only represented around 5% of the verb tokens occurring in each discipline. These results provide additional evidence that general academic vocabulary lists should not be discarded, but rather further investigated to best help EAP practitioners and learners.

The aim of the present paper is to provide insight into the reliability of general academic vocabulary lists by analyzing the collocations and lexical bundles used with general academic verbs in three strongly contrasting academic disciplines: business, linguistics and medicine. More specifically, the objective is to determine the extent to which general academic verbs have shared academic meanings and phraseological patterns. While the phraseology of academic English has been the object of quite a number of studies (see for example, Ackermann and Chen (2013) and Durrant (2009) for an analysis of collocations, and Simpson-Vlach and Ellis (2010) for an analysis of lexical bundles), none of these studies fully answer the criticism leveled against general academic vocabulary lists, as they adopted a textual rather than a lexical approach to phraseology. To best inform what can be

called the specificity debate (i.e. the debate centered around the question as to whether general EAP teaching is worthwhile, and thus the degree of specificity that such courses should adopt) this paper takes general academic verbs as a starting point to better identify their general academic uses vs. their discipline-specific uses.

## 2 Methodology

### 2.1 The Data

This study makes use of a sub-corpus of the *Louvain Corpus of Research Articles*<sup>1</sup> which totals 3,035,510 words and contains 421 research articles from peer-reviewed top-rated journals in three different academic disciplines: business, linguistics and medicine (hereafter LOCRA, BUS, LING and MED) (cf. Table 1).

Table 1: The LOCRA sub-corpus.

Disciplines	Number of texts	Number of words
Business	116	1,053,479
Linguistics	109	1,004,829
Medicine	196	977,202
TOTAL	421	3,035,510

In order to identify the verbs occurring in LOCRA, the three sub-corpora were lemmatized and POS-tagged with WMatrix (Rayson 2009) using the Constituent Likelihood Automatic Word-tagging System (CLAWS7) (Gardside & Smith 1997). As the tagset includes different tags for each verb form (e.g. VV0 for the base form or VVD for the past tense), a Perl program was applied to the CLAWS output so as to simplify the verb tags and conflate them into a single VV tag (cf. Granger & Paquot 2009).

### 2.2 Academic Verb Selection

Rather than selecting verbs that are found in existing general academic vocabulary lists, this paper zooms in on the verbs that stand out as being typical of the academic corpus under study. To do so, I join the forces of two different vocabulary extraction methods which have hitherto never been combined: the keyness analysis (e.g. the AKL, Paquot 2010) and the analysis of traditional frequencies (e.g. the AWL, Coxhead 2000). The novelty of this combined selection procedure is that it takes into account different types of verbs that have never been considered so far in the context of academic English (e.g. *find*, *make* and *see*) (cf. Schutz 2013; 2017). To be considered as general academic verbs, the verbs occurring in LOCRA had to be identified as either key (Scott 2001) or highly frequent across BUS, LING and MED. Key verbs are those which “occur with unusual frequency in a given text” (Scott 2001: 236) when compared to a “strongly contrasting reference corpus” (Tribble 2001: 396). In this study, WordSmith Tools 5 (Scott 2008) and a corpus of fiction writing, viz. the one-million-word fiction sub-corpus of the Baby British National Corpus, were used to extract the key verbs occurring in LOCRA. The verbs that were considered as highly frequent are those that appear among the top verbs covering up to 80% of the total number of verb tokens in each discipline (cf. Coniam 1999).

<sup>1</sup> <https://uclouvain.be/en/research-institutes/ilc/cecl/locra.html>

The comparison of the academic verbs occurring in BUS, LING and MED revealed that 177 academic verbs were used across all three disciplines. These cover 62%, 54% and 69% of the total number of verb tokens occurring in BUS, LING and MED, respectively (cf. Schutz 2013). To reduce this list to one that is more manageable for the purpose of this study, this paper focuses on the top 15 general academic verbs occurring in each discipline. When merging the top 15 academic verbs used in each discipline, the final verb list totals 31 verbs (see Table 2), which represent a sizeable 26% of the verb tokens occurring in LOCRA.

Table 2: Top 31 academic verbs occurring across BUS, LING and MED.

*appear, associate, base, consider, compare, describe, determine, develop, examine, express, find, follow, give, include, increase, indicate, influence, involve, make, observe, occur, perform, provide, receive, relate, report, see, show, suggest, take, use*

### 2.3 The Analysis of Phraseology

For the analysis of phraseology, this study makes use of automated tools to analyze co-occurrence and recurrence. While the former seeks to identify the words that co-occur with a specific node more often than by chance (viz. significant collocates; Sinclair et al. 1970: 150) the latter focuses on “recurrent expressions [of three or more words], regardless of their idiomaticity, and regardless of their structural status” (viz. lexical bundles; Biber et al. 1999: 990). The reason for choosing these two types of analyses is that they describe two different aspects of multi-word units (MWUs) that have rarely been studied concurrently. It is therefore hoped to provide a more complete description of cross-disciplinary phraseology, and thus better inform the specificity debate.

The collocates that are used with the 31 verbs under focus were extracted thanks to the Word Sketch option of the Sketch Engine (SkE; Kilgariff, Rychly, Smrz & Tugwell 2004). This tool automatically extracts the collocates of a specific node (using the logDice measure; Rychlý 2008) and categorizes them according to their grammatical function. As illustrated in Figure 1, Word Sketch identifies, for example, the words *support*, *evidence* and *difference* as the object collocates of *find* in BUS. To ensure the pedagogical relevance of the collocates we focus on, an additional frequency threshold of a minimum of five occurrences with the node was set.

<b>find</b> (verb) Alternative PoS: <b>noun</b> (freq: 2)		BUS freq = <b>1,602</b> (1,056.81 per million)	
<b>object</b>		<b>pro_object</b>	
	<b>38.70</b>		<b>2.81</b>
support	<u>55</u> 10.18	themselves	<u>8</u> 10.75
evidence	<u>36</u> 9.84	them	<u>6</u> 8.58
difference	<u>32</u> 9.04	it	<u>26</u> 8.05

Figure 1: Word Sketch of *find* in BUS.

To compare the collocates used in BUS, LING and MED, a simple three-step procedure was adopted. First, the collocate lists obtained for each verb in each discipline were extracted from Word Sketch. The different collocate lists extracted for each verb were then automatically compared so as to identify the collocates that are used across BUS, LING and MED vs. those that are used in one discipline only, i.e. the potential cross-disciplinary and discipline-specific collocates.

As regards the analysis of recurrence, a Perl script was used to extract the three-to-10-word lexical bundles occurring at least five times with the academic verbs under focus in BUS, LING and MED. After having identified the bundles used with each verb, the script then automatically generated the list of shared and discipline-specific bundles. While this considerably accelerated the extraction procedure, the bundle list then needed to be cleaned up as the output also contained, for example, bundles that did not include the actual verbs under focus (e.g. *increasing number of* or *the following variables*). Because of the extremely large number of bundles that were extracted for some verbs, this last step was restricted to the bundles that were found across BUS, LING and MED.

### 3 Results

#### 3.1 The Analysis of Co-occurrence

The significant collocates extracted from LOCRA were categorized into 10 different grammatical categories: subject, object, modifier, prepositional complement, *wh*- complement, infinitive complement, particle, object complement, adjective complement and *-ing* complement. In the following sections, we give particular attention to the subject and object collocates used in LOCRA, as these are the collocates for which we find the most significant and conclusive findings.

##### 3.1.1 The Shared and Discipline-specific Collocates

Out of the 31 verbs under focus, 22 verbs were found to share significant subject collocates across all three disciplines: *appear, associate, consider, compare, describe, examine, find, follow, include, indicate, influence, involve, make, observe, provide, receive, report, see, show, suggest, take* and *use*. As regards object collocates, 14 verbs were found to show cross-disciplinary similarities: *base, compare, examine, find, give, include, increase, make, perform, provide, report, show, take* and *use*. Tables 3 and 4 list the shared subject and object collocates that were identified in LOCRA and indicate which academic verbs they were generally found to co-occur with. The qualitative analysis of these collocates revealed that most of them could be grouped under various research-related semantic categories. The majority of the shared subjects were grouped under the categories RESEARCHER(S), RESEARCH, FRAMEWORK and INFORMATION (cf. Table 3). Most of the shared objects were grouped under the categories FRAMEWORK/METHOD, INFORMATION, RELATIONSHIP, RESEARCH and PHENOMENON (cf. Table 4). The remaining shared collocates were grouped under the category OTHER.

As can be seen from below, the shared collocates identified in LOCRA clearly relate to the core business of research, irrespective of the discipline. They are all (except for the subjects *it, they* and *we* and the objects *it, place, time* and *detail*) listed in recent general academic vocabulary lists (the AKL and/or the AVL). However, only a couple of the shared collocational pairs identified in LOCRA also appear in existing general academic collocation lists: only 10 verb-object patterns were found to overlap with those listed in the Academic Collocation List (ACL, Ackermann & Chen 2013) (e.g. *use + approach/method/strategy/etc.* and *provide + opportunity/data/support/etc.*). The reason for this small overlap is that Ackermann and Chen adopted a textual approach to phraseology whereas we adopt a lexical approach. In other words, Ackermann and Chen concentrated on the most frequent collocational pairs appearing in academic English, whereas this study focuses on those used with a particular node.



Table 3: Shared subject collocates occurring in LOCRA.

Semantic category	Shared subjects	+	Shared academic verbs
RESEARCHER(S)	we, they, study (metonymic use)	+	<i>compare, consider, describe, examine, find, follow, include, make, observe, provide, report, see, show, suggest, take, use</i>
RESEARCH	study, analysis		<i>appear, include, indicate, involve, show, suggest</i>
FRAMEWORK	model		<i>suggest</i>
INFORMATION	score, evidence, data, factor, result, finding		<i>indicate, influence, show, suggest</i>
OTHER	it, they		<i>appear, associate, provide, receive, show, suggest, use</i>

Table 4: Shared object collocates occurring in LOCRA.

Shared academic verbs	+	Semantic category	Shared objects
<i>use</i>	+	FRAMEWORK/ METHOD	<i>approach, criterion, instrument, measure, method, model, procedure, strategy, system, technique, test</i>
<i>compare, find, give, increase, provide, report, show, use</i>		INFORMATION	<i>data, detail, estimate, evidence, explanation, information, insight, level, number, rate, result, sample, score</i>
<i>compare, examine, find, show</i>		RELATIONSHIP	<i>correlation, difference, effect, relationship</i>
<i>base, include, perform, use</i>		RESEARCH	<i>analysis, study</i>
<i>show</i>		PHENOMENON	<i>pattern</i>
<i>give, make, provide, take</i>		OTHER	<i>advantage, assumption, contribution, decision, it, opportunity, place, rise, time, support</i>

The quantitative analysis of what these collocates cover compared to those that are discipline-specific revealed a rather complex picture of verb patterning: academic verbs appear to show different phraseological preferences according to the discipline they are used in. For example, our results indicate that the verbs under focus show varying degrees of formulaicity according to the discipline they occur in (e.g. the verb *receive* seems to have preferred discipline-specific collocates in medicine, such as *infant/patient/mouse/women + receive + care/therapy/placebo*, but not so much in linguistics). Similarly, some verbs appear to have shared collocates showing different coverage values in different disciplines (e.g. the shared subject collocates of the verb *find*, viz. *we* and *study*, cover 86% of *find*'s total number of subject co-occurrences in medicine while they only cover 36% in linguistics). Despite this finding, our results suggest that, generally speaking, both commonalities and discipline-specificities are important. For quite a number of verbs, whether they show a preference for shared and/or discipline-specific collocates, we found that both types of collocates are often sufficiently frequent to be considered in the context of EAP teaching. Typical examples of such verbs can be found in Table 5.

A closer look at each grammatical category also reveals interesting trends as regards the frequency of shared subject and object collocates. As can be seen from Figure 3, the shared subject collocates often tend to cover more than those that are discipline-specific. Figure 4, on the other hand, shows that the shared object collocates mostly cover as much as or less than the discipline-specific object collocates. Bearing in mind the complexity of cross-disciplinary verb patterning and also the fact that

Table 5: Examples of frequent shared and discipline-specific collocates.

Verbs	Shared subject collocates (coverage/total number of subject co-occurrences)	Discipline-specific subject collocates (coverage/ total number of subject co-occurrences)
<i>show</i>	study, analysis, we, data, etc. (MED = 35%)	biopsy, cell, microscopy, etc. (MED = 25%)
<i>report</i>	study, we, they (LING = 25%)	he, learner, student, etc. (LING = 28%)
Verbs	Shared object collocates (coverage/total number of object co-occurrences)	Discipline-specific object collocates (coverage/ total number of object co-occurrences)
<i>provide</i>	evidence, insight, information, etc. (BUS = 29%)	access, firm, value, etc. (BUS = 22%)
<i>show</i>	evidence, correlation, result, etc. (LING = 19%)	interaction, meaning, variation, etc. (LING = 15%)

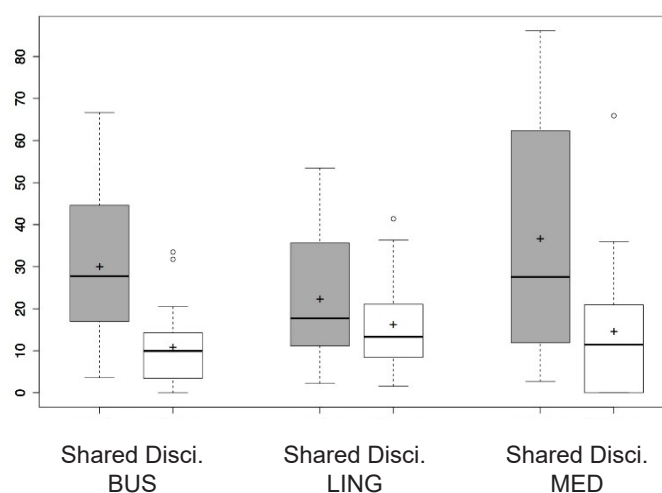


Figure 2: Coverage (%) of the shared and discipline-specific subject collocates used with the 22 verbs sharing subject collocates.

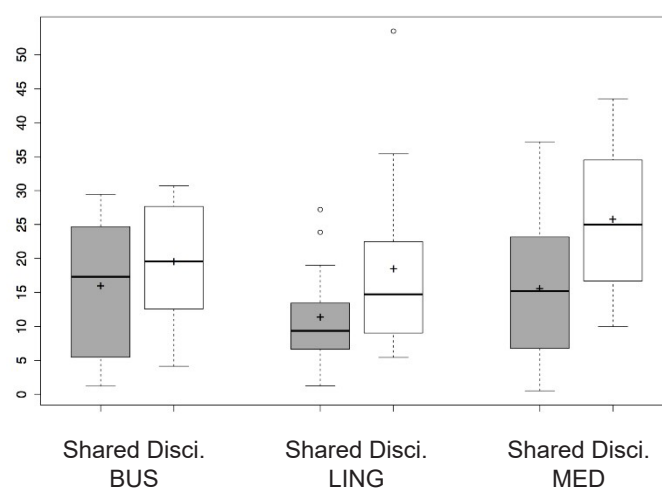


Figure 3: Coverage (%) of the shared and discipline-specific object collocates used with the 14 verbs sharing object collocates.

box plots can oversimplify results, this finding is still particularly interesting given that existing EAP collocation lists contain very few subject-verb collocational pairs: the ACL does not list any subject-verb patterns and Durrant's (2009) collocation list only lists three (*figure + shows*, *we + assume* and *these + suggest*). Our results, however, suggest that, given the high coverage this type of collocate can have for some verbs (e.g. *study/we + find* and *data/result/study/etc. + suggest* in BUS, LING and MED), shared subject-verb collocational pairs should also have a place in general academic collocation lists.

### 3.1.2 Core Academic Meaning and Semantic Variation

While the results presented above revealed that the verbs under focus can indeed show variation in collocational patterning, the contextual analysis of the shared and discipline-specific collocation pairs shows that, despite this, many of the academic verbs under investigation can also be used in similar meanings across BUS, LING and MED to describe/report on results and information, and report on research activities, for example. Only a handful of verbs were found to have discipline-specific uses and meanings.

The collocational patterns used to describe/report on results and information in LOCRA contain the verbs *appear*, *associate*, *give*, *find*, *indicate*, *observe*, *provide*, *relate*, *report*, *show* and *suggest*. For most of these verbs, our study identified both shared and discipline-specific significant collocates. For example, the verbs *find* and *observe* were most often used in the meaning 'discover/notice something after a careful examination of data/results/etc.'<sup>2</sup> no matter the context they were used in (see examples 1-6). To take another example, the verb *report* means 'provide information about something' whether it is a researcher, a study, a respondent, a learner or a patient that reports something. The only verbs for which we found no real discipline-specific collocates were the verbs *indicate* and *suggest*. When performing this rhetorical function, these verbs were used in the following collocational patterns: RESEARCH(ER)/ INFORMATION / FRAMEWORK + *suggest* and RESEARCH(ER) / INFORMATION + *indicate*.

- (1) [...] we **found** the *entrepreneurs* [...] to be heterogeneous with respect to their initial growth intentions. (BUS)
- (2) As a result, hardly any evaluative *words* were **found** in this move of the abstract. (LING)
- (3) Iron *deficiency* was **found** in 46.6% of case patients and 69.4% of controls. (MED)
- (4) *Spreitzer and Quinn (1996)*, for example, **observed** this social support *effect* in Ford's transformational change programme. (BUS)
- (5) In our data, we **observe** that Mandarin 'meiyou' acts as a signal for self-repair or other-repair, particularly in repair types 1 and 2. (LING)
- (6) We also **observed** that greater increases in levels of the 3 biomarkers were associated with significantly higher all-cause and cardiovascular disease mortality [...]. (MED)

The verbs used to report on cognitive and physical research activities are the verbs *base*, *compare*, *determine*, *examine*, *include*, *perform* and *use*. Most of these verbs also occurred with both shared and discipline-specific collocates in LOCRA. However, in this case, cross-disciplinary variation was primarily to be found among the object collocates. This suggests that, while researchers in business, linguistics and medicine can undertake similar research activities (e.g. *compare*, *examine*, *use*), they also show differences in their object of study (e.g. *effectiveness*, *verbs* and *cells*) and methodology (*use + index*, *image* and *primer*). Despite the variation found among their object collocates, all the

<sup>2</sup> Note that all the definitions given in this section are (adapted) from either the *Macmillan Dictionary* or *Longman Dictionary*.

verbs listed above were, in most instances, found to have the same core meaning across BUS, LING and MED. For example, the prepositional verb *base on* was predominantly used in the passive voice in the sense ‘to use something as the thing from which something else is developed’ no matter the context it occurred in, as illustrated in examples 7-9. Similarly, the verbs *determine*, *include* and *perform* were used across BUS, LING and MED in the sense ‘discover something through the examination of evidence/data/etc.’, ‘take something into consideration for the reported study’ and ‘carry out/complete an action or an activity’, respectively.

- (7) Such *strategies* are **based on** two different dimensions [...]. (BUS)
- (8) The *task* is **based on** the alphabetic principle: words that have more sounds need more letters to represent those sounds. (LING)
- (9) Efficacy and safety *analyses* were **based on** the intention-to-treat population (that is, all persons who underwent randomization and received at least 1 dose of medication in the double-blind phase). (MED)

Among the 31 verbs we analyzed in this study, only seven were found to have discipline-specific uses in LOCRA: *develop* (MED), *express* (LING and MED), *give* (MED), *make* (BUS, LING and MED), *perform* (BUS), *receive* (MED) and *take up* (MED). In addition to its shared meaning ‘create a new product/method’ (e.g. with the objects *vaccine*, *approach*, *product* or *model*), the verb *develop* was found to mean ‘begin to be affected by a medical condition’ when co-occurring with subjects such as *cell*, *mouse* and *woman*, and objects such as *cancer*, *diabetes* and *lesion* in MED. When used, for example, with the subjects *clause*, *speaker* and *they* and the objects *proposition* and *meaning* in LING, we found that *express* means ‘utter or state’. In MED, on the other hand, *express* means to ‘produce something’ when used, for instance, with the subjects *cell* and *mice*, objects *gene* and *receptor*, and modifiers *constitutively* and *differentially*. As for *make*, we found that this high-frequency verb is very often used in delexical constructions typical of either BUS, LING or MED. In BUS, *make* co-occurred, for instance, with the nouns *payment*, *investment* and *purchase*. In LING, it co-occurred with *judgment*, *recording* and *suggestion*. Finally, in MED, *make* was used with *diagnosis*, *measurement* and *visit*. Thanks to its discipline-specific modifier collocates *effectively/well/etc.*, we found that the intransitive use of *perform* also has a meaning which seems to be more frequent in BUS: in the sense ‘do something with a particular amount of success’ when discussing the performance of firms and employees, for instance. Finally, the verbs *receive*, *give* and *take up* appeared to be used as technical terms in MED. When used with objects such as *treatment*, *placebo* and *injection*, *receive* and *give* were used in the sense ‘to have/give a particular treatment’. *Take up* was used in the sense ‘to absorb or incorporate into itself’.

### 3.2 The Analysis of Recurrence

While the results presented above already provide considerable insight into the specificity debate, what the analysis of recurrence offers in this study is additional evidence supporting a general approach to academic English. As can be seen from Table 6, 20 of the 31 verbs under focus were found to share frequent lexical bundles across BUS, LING and MED (minimum frequency of 20 occurrences per million words). A closer look at their coverage values reveals that 11 of these verbs are used in shared lexical bundles which cover, on average, at least 15% of the total number of verb tokens occurring in LOCRA. These verbs are *appear*, *associate*, *base*, *consider*, *determine*, *find*, *indicate*, *involve*, *relate*, *show*, *suggest* and *use* (see coverage values in Table 6). In other words, about every six occurrence (at the minimum) of these verbs is used in a shared lexical bundle in LOCRA. As illustrated in Table 7, a closer look at these bundles further shows that many of them represent phrases that would have not been found in an analysis of lexical co-occurrence and that could prove useful to

Table 6: Frequent shared verb-based lexical bundles in LOCRA:  $\geq 20$  occurrences per million words.

<b>appear</b>	<i>appears to be, appear to be, it appears that</i> (mean coverage = 37%)
<b>associate</b>	<i>associated with the, is associated with, associated with a, are associated with, be associated with, were associated with, associated with an, not associated with</i> (mean coverage = 32%)
<b>base</b>	<i>based on the, is based on, based on a, are based on, was based on, is based on the</i> (mean coverage = 39%)
<b>compare</b>	<i>as compared with, compared with the</i>
<b>consider</b>	<i>considered to be</i>
<b>describe</b>	<i>as described in</i>
<b>determine</b>	<i>to determine whether, to determine the, determined by the</i> (mean coverage = 34%)
<b>examine</b>	<i>to examine the</i>
<b>find</b>	<i>we found that, found to be, found that the, found in the, was found to, be found in, were found to, can be found, and found that, been found to, can be found in</i> (mean coverage = 22%)
<b>follow</b>	<i>followed by a</i>
<b>include</b>	<i>included in the, were included in</i>
<b>indicate</b>	<i>indicate that the, results indicate that, indicates that the</i> (mean coverage = 17%)
<b>involve</b>	<i>involved in the</i> (mean coverage = 16%)
<b>provide</b>	<i>to provide a</i>
<b>relate</b>	<i>related to the, is related to, be related to, are related to</i> (mean coverage = 23%)
<b>report</b>	<i>reported in the</i>
<b>see</b>	<i>can be seen, seen in the, see figure #</i>
<b>show</b>	<i>as shown in, been shown to, have shown that, shown in figure, has been shown, shown in table, table # shows, are shown in, shown to be, show that the, has been shown to, as shown in figure, has shown that, have been shown, showed that the, have been shown to, studies have shown, results show that, we show that, # shows the</i> (mean coverage = 26%)
<b>suggest</b>	<i>suggest that the, suggests that the, findings suggest that, results suggest that, this suggests that, we suggest that, suggesting that the, suggested that the, to suggest that, these results suggest, these findings suggest, these results suggest that</i> (mean coverage = 27%)
<b>use</b>	<i>used in the, was used to, can be used, used as a, be used to, by using the, were used to, are used to, used in this, is used to, was used as, was used for, by using a</i> (mean coverage = 17%)

EAP students no matter what their academic discipline: e.g. *appears to be, can be found in, results indicate that, as shown in figure* and *by using the*.

When comparing our shared bundles with those found in an existing list of shared academic lexical bundles, viz. the Academic Formulae List (AFL, Simpson-Vlach & Ellis 2010), we found that our results identified frequent shared lexical bundles for verbs that are not described at all in the AFL: for example, those used with the verbs *compare* (e.g. *as compared with*), *indicate* (e.g. *results indicate that*), *describe* (e.g. *as described in*) and *suggest* (e.g. *these findings/results suggest that*). One of the reasons why these bundles were not found in the AFL is because this list only contains the top 200 bundles used in academic English. Given that academic prose has been shown to prefer NP- and PP-based bundles rather than VP-based bundles (Biber et al. 2004; Biber et al. 1999), it is normal that very few of the top bundles contain verb phrases. It thus appears that our approach to lexical bundles can prove beneficial when analyzing individual words, as it enables the identification of frequent verb-based bundles that would not especially appear at the top of corpus-driven academic bundle lists.



Table 7: Examples of frequent shared lexical bundles in LOCRA.

<i>appears to be</i>	Trust <i>appears to be</i> the central component that enhances perceived quality [...]. (BUS) However, the trend in modern corpus construction <i>appears to be</i> toward bigger and broader. (LING) Elevated pulmonary capillary wedge pressure <i>appears to be</i> common in those with emphysema and may be an important determinant of pulmonary artery pressure in these patients. (MED)
<i>can be found in</i>	Other examples of counter-intuitive results <i>can be found in</i> the empowerment literature and many suggestions have been made as to why empowerment interventions do not succeed. (BUS) Sample items from the language test and the meta-language test <i>can be found in</i> the Appendix. (LING) Different DC subsets <i>can be found in</i> the lung, each with a functional specialization. (MED)
<i>results indicate that</i>	The <i>results indicate that</i> 135 firms [...] went public with founder CEOs. (BUS) The <i>results indicate that</i> [...] English-writing skills show up as early as the second grade. (LING) These <i>results indicate that</i> the assessment of immunogenicity after immunization with DNA alone is not a reliable measure of the priming ability of DNA candidate vaccines. (MED)
<i>as shown in figure</i>	<i>As shown in Figure 1</i> , stock prices rose after the unexpected deaths of such CEOs. (BUS) This consisted of 12 drawings accompanied by a verb and an NP, <i>as shown in Figure 2</i> , where the target response would have been “El vaso se rompi”. (LING) <i>As shown in Figure 4A</i> , apocynin treatment increased survival of ALS mice [...]. (MED)
<i>by using the</i>	It is hoped that <i>by using the</i> organizing framework of competence [...] some clarity is offered regarding the issues which are receiving empirical attention and existing gaps. (BUS) The notion of invariance was investigated <i>by using the</i> moving word task. (LING) All events were coded <i>by using the</i> Medical Dictionary for Regulatory Activities conventions. (MED)

## 4 Conclusion

The purpose of this paper was to take a closer look at the phraseological patterning of academic verbs with the aim of verifying the reliability of general academic vocabulary lists. To do so, this paper drew on an exploratory cross-disciplinary comparison of the collocational patterns and lexical bundles used with a set of general academic verbs. While our study is limited, for example, by the number of academic disciplines and verbs it focuses on, it nevertheless provides valuable insight into an ongoing debate for which, as pointed out by Nhã (2015: 43), EAP scholars currently lack empirical evidence to support their arguments; most scholars have indeed based their arguments (either in favor of or against general EAP teaching) on but a handful of illustrative examples (e.g. de Escorcia 1985; Bruce 2011; Granger & Paquot 2009; Ming-Tzu & Nation 2004). In this study, we provide empirical evidence that, in addition to showing cross-disciplinary differences, general academic verbs also seem to have general academic meanings, collocates (e.g. *study/results* + *suggest*) and lexical bundles (e.g. *it appears that*).

More specifically, the analysis of co-occurrence showed that academic verbs have both shared and discipline-specific significant collocates. However, in many cases, variation in collocational patterning does not appear to affect the core meaning of the academic verbs. While our results revealed a rather complex picture of cross-disciplinary variation, many of the verbs under investigation were used in

the same core academic meaning no matter the context; only seven verbs had real discipline-specific uses in LOCRA (in addition, some had cross-disciplinary uses). Our results also showed that cross-disciplinary similarities were often to be found among subject collocates rather than object collocates. This finding is particularly interesting given the fact that subject collocates receive considerably less attention in EAP teaching materials. This type of collocate would thus deserve further investigation in future research.

While the analysis of recurrence did not provide evidence as compelling as that provided by the analysis of co-occurrence, it did however provide additional support for our main finding: academic verbs are also used in frequent shared lexical bundles. In addition to completing our description of cross-disciplinary phraseology, the results yielded by this analysis are just as important given the salience and systematic functionality of lexical bundles (e.g. Biber & Barbieri 2007), the challenge MWUs represent for learners (e.g. Nation 2001) and the importance of lexical bundles for academic proficiency (e.g. Hyland 2008). It is therefore important to include such patterns in general EAP teaching material to best help EAP learners.

More generally, adopting a lexical approach rather than a textual one to phraseology proved particularly valuable in the context of the present study. While a large majority of frequency-driven studies of EAP phraseology adopt a corpus-driven approach, we decided to concentrate on the phraseological patterning of particular nodes. Not only did this enable us to provide a comprehensive picture of cross-disciplinary language variation, it also helped identify phraseological patterns that are not found in existing lists of general academic MWUs. To establish whether there is such a thing as general academic vocabulary, our study also suggests taking a lexical approach to better determine how vocabulary items are used across academic disciplines.

I acknowledge however that this paper was based on a very small corpus. It would thus be necessary to replicate this study using a corpus of academic English containing a wider variety of disciplines and larger number of texts to better identify, for example, the patterns that are cross-disciplinary and those that are discipline-specific. Nevertheless, it is hoped that this paper has shown how, conducted on a larger scale, cross-disciplinary comparisons along the lines presented here can highlight typical phraseological patterns which EAP teachers could use to raise their students' awareness as to how general academic vocabulary behaves across disciplines and in their own field of study.

## References

- Ackermann, K., Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. In *Journal of English for Academic Purposes*, 12(4), pp. 235-247.
- Baker, M. (1988). Sub-technical Vocabulary and the ESP Teacher: An analysis of Some Rhetorical Items in Medical Journal Articles. In *Reading in a Foreign Language*, 4(2), pp. 91-105.
- Biber, D., Barbieri, F. (2007). Lexical bundles in university spoken and written registers. In *English for Specific Purposes*, 26(3), pp. 263-286.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. In *Applied Linguistics*, 25(3), pp. 371-405.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, England: Longman.
- Bruce, I. (2011). *Theory and Concepts of English for Academic Purposes*. New York: Palgrave Macmillan.
- Campion, M., Elley, W. (1971). *An academic vocabulary list*. Wellington: New Zealand Council for Educational Research.
- Coniam, D. (1999). Second language proficiency and word frequency in English. In *Asian Journal of English Language Teaching*, 9, pp. 59-74.

- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. In *English for Specific Purposes*, 23(4), pp. 397-423.
- Coxhead, A. (2000). A New Academic Word List. In *TESOL Quarterly*, 34(2), pp. 213-238.
- de Chazal, E. (2013). The general-specific debate in EAP: which case is the most convincing for most contexts? In *Journal of Second Language Teaching and Research*, 2(1), pp. 135-148.
- de Escorcia, B. (1985). ESP and beyond: a quest for relevance. In R. Quirk, H. G. Widdowson (eds.), *English in the World: Teaching and learning the language and literatures*. Cambridge: Cambridge University Press, pp. 228-237.
- Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. In *English for Specific Purposes*, 28(3), pp. 157-169.
- Gardner, D., Davies, M. (2013). A New Academic Vocabulary List. In *Applied Linguistics*, 34(3), pp. 305-327.
- Garside, R., Smith, N. (1997). 'A hybrid grammatical tagger: CLAWS4'. In R. Garside, G. Leech & A. McEnery (eds.) *Corpus annotation: linguistic information from computer text corpora*. New York: Addison Wesley Longman, pp. 102-121.
- Ghadessy, P. (1979). Frequency counts, word lists, and material preparation: a new approach. In *English Teaching Forum*, 17(1), pp. 24-27.
- Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. In *Oslo Studies in Language*, 9(3), pp. 9-27.
- Granger, S., Paquot, M. (2009). In search of a General Academic Vocabulary: A Corpus-driven Study. In K. Katsampoxaki-Hodgetts (ed.) *Options and Practices of LSP Practitioners*. Crete: University of Crete Publications, pp. 94-108.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. In *English for Specific Purposes*, 27(1), pp. 4-21.
- Hyland, K., Tse, P. (2007). Is there an "Academic vocabulary"? In *Tesol Quarterly*, 41, pp. 235-253.
- Kilgariff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*. Bretagne: Université de Bretagne Sud, pp. 105-116.
- Longman Dictionary of Contemporary English Online*. Accessed at: <https://www.ldoceonline.com> [31/03/2018].
- Lynn, R. W. (1973). Preparing word lists: a suggested method. In *RELJ Journal*, 4(1), pp. 25-32.
- Macmillan English Dictionary Online*. Accessed at: <http://www.macmillandictionary.com> [31/03/2018].
- Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. In *English for Specific Purposes*, 28(3), pp. 183-198.
- Ming-Tzu, K., Nation, P. (2004). Word meaning in academic English: Homography in the academic word list. In *Applied Linguistics*, 25(3), pp. 291-314.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nhã, V. T. T. (2015). Should an ESP Course be Specific or General? A Literature Review of the Specificity Debate. In *VNU Journal of Science: Foreign Studies*, 31(4), pp. 37-45.
- Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens, & S. Gozdz-Rozkowski (eds.) *Corpora and ICT in Language Studies*. Frankfurt am Main: Peter Lang, pp. 127-140.
- Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.
- Praninskas, J. (1972). *American university word list*. London: Longman.
- Rayson, P. (2009). *Wmatrix: a web-based corpus processing environment*, Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/>.
- Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka, A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 6-9.
- Schutz, N. (2013). How specific is English for Academic Purposes? A look at verbs in business, linguistics and medical research articles. In G. Andersen, K. Bech (eds.) *English Corpus Linguistics: Variation in Time, Space and Genre*. Amsterdam: Rodopi, pp. 237-257.
- Schutz, N. (2017). *Verbs in English for Academic Purposes: a cross-disciplinary corpus driven study*. PhD thesis. Université catholique de Louvain, Louvain-la-Neuve, Belgium.
- Scott, M. (2001). Comparing corpora and identifying key words, collocations and frequency distributions through the WordSmith Tools suite of computer programs. In P. Ghadessy, A. Henry & R. Roseberry (eds.) *Small corpus studies and ELT*. Amsterdam: John Benjamins, pp. 47-67.

- Scott M. (2008). *WordSmith Tools 5*. Oxford: Oxford University Press.
- Simpson-Vlach, R., Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. In *Applied Linguistics*, 31(4), pp. 487-512.
- Sinclair, J., Jones, S. & Daley, R. (1970). *English lexical studies*. Department of English, University of Birmingham.
- Tribble, C. (2001). Small corpora and teaching writing: towards a corpus-informed pedagogy of writing. In M. Ghadessy, A. Henry & R. Roseberry (eds.) *Small corpus studies and ELT: theory and practice*. Amsterdam: Benjamins, pp. 381-408.
- Xue, G., Nation, I. (1984). A University Word List. In *Language Learning and Communication*, 3(2), pp. 215-229.

# Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX

*Arvi Tavast, Margit Langemets, Jelena Kallas, Kristina Koppel*

*Institute of the Estonian Language, Tallinn*

*E-mail: arvi@tavast.ee, margit.langemets@eki.ee, jelena.kallas@eki.ee, kristina.koppel@eki.ee*

## Abstract

The Institute of the Estonian Language is developing EKILEX, a new dictionary writing system for both semasiological dictionaries and onomasiological termbases. While the long-term vision is to have a single data source that provides consistent information about Estonian, the system also needs to cope with the multitude of existing datasets. In this paper, we present work in progress on modelling the data and importing an initial sample of legacy dictionaries. The data model is based on an m:n relation between words and meanings, which are both unified across dictionaries, even while there still are separate dictionaries in the system. What is dictionary-specific is only the mapping between word and meaning. The importing of dictionaries has revealed various issues with data quality: ambiguities, underspecification, inconsistencies and conflicts. These need to be dealt with, if the long-term vision is to be achieved. We also outline the next steps of human- and machine-readable publishing, corpus connection and quantification (frequency, salience measures, etc.).

**Keywords:** data modelling, dictionary portal, interoperability, linked data, Estonian

## 1 Introduction

The Institute of the Estonian Language has been publishing dictionaries and termbases for decades, providing a comprehensive description of Estonian from a variety of perspectives. At the same time, the state of the art in lexicography has evolved from paper to electronic, introspective to empirical, manual writing to corpus-based generation, normative to descriptive, binary to quantitative, and human only to machine-readable. Software and storage formats have changed too, and one of the reasons for this is increased mutual awareness between linguistics and information technology.

Departments and working groups of the Institute have had a high degree of autonomy in compiling dictionaries, thus leading to the possibility of publishing inconsistent information, duplicating each other's work, and storing data in a way that is only semi-structured. Three separate dictionary writing systems are used for historical reasons, the dictionary data models are far removed from each other, and the whole dictionary system has gradually moved away from current thinking in lexicography.

The Institute has thus reached a point where changes are inevitable, first in the working methods, but consequently also in the tools used. In August 2017, development work was started for EKILEX, the Institute's new dictionary writing system, with the aim of addressing the most pressing issues and supporting the necessary changes in working methods.

In this paper, we report on the work in progress from the point of view of data modelling, including the importing of representative legacy dictionaries as a stress test for the new model.

First we describe what the existing datasets look like, and where the problems are that have caused the Institute to initiate the development of yet another dictionary writing system. We continue by referencing currently existing standards for lexical data representation. This is followed by three sections about



the work in progress itself: data modelling, including comparison to the referenced standards, data import and data harmonization. In the Discussion section, we explain the rationale behind some of the more difficult or controversial design choices, as well as a number of lessons already learned during the project. We conclude by outlining some directions for future work: electronic publishing for humans, connecting dictionaries to corpora, machine-readable publishing, and quantification of lexical data.

## 2 The Current Situation

The Institute is currently using three separate dictionary writing systems for its dictionaries and termbases:

- EELex<sup>1</sup> (Langemets, Loopmann & Viks 2010; Jürviste et al. 2011) was developed in-house from 2003 to 2015 and currently holds more than 70 dictionary databases of different types. It started out as an XML database, but for performance reasons was later transferred to a mixed model storing chunks of XML in a relational database. EELex predominantly uses semasiological data models and is highly customizable. At a late stage in its development support for onomasiological data structures was added, but it has been rarely used. For electronic publishing, a separate web interface is developed for each dataset, with automatic nightly data transfers.
- Termeki<sup>2</sup> was originally developed from 2007 to 2015 by Werkdata Ltd, and is still available commercially as termbases.eu.<sup>3</sup> Since 2012, a contract with Werkdata has allowed the Institute to provide it for free to Estonian terminologists, and it has mainly been used outside of the Institute. It is a relational database system with a partially customizable onomasiological data model, and has been used for about 40 termbases and one bilingual general language dictionary. Electronic publishing is implemented by allowing anonymous users restricted access to the same database.
- Multiterm,<sup>4</sup> a commercial product using XML technology and a partially customizable onomasiological data model, is used for two major termbases by the terminology department at the Institute. Electronic publishing in our current setup requires manual data transfers, which are only undertaken once a month.

Disparate data models have been used, especially in EELex, where a new data model has been custom developed for each new dictionary. Such flexibility was originally been designed to accommodate the heterogeneous wishes of dictionary authors, and has fulfilled this objective well: each author has obtained a data model of their choice. However, the results are not necessarily in line with current thinking in lexicography, the datasets are disconnected, information is duplicated and inconsistent across datasets, and the same information may be located differently in the model depending on the dataset.

All three have data models with a 1:n relation between form and meaning. One word has several meanings in the semasiological case, and one concept has several terms in the onomasiological case. A shortcoming of both is non-normalized data: information on the n-side of the 1:n relation is duplicated, causing inconsistencies due to human error (see Figure 2 for examples).

Especially in the two XML-based systems, the elements on the n-side of the 1:n relation are mostly plain text values, rather than entity references, making them ambiguous. In some newer datasets, homonym and meaning numbers may be included, but mostly the reference only consists of the target headword as a character string. This is understandable, considering that the only use case the authors

1 <https://eelex.eki.ee> [18.5.2018]

2 <https://term.eki.ee/> [18.5.2018]

3 <https://www.termbases.eu/> [18.5.2018]

4 <https://www.sdl.com/software-and-services/translation-software/terminology-management/sdl-multiterm/> [18.5.2018]

originally had in mind was a human reader, who has no difficulty navigating various meanings of a word. In addition to problems with machine-readable publishing, data reuse and linking of dictionaries, this solution also makes it impossible to automatically enforce internal consistency.

There are violations of atomicity and other abuses of each data model, for instance a definition and its source(s), or multiple definitions, in a single definition field; duplicated classifier codes with one of them containing a typo; domain labels entered in the pronunciation field due to excessive difficulty of using the domain classifier, and so on.

Regarding electronic publishing, a major issue is that each dataset has a separate public interface, in the worst case requiring the user to perform 130+ searches in separate dictionaries with the same search term. There is no machine-readable publishing, apart from custom exports performed at the request of prospective users of the data. For ESTERM<sup>5</sup> and MILITERM,<sup>6</sup> the two termbases compiled in Multiterm, their monthly publishing interval is not nearly enough to serve current needs. There are also performance and usability issues due to architectural choices made years ago, including limited browser compatibility.

Recognizing these issues, the Institute has started developing EKILEX, a dictionary writing system to replace all three current systems for both semasiological and onomasiological data, and importing existing datasets into the new system.

### 3 Prior Work

Modern lexicography has shifted its focus from compiling stand-alone dictionaries to making lexicographic data findable, accessible, interoperable and reusable (FAIR<sup>7</sup> data). Lexicographers thus need to pay more attention not only to the quality of lexicographic data but also to the data modelling of lexicographic databases.

There are several frameworks that can be used as a starting point for the database model. The most common are Lexical Markup Framework (LMF; ISO 24613:2008)<sup>8</sup> and Text Encoding Initiative (TEI XML)<sup>9</sup> for lexical resources, and TEI-Lex0 Initiative (Bański, Bowers & Erjavec 2017) for encoding of retro-digitized dictionaries. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of a large number of individual electronic resources to form extensive global electronic resources (Francopoulo et al. 2006). The Text Encoding Initiative (TEI) is aimed at equipping scholars with markup suitable for describing the majority of textual forms (concerning lexicography, especially for printed dictionaries) and analytic approaches, and providing extension capabilities to encompass new or infrequently found phenomena. TEI-Lex0 aims at ‘formulating guidelines for the encoding of retro-digitized dictionaries by streamlining and simplifying the recommendations of the “Print Dictionaries” chapter of the TEI Guidelines’ (Bański et al. 2017: 485). LMF is widely used for building lexical resources, see e.g. Borin et al. 2012.

There are also models that use ontologies and are geared towards the conversion of lexical resources to linked data. These are the LEXicon Model for ONtologies (*lemon*)<sup>10</sup> (McCrae et al. 2012) and its

5 <http://termin.eki.ee/esterm/> [18.5.2018]

6 <http://termin.eki.ee/militerm/> [18.5.2018]

7 <https://www.force11.org/group/fairgroup/fairprinciples> [18.5.2018]

8 <http://www.lexicalmarkupframework.org/> [18.5.2018]

9 <http://www.tei-c.org/index.xml> [18.5.2018]

10 <http://lemon-model.net/> [18.5.2018]

recently developed OntoLex-Lemon model<sup>11</sup> (McCrae et al. 2017). *lemon* is a model for modelling lexicon and machine-readable dictionaries and is linked to the Semantic Web and the Linked Data cloud. Bosque-Gil et al. (2016) claim that *lemon* is a de-facto standard for representing lexical information in the Web of Data. This model was tested in several lexicographic projects and has proved its success. Tiberius and Declerck (2017) explored reusing, improving and optimizing a dictionary of contemporary standard Dutch (ANW) by porting some of its elements into modules of *lemon*. They claim that encoding information in *lemon* has a number of advantages, including better modularization of the data, linking to other (lexical) data as well as providing improved access to data. *lemon* has been chosen as the backbone of BabelNet's<sup>12</sup> lexical knowledge linked data representation.<sup>13</sup> McCrae et al. (2017: 590) state that dictionaries represented with *lemon* or OntoLex-Lemon can be easily integrated with other resources previously converted to the Resource Description Framework (RDF)<sup>14</sup> without any remodeling efforts.

## 4 Data Modelling for EKILEX

Development of EKILEX was started in August 2017 in cooperation with the software house TripleDev Ltd, and the first project stage with currently committed funding will last until the end of 2018.

Initial requirements for the data model are the following:

- Describe language, as opposed to describing dictionaries: combine legacy dictionaries into a single data source about the language, and treat both words and meanings as existing independently of whether any dictionary includes them or not.
- Represent both semasiological and onomasiological data.
- Accommodate all existing dictionaries and termbases.
- Enforce best practices in both lexicography and terminology.
- Support the authors in maintaining data integrity.
- Comply with any current or future standard of data exchange.

The long-term vision is to have a single data source that provides consistent and comprehensive information about Estonian words, combining the research done at all departments and working groups of the Institute. In that ideal situation, each author or working group would be enriching the database with, for example, collocations, Chinese translations, normative recommendations or other data according to their expertise, instead of working in isolation on a collocations dictionary, Estonian-Chinese dictionary or normative dictionary.

Realistically, however, we also need to cope with the current transition stage of still having a multitude of dictionaries, each with their own ideology, working methods, legacy data and (administrative or financial) publishing requirements. The authors are aware of the problems described above and unification is their long-term goal.

Considering this gap between vision and reality, the process agreed for the project is the following:

- Make sense of existing dictionaries with their peculiarities, and import them as they are. Only correct errors (duplicated or non-structured data) that can be corrected automatically, or that the

<sup>11</sup> <https://www.w3.org/community/ontolex/> [18.5.2018]

<sup>12</sup> <http://babelnet.org/> [18.5.2018]

<sup>13</sup> <http://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/> [18.5.2018]

<sup>14</sup> <https://www.w3.org/RDF/> [18.5.2018]

authors are able and willing to manually correct within the project schedule. The results must be publishable as separate dictionaries.

- While the imported material still contains inconsistencies, provide a clear path towards unification, so that authors will be able to use the new system to gradually improve data quality, by reconciling conflicts found within their own or between datasets. Already at this point the results must also be publishable as a single dictionary.

The development of EKILEX uses an agile methodology (Scrum<sup>15</sup>) and is driven by priorities set by the stakeholders as expressed in the initial discussions and during biweekly sprint planning meetings. For data modelling, we did not start from a ready-made standard, but analyzed customer requirements instead, and optimized our solution to the particular situation of the Institute. In the following sections we describe the main design choices, comparing them to LMF and OntoLex-Lemon where applicable.

#### 4.1 Word and Meaning

Considering the inherent data duplication issues of 1:n models for both semasiology and onomasiology, we use instead an m:n relation between word and meaning: one word can have several meanings, and one meaning (concept) can be referred to by several words (see Figure 1). This could also be described as reuse of word and meaning data. In relational database terms, this is implemented using a link table between Word and Meaning tables.

While this linking entity, called Lexeme in our model and representing ‘this word in this meaning’, started out as a purely technical link table, it turned out to be the central point of our data model in terms of relating to other entities: the majority of data items are parameters of the Lexeme. In OntoLex-Lemon, our Lexeme corresponds to Lexical Sense and works in the same way, “mapping from a word to a concept” (McCrae et al. 2017).

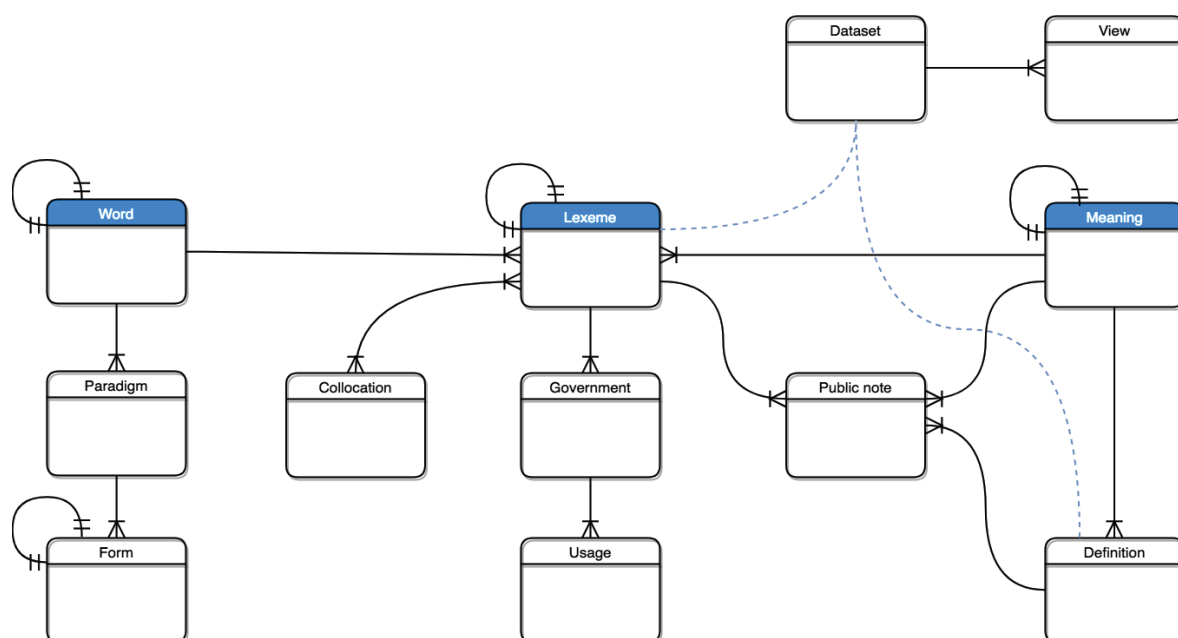


Figure 1: Simplified data model of EKILEX, highlighting the m:n relation between Word and Meaning through the Lexeme link table. Note that only Lexeme and Definition explicitly belong to a Dataset (dictionary or termbase). All other entities are common to all dictionaries in the EKILEX model, possibly being associated with a dictionary through Lexeme or Meaning.

15 <https://www.scrum.org/resources/what-is-scrum> [18.5.2018]

## 4.2 Dictionaries

During the current transition phase with multiple dictionaries, the Lexeme also carries dictionary-specific information, making its content effectively ‘this word in this meaning as described in this dictionary’.

So we are able to keep both words and meanings independent of dictionaries. Indeed, this makes theoretical sense too: words belong to the language, not any particular dictionary, and the same for meanings. What belongs to dictionaries is only the description of the relationship between words and meanings.

This design also works well with our vision-reality gap. When importing legacy dictionaries and finding two words with the same form, in most cases we have no way of telling whether these are two homonyms, one polyseme or simply data duplication. We can move on by importing them as two homonyms for the time being, with the option of deciding to combine them at a later stage. We apply a similar approach to meanings. During the import process we have no machine-readable information about meanings whatsoever, so whenever finding a meaning in the legacy files, we import it as a new meaning. Since words and meanings are dictionary-agnostic, the process of combining duplicates across dictionaries is exactly the same as within a dictionary.

## 4.3 Lexical relations

Lexical relations are represented using two distinct methods. The first is used for synonyms and translation equivalents, defined as words with the same meaning within one language or across languages. These are actually connected to the same meaning. There is no explicit synonymy or equivalence relation in addition to the connection through the meaning.

All other lexical relations are expressed using relations between meanings. So there are no explicit antonymy, hyperonymy, and so on relations between words either. Instead of recording that “dog” is a hyponym for “animal”, for instance, we record that dog is a type of animal. This is transparent for the user, so dictionary authors may continue thinking in lexical relations if they wish, even if they are stored as conceptual relations in the database.

## 4.4 Morphology and other dictionary-agnostic linguistic information

As the long-term vision no longer contains separate dictionaries, we are already moving towards centralizing some of the data elements.

One of the data categories that does (or should) not depend on the dictionary is morphology. While there are known differences between dictionary authors in which forms they consider legitimate, it does not make sense to list those differences without explanation. They should be either reconciled or, if the differences continue to be important for the authors, tagged according to their normative, stylistic or other status. In our model, a word has one or more paradigms (inflectional patterns in LMF terms), each containing one or more forms. Forms have written representations and may have various types of phonetic transcriptions and links to sound files.

The word itself does not have any linguistic representation. Instead, one or more of its forms can be marked as canonical, and the word gets its representation(s) from there. Forms can also be quantified, e.g. to not show rare forms to the L2 learner, or explicitly tagged as suitable for some use case.

Other clear examples of dictionary-agnostic data items include collocations and word-formation. The Dictionary of Estonian (DicEst), due to be published in 2018, is the first dictionary where these two will not be written from scratch, but reused from the Collocations Dictionary (COLL,



Kallas et al. 2015) and the Dictionary of Word-Families (Vare 2012), respectively. In EKILEX, the relations between linguistic items (e.g. derivational relations, compounds and their subterms, collocations, etc.) are represented using database relations between the corresponding entities.

## 5 Data import and harmonization

The minimum required selection of datasets to be imported during the first project stage was the following:

- At least one representative dataset from each source (EELex, Termeki, Multiterm), to verify that this source is readable and get an overview of the problems.
- The Dictionary of Estonian as the largest and most modern general dictionary, to serve as a basis for the database backbone.
- The Collocations Dictionary, due to its specific requirements and the fact that it is scheduled to be completed by November 2018 and published using the new system.
- Three resources with the Estonian-Russian language pair, as a requirement from one of the financiers.

When talking of 100+ databases the process of harmonization is the key concept. Harmonizing the lexical data involves an iterative process of capturing, defining, analyzing and standardizing the data. Harmonizing definitely improves the quality of the data by eliminating redundancies, inconsistencies and duplications, as well as facilitating the exchange of data and improving automation by ensuring interoperability (see Figure 2). The problem is that to a large extent this work is to be done manually by lexicographers editing the dictionaries, or half-manually, using some simple in-house tools for editing and adjusting the lexical data.

The backbone of the new EKILEX will be the corpus-based comprehensive scholarly Dictionary of Estonian (DicEst), which has been compiled at the Institute of the Estonian Language since 2010, and will be published online in our new dictionary portal Sõnaveeb ('Web of Words') in autumn 2018. The dictionary focuses on written Estonian, being the descriptive, not the normative dictionary. There are ca 110,000 words in the dictionary and when published it will then be constantly updated. Most of its elements have been imported into the new EKILEX model. The morphosyntactic properties of the words (for all dictionaries) will be imported from the Morphological Database that is currently being developed at the Institute.

The core entities of Word and Meaning, as well as more peripheral Morphology, Collocation, Usage Example and Etymology, and the like, are common to all dictionaries in the EKILEX model (note that the Lexeme is dictionary-specific, see Figure 1). This means that during the import process data needs to be unified across the legacy dictionaries – e.g. importing a headword as many times as it has legitimate homonyms, not as many times as it is found in the 100+ dictionaries to be imported. This is a nontrivial task even for the relatively clear case of homonyms. Not only may dictionaries differ in their level of detail for a particular headword, authors may also have various working definitions of what to consider homonyms in the first place. We currently use a combination of morphology-based word sense disambiguation, manual disambiguation and organizational measures (persuading authors to reach an agreement) for unifying the word list. Next in line for unification are example sentences, collocations and etymologies, followed finally by meanings.

The first task lexicographers were involved in with EKILEX was aligning and linking at the lemma (homonym) level. With the help of a special mini-tool we could connect homonyms across many dictionaries. Table 1 shows three homonyms and different forms of 'luup' from seven dictionaries.

In the dictionary portal, when searching ‘luup’, the user will get three homonyms (‘luup1’, ‘luup2’, ‘luup3’). Each of these is connected to the content (all data from entries) from several dictionaries. The morphophonetic data (the degree of quantity) in ET-RU (learners) and ORTH is harmonized via the Morphological Database.

Table 1: Three homonyms in seven contemporary dictionaries. DicEst = *Dictionary of Estonian* (to be published in 2018, the backbone of EKILEX), BasicDic = *Basic Estonian Dictionary* (2014), ET-RU (learners) = *Estonian-Russian Dictionary of Standard Estonian for Learners* (2011), ET-RU (general) = *Estonian-Russian I-V* (1997-2009), COLL = *Estonian Collocations Dictionary* (to be published in 2018), ET-FI (general) = *Estonian-Finnish* (to be published in 2018), ORTH = *The Dictionary of Standard Estonian ÕS 2018* (to be published in 2018).

DicEst	BasicDic	ET-RU (learners)	ET-RU (general)	COLL	ET-FI (general)	ORTH
<b>luup1</b> ‘loupe’	–	.luup I	luup I	luup	luup1	l’uup 1.
<b>luup2</b> ‘sloop’	–	.luup II	luup II	–	luup2	l’uup 2.
<b>luup3</b> NEW! ‘looper’	–	–	–	–	luup3	–

Another major challenge for importing existing dictionaries is that the information in them is often ambiguous or underspecified. Collocations, lexical relations and other references to entities in the same dictionary are increasingly expressed using relations, not unstructured text any more, but the target of that relation is still a string of characters, not an object reference. We solve these case by case. When first attempting to import a dataset, such ambiguities are logged for the dictionary owner to review and decide what to do. Some can be resolved using hints found elsewhere in the data, some can be manually disambiguated before the next import attempt, and for some, the dictionary owner may decide to omit them from the import altogether. The rest are usually more labor-intensive to resolve, so we import the ambiguity as it is, and leave data harmonization to be performed at a later time, when it is already in the new system.

Besides, lexicographers have been faced with the idea of ‘linking at sense level’. This is not an easy process, but EKILEX goes a step further than linking, actually combining equivalent meaning entities from separate dictionaries into a single entity. As a result, that single meaning entity will link things such as mini-definitions or glosses to longer definitions, collocations to senses of their counterparts, translation equivalents (from bilingual dictionaries) to senses in monolingual dictionaries, and so on. When looking closer at *luup1* ‘loupe’ we recognize instances that we have to harmonize (Figure 2):

**Five definitions** (in Estonian, all defining the same sense ‘loupe’):

- lihtne optikariist, mis annab esemeist suurendatud kujutise [DicEst]
- [BasicDic]
- [ET-RU (learners)]
- suurendusklaas [ET-RU (general)]
- suurendav optikariist [COLL]
- lihtne optikariist, mis annab esemeist suurendatud kujutise [ET-FI (general)]
- suurendusklaas [ORTH]

**Synonyms or candidates for synonyms** (in Estonian):

- suurendusklaas [DicEst, explicitly marked as a synonym]
- suurendusklaas [ET-RU (general), originally encoded as a definition]
- suurendusklaas [ORTH, originally encoded as a definition]

**Translation equivalents** (Russian, Finnish):

лупа, увеличительное стекло [ET-RU (learners)]  
 лупа [ET-RU (general)]  
 luuppi, suurennuslasi [ET-FI (general)]

**Usage examples** (in Estonian, almost the same combinations recurring in slight variations):

kümnekordse suurendusega luup [DicEst]  
 vanahärra uuris fotosid läbi luubi [DicEst]  
 tugev / terav / suurendav luup [COLL]  
 luupi kasutama / luubiga uurima / luubiga vaatama / luubiga lugema [COLL]  
 kümnekordse suurendusega luup [ET-FI (general)]  
 vanahärra uuris fotosid läbi luubi [ET-FI (general)]  
 uurib luubiga, läbi luubi postmarke [ORTH]

**Translations of the usage examples** (translations into Russian, Finnish):

kümnekordse suurendusega luup – kymmenkertaisesti suurentava suurennuslasi [ET-FI (general)]  
 vanahärra uuris fotosid läbi luubi – vanhaherra tutki valokuvia luupilla [ET-FI (general)]

Figure 2: Instances across dictionaries to be harmonized in the case of *luup* 'loupe'.

We can observe how fuzzy the boundary is between (short) definitions and synonyms, e.g., the term *suurendusklaas* 'magnifying glass' appears to serve as both in different original encodings. The duplication of the same material in different dictionaries has been unavoidable when compiling printed dictionaries as well as standalone and strictly separated dictionary databases (as in EELex up to now). In the case of the EKILEX model these are the inconsistencies.

## 6 Discussion

The EKILEX project has brought up a number of issues to be addressed and decisions to be made. Some of the choices described above have not been straightforward at all, and some have even been revisited and changed as a result of new information.

A major discussion point was whether to make the model recursive, i.e., unify all form-related entities (Word, Collocation, Usage Example, and maybe even Definition) into what is now the Word. We decided otherwise, and to have separate entities for each. The reason was that while these entities are theoretically similar and do share some important properties, notably that of having a meaning, many properties are not shared and the business logics applied to them are still very different. We thus chose a wider and shallower model over deep recursion, to keep queries and program logic simpler. Recursive RDF can still be exported from our model if needed.

The central idea of our model is the m:n relation between form and meaning. We do not know of any dictionary writing system that would implement this idea, at least not as radically as EKILEX (by not having any synonymy or equivalence relations at all). However, the idea itself is not new. In a well-hidden form it already exists in the LMF standard, where the Synset entity can be construed to correspond to what we call the Meaning. So while we do agree with Borin et al. (2012: 3599) that at first glance LMF looks unusably semasiological, it seems that theoretically the Synset entity there could be used to represent onomasiological data too, similar to Wordnet. In OntoLex-Lemon, the idea is more visible in the form of the Lexical Concept, relating m:n to the Word.

Despite prior familiarity with the data models of existing datasets, we underestimated the workload of data import. The initial plan was to complete the first round of imports by the end of 2017, but as of

this writing in late March 2018 it is still not completed. Technical implementation is not the only or perhaps not even the main reason for delays. Attempts to first create a mapping between the models and then to actually load the existing XML files have revealed decisions that authors were able to ignore in the semi-structured model, but that have to be decided now. The importing activity has even sparked discussions on the principles and objectives of some datasets, like who is the target group and what are their needs. Such questions do not have a single correct answer, resulting in lengthy discussions delaying the project.

There is a constantly nagging gap between the long-term vision and what is currently possible, given the legacy data. Namely, when we find words with the same written representation in several datasets when importing, and there is no information on how to combine them, we import them as homonyms. This is an obvious temporary solution for insiders familiar with the import process. For normal users, it looks like a simple UI issue (“you are displaying this word too many times”), while solving it would actually require a major undertaking of manual sense-level linking of each dataset to the backbone. This linking is still firmly in the plans, but the workload is daunting.

From the linked data perspective, we are not linking the various dictionaries of the Institute. Instead, we combine them into a single dictionary or Lexicon in the OntoLex-Lemon sense. This can then be published as linked data, if needed. The reason is that unlike the global community of linked linguistic data, dictionary authors are (or at least should be) under common management, following the same objectives and working methods, and capable of cooperation. This creates an opportunity to provide the added value of actually making the dictionaries consistent and non-redundant, in addition to making them link to each other.

## 7 Next steps

We are continuing with the process of data import and should release the first version of the dictionary writing system for lexicographers and terminologists in November 2018. In addition, the following next steps have been planned.

### 7.1 Dictionary portal

The user interface and the types of access to data depend very much on the data model behind the actual data. Access to data in dictionary portals ranges from searching different dictionaries (via linking) to searching in the data within the entry (Boelhouwer, Dykstra & Sijens 2017: 755). The user is generally expected to be capable of identifying and classifying dictionaries according to type. However, this might not be the best premise, as quite often there are dozens of dictionaries and databases on a website (e.g. the Estonian dictionary website<sup>16</sup>, European Dictionary Portal<sup>17</sup>), which makes it difficult for the user to decide which one is the right one.

Our near-future EKILEX-based dictionary portal Sõnaveeb (‘Web of Words’, to be launched in autumn 2018) is meant to serve human users as an aggregator with items of content collected to one web page and enabling access to data within several dictionaries. These days, when searching for what a word means or how it is translated, people do not necessarily realize that they are searching a dictionary. They are just looking for the answers to their questions. The variety of data now available means that it is possible to meet both learners’ productive as well as receptive needs.

<sup>16</sup> <http://portaal.eki.ee/sonaraamatud.html> [18.5.2018]

<sup>17</sup> <http://dictionaryportal.eu/> [18.5.2018]

The new portal is linked to the new Estonian Corpus for Learners 2018 (etSkELL)<sup>18</sup>. The Corpus was compiled using the Estonian GDEX module (Koppel 2017) to filter the sentences in the Estonian National Corpus 2017, which is the largest and newest Estonian corpus (about 1.1B tokens) in Sketch Engine (Kilgarri et al. 2004). GDEX (Good Dictionary Example, Kilgarri et al. 2008; Kosem et al. 2018) scores sentences according to how well they meet predefined conditions. All sentences that met the conditions of the hard classifiers were collected into the Corpus for Learners (others were removed). All sentences were then scored and reordered (with the highest scores at the top) using the soft classifiers of the Estonian GDEX module. The resulting Corpus contains about 248,000 words and about 25M sentences that derive from various media and scientific texts, fiction, Estonian Wikipedia and the Estonian Coursebook Corpus of CEFR-graded sentences.

When searching for a word, the portal Sõnaveeb directs the query to the corpus query system Korp<sup>19</sup> using an API, and a certain number of authentic example sentences is presented.

## 7.2 Machine-readable publishing

Since EKILEX stores data in a structured and normalized form, there will always be a mapping from our database to any existing or future standard of data exchange, including any that will be developed in the ELEXIS project. The mapping to OntoLex-Lemon is especially straightforward. Over the years, the Institute has received and fulfilled several requests for wordlists, usually in very simple text formats. While there have been no requests to access the Institute's datasets as linked open data, providing such access is technically possible.

## 7.3 Quantification

A future development that we want to prepare for with this data model is quantification. The model allows any relation to be quantified, from morphological preferences to the relation between a Word and its Meaning. The collocations that we import, for example, already have empirical frequency and salience measures attached, widening the selection of possible display methods for collocations. These measures themselves may pose additional temporary challenges, like undifferentiated frequency counts for homonyms due to lack of semantically tagged corpora, but we believe in empirically based quantification in the long term, and have already left room for this in the data model.

# 8 Conclusion

In this paper, we presented the data model of EKILEX, a new dictionary writing system for both semasiological dictionaries and onomasiological termbases. We also discussed various issues that arose in the process of importing legacy dictionaries into the new system, such as issues with data quality: ambiguities, underspecification, inconsistencies and conflicts.

The Institute of the Estonian Language has been using three separate DWSs (EELex, Termeki, Multiterm) for its dictionaries and termbases, which has resulted in disconnected datasets and duplicated, inconsistent information across these. All three DWSs have a 1:n relation between form and meaning – one word has several meanings in the semasiological case, and one concept has several terms in the onomasiological case. The data model of EKILEX on the other hand is based on an m:n relation between words and meanings – one word can have several meanings and one meaning (concept) can

18 <https://etskell.sketchengine.co.uk/run.cgi/skell> [18.5.2018]. The authors would like to thank Jan Michelfeit for compiling the corpus.

19 <https://korp.keeleressursid.ee/> [18.5.2018]



be referred to by several words. Words and Meanings are linked by Lexemes, which carry dictionary-specific information and which in our model represent ‘this word in this meaning as described in this dictionary’. We are keeping both Words and Meanings independent of dictionaries, both belonging to the language, not to any particular dictionary. Dictionary data is held by the Lexeme, i.e. the description of the relationship between Words and Meanings.

The long-term vision is to have a single data source (EKILEX) that provides consistent and comprehensive information about Estonian words, combining the research done at all departments and working groups of the Institute. The backbone of the new EKILEX will be the corpus-based comprehensive scholarly *Dictionary of Estonian* (DictEst), and other linguistic items (morphology, compounds, derivational relations, collocations, etymology) will be linked with DictEst.

## References

- Bański, P., Bowers, J., & Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, V. Baisa (eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference, Leiden, 2017*. Brno: Lexical Computing CZ s.r.o. Accessed at: <https://elex.link/elex2017/proceedings-download/> [18.5.2018].
- Boelhouwer, B., Dykstra, A., & Sijens, H. (2017). Dictionary portals. In P. A. Fuertes-Olivera (ed.), *The Routledge handbook of lexicography*. London and New York: Routledge, pp. 754–766.
- Borin, L., Forsberg, M., Olsson, L.-J., & Uppström, J. (2012). The open lexical infrastructure of Språkbanken. In *Proceedings of Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pp. 3598–3602. Accessed at: [http://www.lrec-conf.org/proceedings/lrec2012/pdf/249\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/249_Paper.pdf) [18.5.2018].
- Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., & Aguado-de-Cea, G. (2016). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, pp. 65–73. Accessed at: <http://www.citeulike.org/user/jgracia/article/14024090> [18.5.2018].
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., & Soria, C. (2006). Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Accessed at: [http://www.lrec-conf.org/proceedings/lrec2006/pdf/577\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/577_pdf.pdf) [18.5.2018].
- Jürviste, M., Kallas, J., Langemets, M., Tuulik, M., & Viks, Ü. (2011). Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In *Electronic lexicography in the 21st century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011*, pp. 106–112. Accessed at: <https://dialnet.unirioja.es/servlet/articulo?codigo=4567368> [18.5.2018].
- Kallas, J., Kilgariff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Ljubljana; Brighton: Trojina, Institute for Applied Slovene Studies; Lexical Computing Ltd.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.), *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425–432.
- Kilgariff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*, 105–115.
- Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks [Automatic detection of good dictionary examples in Estonian learner's dictionaries]. *Eesti Rakenduslingvistika Ühingu Aastaraamat [Estonian Papers in Applied Linguistics]*, 13, 53–71. <https://doi.org/10.5128/ERYa13.04> [18.5.2018].
- Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J., & Tiberius, C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*.
- Langemets, M., Loopmann, A., & Viks, Ü. (2010). Dictionary management system for bilingual dictionaries. In S. Granger, M. Paquot (eds.), *eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, pp. 425–429.

- McCrae, J. P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), 701–719.
- McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference, Leiden, 2017*. Brno: Lexical Computing CZ s.r.o., pp. 587–597. Accessed at: <https://elex.link/elex2017/proceedings-download/> [18.5.2018].
- Tiberius, C., & Declerck, T. (2017). A lemon model for the ANW dictionary. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference, Leiden, 2017*. Brno: Lexical Computing CZ s.r.o., pp. 237–251. Accessed at: <https://elex.link/elex2017/proceedings-download/> [18.5.2018].
- Vare, S. (2012). *Eesti keele sõnapered [Dictionary of Estonian Word-Families]*. Accessed at: <http://www.eki.ee/dict/sp/> [18.5.2018].

## Acknowledgements

This work has been supported by receiving funding from the European Regional Development Fund. Project EKI-ASTRA 2014-2020.4.01.16-0034



# On the Interpretation of Etymologies in Dictionaries

**Pius ten Hacken**

Leopold-Franzens-Universität Innsbruck

E-mail: [pius.ten-hacken@uibk.ac.at](mailto:pius.ten-hacken@uibk.ac.at)

## Abstract

Etymological information is an expected type of information in historical dictionaries, but it also appears in many general dictionaries, while it is the key information in etymological dictionaries. Etymologies are generally considered to trace the history of words. However, the notion of a *word* in this statement is an abstraction in more than one way. First, the questions of which forms and which meanings should be placed together as a word does not have an obvious answer. Moreover, the question of which words there are in a language at a particular time cannot be answered on a purely empirical basis. In the light of such observations, I show that what is recorded in an etymology can best be interpreted as the history of the motivation speakers had for the combination of a particular form with a particular meaning. This does not subtract from the value of etymological information, but gives a linguistically sound interpretation of what etymologists have tried to achieve.

**Keywords:** etymology, historical dictionaries, dictionary interpretation

## 1 Introduction

Etymological information is a standard type of information for historical dictionaries, and it is not accidental that Considine (2013) treats etymology and historical dictionaries in one chapter. However, etymological information is also common in general dictionaries, and it is central in specialized etymological dictionaries. An example of an etymology in a historical dictionary is the one for *translation* in the OED (2018), given in (1).

### (1) *translation*

< Old French *translation* (12th cent. in Godefroy *Compl.*), or < Latin *translātiō-em* a transporting, translation, noun of action < *translāt-*, participial stem of *transfere* to TRANSFER V.

In (1), we see that there are two possible historical paths, one from Old French and one from Latin. For the Latin word, which underlies the Old French one, a morphological analysis is given.

In etymological dictionaries, e.g. Kluge (1995), information is much more extensive. Entries do not only give the forms recorded in various periods or reconstructed for periods from which no records are available, but also give arguments for or against the hypothesis of certain relationships. Moreover, entries may be followed by a list of cognates in other languages and references to discussions of the etymology of the word in the scientific literature.

Because of space constraints, general dictionaries tend to be more selective in the etymological information than historical dictionaries. The COED (2011) thus does not give an etymology for *translation*, but only (2) in the entry for *translate*.

### (2) *translate*

ORIGIN ME: from L. *translat-*, *transfere* (see TRANSFER)

It is clear that (2) summarizes information from (1), which is not surprising given the relationship between the COED and OED. The information about the possible Old French origin is only given in the etymology of *transfer*.

Here, I will discuss the interpretation of statements such as (1) and (2), as well as more elaborate etymologies such as given in Kluge (1995). First, section 2 will address the question of what constitutes a word. This question is central to the argument, because etymologies are given for words, e.g. *translation* in (1), and they refer to words, e.g. *transfere* in (1). Then, section 3 will turn to the sources of etymological information. In section 4, I argue for a particular interpretation of etymological statements that is in line with linguistic insights about the nature of words. Finally, section 5 summarizes how etymologies in dictionaries can be seen as well-founded pieces of information.

## 2 The nature of words

Etymology is concerned with the history of words. What constitutes a word, however, is not empirically verifiable, because a *word* is an abstraction from empirical data. This problem is recognized explicitly by Durkin (2016: 236). In order to understand the different dimensions of abstraction involved, it is useful to start from Saussure's (1916) theory of the word as a *signe*, consisting of a *signifiant* (form) and *signifié* (meaning). Three dimensions of abstraction can be identified. In section 2.1, I will address the abstraction of historical stages, in section 2.2, the synchronic extent of the form and meaning, and in section 2.3, the abstraction of named languages, such as English.

### 2.1 The history of a word

In determining what is a word, a first dimension of abstraction is the historical one. This can be illustrated with the start of the entry for *lügen* ('lie, tell lies') in Kluge (1995), given in (3).

(3) **lügen** *stV* (< 8. Jh.). Mhd. *liegen*, ahd. *liogan*, as. *liogan* aus g. \**leug-a-*

In (3), the verb is characterized as a strong verb (*stV*) attested from the 8<sup>th</sup> century onwards. Five forms are given. The modern form is the headword. It is followed by documented forms from Middle High German (mhd., 1050-1350), Old High German (ahd., 750-1050) and Old Saxon (as., 8<sup>th</sup>-12<sup>th</sup> century), as well as a reconstructed earlier Germanic form. In etymologies, reconstructed forms for which no corpus evidence is available are indicated by an asterisk. The question is, then, in what sense the five forms given in (3) are forms of the same word.

Saussure (1916) objects to statements about the history of individual sounds. He gives the example of Latin *conficio* ('produce') and *facio* ('make'), where it would be wrong to say that the *-a-* in *facio* has become an *-i-* in *conficio*. We have to consider the entire system of oppositions synchronically before comparing the systems diachronically. This not only applies to sounds, but also to words. Looking at the history of individual words, as in (3), is therefore problematic. How can we determine that they are actually the same word? What is the status of the individual word forms in (3)? Even in a fairly straightforward case as in (3), grouping together these historically attested and reconstructed forms as a word is an abstraction. Although it is possible to come up with reasons for a particular grouping, no such grouping can be taken as an empirical fact. Groupings only emerge from the combination of a theory and empirical data.

Durkin (2016: 237-241) gives a more elaborate example involving *post*. The OED gives 22 homonyms of different word classes, some derived from Italian *posta* ('postal service') or *posto* ('position'), both ultimately participial forms of *porre* ('put'), others from the Latin preposition *post* ('after').



Establishing the relationships among these homonyms is also complicated by the fact that originally unrelated words of a similar form have influenced each other. This is also visible in (3), where the rounded vowel in *lügen* is a development that does not follow from a general sound law, but probably resulted from the need to distinguish *lügen* from *liegen* ('lie, be in a flat position, be located').

In sum, stating that two occurrences of forms are historically linked to the same word is an abstraction. A word is no empirical entity that exists over the course of time. Grouping historical occurrences together as a single word is thus a theoretical decision.

## 2.2 The boundaries of *signifiant* and *signifié*

Apart from the diachronic dimension discussed in section 2.1, words are also synchronically abstractions. In fact, a second dimension of abstraction concerns the synchronic extent of the unit referred to in an etymology. Saussure (1916) takes the opposition between *signes* as their determining factor. The *valeur* ('value') of a sign is that it is different from other signs. Both for the form and for the meaning of a Saussurean sign, the question arises how the boundaries are determined. In terms of the *signifiant*, the question is to what extent forms from the same historical stage of a language are combined as belonging to the same word. In lexicography, it is common to assume that headwords are what Matthews (1974: 22) calls *lexemes*. A lexeme includes all word forms of an inflectional paradigm. This raises the question as to which criteria are used to determine inflectional paradigms, i.e. how to distinguish inflection from word formation. As explained in ten Hacken (2014), this is not an empirical question and it can only be resolved on the basis of authority. In recording the etymology of a word, it is not necessary to have as precise an answer in each instance as required in theories of the lexicon or morphology. Ultimately, etymologies apply to individual word forms, but they are fairly trivial if the word forms are regularly formed. Whether *abridged* is taken as a verb form or an adjective does not affect its origin beyond its relation to the verb *abridge*. Conversely, in the case of suppletion, forms classified as inflected word forms should also be explained separately. The etymology of *bad* does not cover *worse*.

A more challenging issue is determining the extent of the *signifié*. Here, the question is which meanings are sufficiently distinct to require a separate unit of description. This is a form of the well-known question of the distinction between polysemy and homonymy. As the representation in one entry or as two entries is a decision that lexicographers have to make on a regular basis, the issue is discussed extensively in the context of dictionary making, e.g. Atkins and Rundell (2008: 265-316) and Koskela (2016). In the case of etymology, the problem is complicated by diachronic variation. Durkin (2016) discusses both cases of lexical merger and cases of lexical split. In the former case, a single word has two underlying forms with different etymologies. Durkin (2016: 246) mentions the transitive and intransitive readings of the verb *melt*, which have different Old English correlates. In the latter case, forms in free variation are assigned contrastive meanings. Durkin (2016: 248-251) discusses the case of *metal* and *mettle*, which until the 18<sup>th</sup> century were spelling variants of the same word.

In sum, both the form and the meaning of a word constitute abstractions. As there is no particular need to group word forms into lexemes for etymology, the synchronic grouping of forms does not pose big problems. However, the distinction between polysemy and homonymy, which is recognized as a lexicographic problem without a clear empirical solution, directly influences the units for which etymologies are given.

## 2.3 Named languages

Finally, the third dimension of abstraction relates to the system the word is part of. Independently of diachronic variation and the question of polysemy and homonymy, the status of a word as a

component of a language raises questions to which there are no straightforward answers. The significance of this dimension can be illustrated with the difficulty of answering the question as to whether a particular form is a word of English. As an example, let us consider the status of *hypernym* as opposed to *hyperonym*. Many non-specialists will react to the question of whether *hypernym* is a word of English by consulting a dictionary. In fact, the OED (2018) has an entry for *hypernym*. Lexicographers realize, of course, that the inclusion in a dictionary is the result of a decision by a lexicographer., and many will claim that this decision is taken on the basis of the occurrence in a corpus. As I showed in ten Hacken (2012), however, while a corpus can be used to justify a decision, it cannot provide an empirical basis that replaces a decision based on other (i.e. theoretical) considerations. Thus, COCA (2018) gives four occurrences of *inforamtion*, but nobody will claim that this means it is a new word of English. Obviously, they are four errors, where *information* was meant. In the case of *hypernym*, the question is whether it is an error for *hyperonym*. The fact that neither occurs in COCA cannot be used to take a decision. The only proper basis for a decision is a speaker's linguistic competence. However, linguistic competence is organized so as to support the use of language in communication, not its systematic description. Moreover, the existence of a particular word in one person's competence does not predict the existence of the same word in someone else's competence, even if they both speak English. Competence is inherently individual, and nobody's linguistic competence can be equated with *English*.

This does not mean that lexicographers are bound to take arbitrary decisions, but they have to take responsibility for their decisions. It is not possible to hide behind a corpus, but at least in the case of *hyper(o)nym*, it is possible to support the decision by appealing to considerations of analogy and etymology. The opposite of *hyper(o)nym* is *hyponym*. We also have pairs such as *hypotension* and *hypertension*. This seems to favor *hypernym*. On closer inspection, however, *-nym* is not a good parallel to *tension*. We find words such as *synonym* and *antonym*, where the first components are clearly not *\*syno-* and *\*anto-*, but *syn-* and *anti-*. Therefore, we get a better analysis if we assume that the second element is *-onym* in all of these words, which supports *hyperonym* as the correct form. This conclusion is not surprising, because the underlying Greek word is ὄνομα ('name'). Therefore, *hyperonym* is also the etymologically correct form. In this example, etymology supports a theoretically informed decision as to what constitutes a correct word of English.

## 2.4 Conclusion

In sum, we can say that *word* is an abstraction on three dimensions. When we consider the historical development, the question of whether two occurrences at different times are to be considered occurrences of the same word cannot be fully answered by empirical means. Similarly, the boundaries of the variation in form and the extent of the meaning cannot be determined in a fully empirical way. Finally, whether something is a word of a particular named language cannot be decided without an appeal to authority.

## 3 Sources of etymological information

Before turning to the interpretation of etymologies, it is useful to consider the sources on which they are based. In a way not unlike the question of how we can determine whether *hypernym* is a word of English, we can appeal to three types of information for the compilation of an etymology. First, etymologists have their intuitions a speakers, supplemented by their experience and knowledge built up in the course of their work. Secondly, corpora can be used. Thirdly, analyses by others, published in dictionaries or scholarly articles, can be appealed to.

In scientific work, using one's own intuitions as a source of information is all but inevitable, but appealing to them explicitly is not generally accepted. As a result, while intuition is used to arrive at a conclusion, the presentation of the results follows a different logic. This distinction between the 'logic of discovery' and the logic of presentation is one of the central insights of Popper (1959), which has since been accepted in the mainstream of the philosophy of science. In the case of language, the situation is somewhat different, because linguistic competence, which underlies any other realization of language, exists in the speaker's mind. On the basis of this insight, Chomsky (1957: 15) introduced grammaticality judgements as a source of data for linguistics. Ten Hacken (2007: 54-57) gives an overview of the issues this raises, and places this type of data in a broader context. In the case of etymology, grammaticality judgements are of little use. Decisions have to be made on what to investigate and how to interpret data collected from other sources, but etymology is not part of linguistic competence in the same way as rules of syntax are. The use of intuition by an etymologist is more like its use by an astronomer who builds instruments to make observations that cannot be made without them, but has to know which instruments to build, how to use them and how to interpret the observations.

Corpora are an important category of instruments used by etymologists, and are an indispensable source of historical data. However, one of the problems of corpora is that they cannot be representative of a language. This is a consequence of the fact that a language is not an empirical entity. The nearest empirical entity is a speaker's competence, but in a large corpus the performance of many different speakers is mixed. Moreover, the performance recorded in a corpus is not a direct, proportional reflection of the underlying competence. What speakers say or write is determined by various other factors interacting with their linguistic competence, e.g. the situation they are in and the aims they have. The realization of the intended performance may also be hampered by interfering factors, leading to what the speaker will identify as errors. Therefore, when an expression appears in a corpus, it may be a performance error, and when an expression does not appear in a corpus, there may be many different reasons for this, unrelated to its status as a correct expression for a particular speaker. Moreover, the etymologist working with historical corpora cannot ask the speakers of the historical period to check this.

A dictionary is an important source of information in the search for an answer to the question of whether a particular word exists in English, because it records the results of the analysis of corpora and linguistic intuitions by trained lexicographers. For etymologists, the sources of this type are much more varied. Traditionally, scholarly articles have been devoted to individual etymologies, e.g. Spitzer (1950). Although the relative weight of etymology as a field of study within linguistics has declined, we still find volumes dedicated to the presentation of individual etymologies, e.g. Hansen et al. (eds.) (2017). In etymological dictionaries such as Kluge (1995), many references to such sources are given to support the proposed etymologies. In addition, historical knowledge about language contact, material culture, and the exchange of ideas is an important source of indirect evidence to support or contest a hypothesis. As such, the work of an etymologist can be compared to the work of a zoologist studying the evolutionary development of species on the basis of fossils. Here the recorded data take the role of the fossils to be interpreted.

## 4 The mechanisms of etymological explanation

In his systematic overview of etymological mechanisms, Durkin (2009) devotes separate chapters to word formation, borrowing, change in word form and semantic change. In sections 4.1 to 4.4, I will discuss an example of each of these from English and German. For English, I will use the OED (2018) and for German Kluge (1995). In section 4.5, I will address the question of how these mechanisms relate to a speaker's competence and to the speech community.

#### 4.1 Word formation

As an example of word formation in an etymology, let us consider *fiver* in (4).

- (4) *fiver*  
 < *five* adj. and n. + *-er* suffix<sup>1</sup>

The OED (2018) is very systematic in giving word formation etymologies, even if the origin of the word is relatively straightforward. In the electronic version, the references to *five* and *-er* are hyperlinked to the relevant entries. This is particularly important for suffixes, as they are often highly ambiguous. In fact, the OED (2018) gives six homonyms of *-er*. (4) illustrates how an etymology is incomplete. It is not explained why a *fiver* is a banknote, because this cannot be deduced from the component parts, and etymologies generally do not cover such meaning components.

In German dictionaries, word formation etymologies are much less systematically covered. Thus, the compound *Elternabend* ('parents' evening') needs to be in a dictionary because of its specialized meaning, but the etymology is not recorded in any of the German dictionaries I consulted, because the component parts *Eltern* ('parents') and *Abend* ('evening') are immediately recognizable. In Kluge (1995), such words are not included.

As I argued in ten Hacken (2013a), the under-specification of the meaning of word formation outputs is a core property of word formation rules. On hearing a word formation output, speakers recognize that it is a new word and use the rule and the components to narrow down the range of possible meanings, but they also look in the context of use for a candidate concept that may be named by it. This process is recorded in etymologies only as a reference to a word formation rule, as in (4), or not at all, as for German *Elternabend* and other compounds.

#### 4.2 Borrowing

As an example of borrowing, let us consider the German *Fiasko* and English *fiasco*. Kluge (1995) gives the etymology in (5) (non-matching brackets corrected).

- (5) *Fiasko*  
 (< 19. Jh.). Entlehnt aus it. (*far*) *fiasco* 'durchfallen', eigentlich 'Flasche (machen)', zu it. *fiasco* m. 'Flasche' aus spl. *flasco* 'Weinkrug', aus wg. \**flaska* 'Flasche'.

The meaning of 'fiasco' developed from an Italian verbal expression with *far(e)* ('make'). On its own, the central meaning of Italian *fiasco* is 'bottle'. Going beyond modern Italian, (5) gives corresponding forms in Late Latin ("spl.") and West-Germanic ("wg."), the latter reconstructed. These forms have no relation to the German and English meaning. The reconstructed West-Germanic form underlies the normal words for 'bottle' in German, Dutch and Frisian. The Late Latin form means 'wine jug', which ties in with the prototype of an Italian *fiasco*, which is a wine bottle with a round body, the bottom part covered by a straw basket, traditionally common in the Chianti area. The OED (2018) gives the etymology in (6).

- (6) *fiasco*  
 < (in sense 2 through French) Italian *fiasco* (see FLASK n.<sup>2</sup>) lit. 'a flask, bottle'.

As a historical dictionary, the OED starts the description of *fiasco* with the sense of the Chianti bottle and, for the more common sense of the word, records the intermediate stage of French. The further etymology of *fiasco* is discussed in the hyperlinked entry of *flask*. Both Kluge's (1995) (5) and OED's (6) are followed by remarks about the relationship between the 'bottle' reading and the 'fiasco' reading in Italian. An important property of borrowing illustrated in (5) and (6) is that the central meaning



of the Italian noun does not play a role in the meaning of the German and English borrowing. Zingarelli (1988) gives the ‘fiasco’ reading as the fourth, after three readings based on ‘bottle’. The OED entry dates from 1895. It also records the ‘bottle’ reading with a single quotation. It is not obvious to what extent this reading has been added for etymological reasons. The BNC has 232 occurrences of *fiasco*, none of which is of the ‘bottle’ reading.

### 4.3 Change of form

Let us now turn to the mechanism of the change of form. We have seen some examples of changes in word form in (3) and (5). This type of information is more important for words where no word formation rule or borrowing can be used as a starting point of the etymology. An example is *Buch*, for which Kluge (1995) gives the etymology in (7).

(7) *Buch*

(< 8. Jh.). Mhd. *buoch*, ahd. *buoh* f./n./m., as. *bōk* (s. u.) aus g. *\*bōk-(ō) f.*, auch in [...]

The etymology in (7) consists of two parts, of which only the first is quoted. This first part gives a linear historical development in reverse chronological order. The language stages are the same as in (3). The *ahd* (‘Old High German’) variety is contemporary with the variety indicated by *as* (‘Old Saxon’), with the former in the southern part of the German-speaking area, the latter in the northern one. This historical development is followed by a list of cognates introduced by ‘auch in’ (‘also in’). The OED’s (2018) etymology of *book* in (8) uses a different strategy.

(8) *book*

Cognate with Old Frisian *bōk* book, Old Dutch *buok* large written document, book (Middle Dutch *boec* book, document, Dutch *boek*), Old Saxon *bōk* book, writing tablet (Middle Low German *bōk*, *būk*), Old High German *buoh* book, written text, scripture, (in an isolated attestation) letter of the alphabet (Middle High German *buoch*, German *Buch*), Old Icelandic *bók* book, story, history, Old Swedish *bok* book (Swedish *bok*), Old Danish *bok* book (Danish *bog*), and also (in a different declension: feminine *ō* -stem) Gothic *bōka* letter (of the alphabet), in plural *bōkōs* also in the sense ‘(legal) document, book’ (perhaps also in singular in this sense, as indicated by the compound *frabauhtabōka* document of sale), probably < the same Germanic base as BEECH *n*.

Instead of outlining a linear history at the outset, (8) only gives a long list of cognates in related languages. The order of the languages is roughly from more closely related to English to somewhat further removed. For each language the oldest recorded form is given, along with the later forms developing from it in brackets. There is a lot of overlap of (8) with the information in Kluge’s (1995) etymology after *auch in* in (7). The recorded forms in (7) are also included in (8). The difference is that (7) gives a stronger sense of linear development and includes a reconstructed form.

### 4.4 Semantic change

The final etymological mechanism is semantic change. We have come across an example in which semantic change played a role in (5) and (6). The nature of the connection between the ‘bottle’ reading and the ‘fiasco’ reading of the Italian *fiasco* is a matter of debate. However, in this case it is in a sense obvious that the two are readings of the same word, because speakers know both senses of the word. In (8), we also encountered some examples of semantic variation, but in this case the senses are much closer to each other. A general problem with semantic change as an etymological mechanism is that the meaning of a word is not recorded in a corpus, but only arises through interpretation of the form. In general, meaning only exists in speakers and hearers, not in texts and utterances, and this is



the reason people can misunderstand each other. The problems this causes for an etymologist can be illustrated by the verb *spear*. Suppose we have an example like (9) in our corpus.

(9) The poet tells us how the King saw his men speared.

In (9) we have an edited example from the OED (2018: *spear*<sub>v</sub>). The meaning of *spear* can be inferred on the basis of the current meaning of the word and the context of (9), perhaps supplemented with knowledge of the use in other contexts. However, it is not possible to decide with absolute certainty whether *spear* in (9) means ‘attack with a spear’, ‘wound with a spear’, or ‘kill with a spear’. Distinguishing these meanings would ultimately require asking the author what he or she meant. For historical corpora, this is generally not possible. This is one reason why semantic considerations are usually subordinate to formal similarity in etymologies. The meaning of words is invoked mainly in order to show that we are dealing with the history of the same word. Sense extensions such as that of the Italian *fiasco* in (5) are essential to trace the history of German *Fiasko* beyond the borrowing from Italian. Meaning changes can also be of the type exemplified by the sense extensions of English *ride* and *drive* or German *reiten* and *fahren* to new types of vehicle. As such, *bicycle* collocates with *ride* in English, which is also used for horses, whereas in German it is not *reiten*, which is used for horses, but *fahren* (‘drive’) that is used with *Fahrrad* (‘bicycle’). The German *fahren* is used with vehicles in general, and also with boats and ships. In Polish, the same verb, *jeździć* (‘go’), collocates with *koń* (‘horse’), *rower* (‘bicycle’) and *samochód* (‘car’). Such developments are the result of how new means of transport are included in existing classification schemes incorporated in the logic of collocation. Though potentially highly interesting from a cross-cultural perspective, they are generally not covered in etymologies.

#### 4.5 Speakers and speech communities

Having considered the four main mechanisms used in etymologies, let us now turn to their implementation. Given that the underlying empirical reality of language is a speaker’s linguistic competence, it must be determined how each of the mechanisms relates to this competence. Three of the mechanisms, word formation, borrowing and sense extension, correspond to the key mechanisms for naming new concepts. Elaborating on this onomasiological perspective, Štekauer (1998: 5) takes the speech community rather than the individual speaker’s competence as the point of departure, but ten Hacken and Panocová (2011) show that there is no genuine opposition involved, because naming actions perceived as actions of a speech community are actually performed by individual speakers. A competence-based view of language is not incompatible with the recognition of speech communities. As long as we accept that labels such as “Mhd.” in (7) and “Old Frisian” in (8) do not refer to empirical entities with clear boundaries but are fuzzy concepts whose boundaries cannot be determined exactly, there is no problem with their use in etymologies.

The three mechanisms for naming new concepts show different degrees of regularity, but in all cases the meaning of their output is determined by what I called *onomasiological coercion* (ten Hacken 2013b). This means that the meaning is the concept to be named rather than what is predicted by the mechanism used. Word formation is a rule-based mechanism. The rules are implemented in individual speakers. They are used both in coining new words and in understanding them. Here *new* means ‘new to the speaker’. Sense extension is very similar, and includes, for instance, metaphor and metonymy. These seems to be less rule-based than word formation, but if we take into account that the actual meaning of a word formation output is determined in large part by the concept to be named, the difference with metaphor and metonymy is less striking. Borrowing is also driven by the need to name concepts, which explains why it is not an entire word with all of its senses that is borrowed, but only a name attached to a concept, as in the case of *fiasco*. In each of the three mechanisms, naming is performed on a case-by-case basis.

Sound change is of a very different nature. As explained by Beekes (1995), it underlies much of the historical-comparative work in linguistics originating in the 19<sup>th</sup> century. Unlike word formation rules, rules of sound change are not realized in the speaker's competence. A rule of sound change is a generalization about historical developments observed by linguists. It is thus worth considering what exactly constitutes the empirical basis underlying such generalizations. Clearly, it cannot be the historical change of a word, because a word is not an empirical entity, as we saw in section 2. In historical linguistics, words are studied primarily as realizations in performance. However, generative linguists have also linked these data to the underlying competence, e.g. Lightfoot (1999). From a historical perspective, performance and competence are in a cyclic relationship. An individual speaker's competence underlies their performance. At the same time, the origin of the competence in an individual speaker is the language acquisition process, which is based on observation of the performance of speakers in the environment. What happens in a sound change is that performance is observed to change gradually. For individual speakers, this does not have to mean that their competence changes. The sound image of a word is a prototype used to interpret perceived speech and produce the word in performance. Without changing the prototype, a certain shift in perceived speech can be accommodated. For a new generation, the prototype will be calibrated anew on the basis of the performance of the community, which has gradually shifted. In this way, changes may take place gradually. A crucial difference to the naming mechanisms is that at no point does a decision by a speaker have to be taken.

## 5 Conclusion

In ten Hacken (2009), I showed that a dictionary cannot be a description of the vocabulary of a language. The core of the argument is similar to the point made in section 2 that a word is an abstraction. For general dictionaries, I proposed that they should be interpreted as tools for users to solve problems with. Such a conclusion is broadly in line with lexicographic approaches, such as that outlined by Bergenholtz and Bergenholtz (2011), although they arrive at it from a very different background.

If a dictionary is a tool, a central question in the interpretation of etymologies is who needs etymologies. In fact, this is Svensén's (2009: 333) starting point in his chapter on etymology. Svensén takes etymology to be less important from a utilitarian perspective, but is motivated by an interest in "facts of language". This raises the question of what these facts are, given that words as elements of a language are not empirical entities.

Whereas etymology is traditionally oriented towards the history of words, the discussion of the methods and mechanisms above suggests a different interpretation. When Saussure (1916) proposed his theory of the word as a *signe*, he stipulated that the relationship between the *signifiant* and the *signifié* is arbitrary. If we consider the *signifiant* and the *signifié* as entities that can be related, this is no doubt correct. After all, different languages have different words for the same concept. However, the word is not an entity but an abstraction. This is the same for the *signifiant* and the *signifié*. They exist in speakers' minds, but not as speaker-independent entities. For the individual speakers, the relationship between the form and meaning of a word is *not* arbitrary. It is in most cases determined by the performance of other speakers in the environment at the time of language acquisition, itself based in the competence of these speakers. As long as the result is immediately recognizable as 'the same word', this process is not remarkable. What is described in etymology is the cases where there is a change in this transfer.

Therefore, etymology can be seen as the historical record of the motivation of the relationship between the form and meaning of a word. The standard assumption is that when a speaker adds a word to their competence, this is taken over from the competence of other speakers in the speech

community. This corresponds to the situation where the performance of other speakers is interpreted in accordance with how they intended it. A large proportion of this vocabulary extension takes place in childhood. Change of pronunciation may occur as a gradual side effect of the calibration of the prototype on the basis of performance. Change of meaning may occur for the same reason. These are the gradual processes described in etymology. Naming acts are the more striking etymological facts. Here, word formation, borrowing or sense extension are used to name a new concept. *New* in the context “new concept” is of course also speaker-dependent. The difference with regard to gradual phonological and semantic changes is that no model is available in the competence of speakers in the same speech community for a new concept that requires an act of naming.

## References

- Atkins, B.T. Sue & Rundell, Michael (2008), *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.
- Beekes, Robert S.P. (1995), *Comparative Indo-European Linguistics: An Introduction*, Amsterdam: Benjamins.
- Bergenholtz, Henning & Bergenholtz, Inger (2011), ‘A dictionary is a tool, a good dictionary is a monofunctional tool’, in Fuertes-Olivera, Pedro & Bergenholtz, Henning (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*, London: Continuum, pp. 187-207.
- Chomsky, Noam (1957), *Syntactic Structures*, Den Haag: Mouton.
- Considine, John (2013), ‘Researching Historical Lexicography and Etymology’, in Jackson, Howard (ed.), *The Bloomsbury Companion to Lexicography*, London: Bloomsbury, pp. 148-164.
- Durkin, Philip (2009), *The Oxford Guide to Etymology*, Oxford: Oxford University Press.
- Durkin, Philip (2016), ‘Etymology, Word History, and the Grouping and Division of Material in Historical Dictionaries’, in Durkin, Philip (ed.), *The Oxford Handbook of Lexicography*, Oxford: Oxford University Press, pp. 236-252.
- ten Hacken, Pius (2007), *Chomskyan Linguistics and its Competitors*, London: Equinox.
- ten Hacken, Pius (2009), ‘What is a Dictionary? A View from Chomskyan Linguistics’, *International Journal of Lexicography* 22: 399-421.
- ten Hacken, Pius & Panocová, Renáta (2011), ‘Individual and Social Aspects of Word Formation’, *Kwartalnik Neofilologiczny* 58: 283-300.
- ten Hacken, Pius (2013a), ‘Semiproductivity and the place of word formation in grammar’, in ten Hacken, Pius & Thomas, Claire (eds.), *The Semantics of Word Formation and Lexicalization*, Edinburgh: Edinburgh University Press, pp. 28-44.
- ten Hacken, Pius (2013b), ‘Compounds in English, in French, in Polish, and in General’, *SKASE Journal of Theoretical Linguistics* 10: 97-113.
- ten Hacken, Pius (2014), ‘Delineating Derivation and Inflection’, in Lieber, Rochelle & Štekauer, Pavol (eds.), *The Oxford Handbook of Derivational Morphology*, Oxford: Oxford University Press, pp. 10-25.
- Hansen, Bjarne Simmelkjær Sandgaard; Whitehead, Benedicte Nielsen; Olander, Thomas & Olsen, Birgit Anette (Eds.) (2017), *Etymology and the European Lexicon: Proceedings of the 14th Fachtagung der Indogermanischen Gesellschaft, 17-22 September 2012, Copenhagen*, Wiesbaden: Reichert Verlag.
- Kluge (1995), *Etymologisches Wörterbuch der deutschen Sprache*, 23th edition, edited by Elmar Seebold, Berlin: De Gruyter.
- Koskela, Anu (2016), ‘Identification of homonyms in different types of dictionaries’, in Durkin, Philip (ed.), *The Oxford Handbook of Lexicography*, Oxford: Oxford University Press, pp. 457-471.
- Lightfoot, David (1999), *The Development of Language: Acquisition, Change, and Evolution*, Oxford: Blackwell.
- Matthews, Peter H. (1974), *Morphology: An Introduction to the Theory of Word Structure*, Cambridge: Cambridge University Press.
- OED (2018), *Oxford English Dictionary*, Third edition, edited by John Simpson, www.oed.com.
- Popper, Karl R. (1959), *The logic of scientific discovery*, London: Routledge.
- Saussure, Ferdinand de (1916), *Cours de linguistique générale*, Charles Bally & Albert Sechehaye (eds.), Édition critique préparée par Tullio de Mauro, Paris: Payot, 1981.

- Spitzer, Leo (1950), 'On the etymology of *pet*', *Language* 26: 533-538.
- Štekauer, Pavol (1998), *An Onomasiological Theory of English Word-Formation*, Amsterdam: Benjamins.
- Svensén, Bo (2009), *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*, Cambridge: Cambridge University Press.
- Zingarelli, Nicola (1988), *Vocabulario della lingua italiana*, 11th edition, edited by Miro Dogliotti and Luigi Rosiello, Zanichelli, Bologna.





# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function

**Christina Mutter, Aleksander Wiatr**

*Ludwig-Maximilians-Universität München*

*E-mail: christina.mutter@lmu.de, aleksander.wiatr@lmu.de*

## Abstract

This paper describes the long-term research project VerbaAlpina of Munich University, which has been funded by the German Research Foundation (DFG) (<http://gepris.dfg.de/gepris/projekt/253900505>) since October 2014. The project investigates the Alpine lexis of three conceptual domains in the Alpine region where dialects and languages belonging to three large language families (Germanic, Romance and Slavonic) are spoken. This paper emphasizes one of the project's main functional areas, its lexicographic function, which serves to gather, process, access and visualize lexical data. To this end, data from traditional linguistic atlases and dictionaries as well as recent data gathered via the project's crowdsourcing tool first have to undergo a process of systematic data processing to fit the unified structure of the relational database (MySQL). This process can be subdivided into three major steps: transcription, tokenization and typification. Apart from the multi-directionality of the project which collects, documents and disseminates structured linguistic and ethnographic data, VerbaAlpina also provides an innovative online publishing platform that will prove sustainable and can be easily cited.

**Keywords:** digitalization; crowdsourcing; interlingual geolinguistics

## 1 Project Description

### 1.1 Area Under Investigation: The Alpine Region

The project “VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit” (VerbaAlpina. The Alpine cultural region reflected through its multilingualism) seeks to investigate the linguistic and cultural area of the entire Alpine region from a transnational perspective through selective analyses. The geographical region of investigation is limited to the territorial borders defined by the Alpine convention<sup>1</sup>. This covers a surface area of 190,600 km<sup>2</sup> and encompasses parts of six different countries (Austria, Italy, France, Switzerland, Germany and Slovenia) as well as two entire countries (Liechtenstein and Monaco). The Alpine region is characterized by its ethnographic and topographic homogeneity, and at the same time by its strong linguistic heterogeneity. This linguistic heterogeneity, which includes three large language families (Germanic, Romance and Slavonic), has made the region a topic of interest for linguists. Accordingly, VerbaAlpina (in the following text often abbreviated as VA) focuses on the following languages and their respective dialects: German, French, Italian, Slovenian, Franco-Provençal, Romansh, Ladin, Friulian and Occitan (c.f. Krefeld/Lücke 2014b: 189).

1 <http://www.alpconv.org/en/convention/default.aspx>, [last access: 23.03.2018]: “The Alpine Convention is an international treaty between the Alpine Countries (Austria, France, Germany, Italy, Liechtenstein, Monaco, Slovenia and Switzerland) as well as the EU, for the sustainable development and protection of the Alps.”

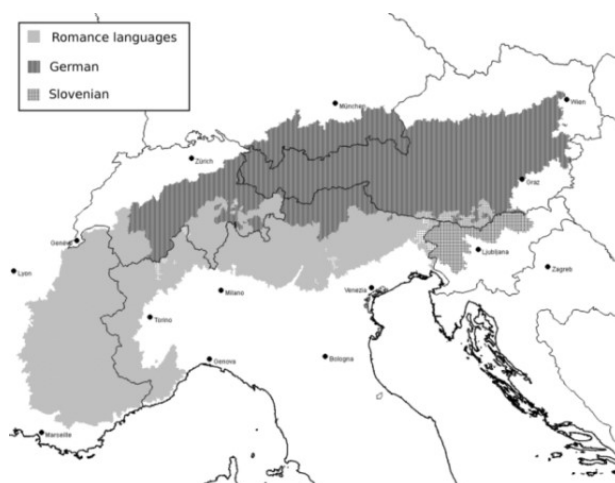


Figure 1: The area under investigation in VerbaAlpina corresponding to the Alpine region as defined by the Alpine Convention. The data collection of the three conceptional domains is broken down into three stages.

Stage one (from October 2014 to October 2017) focused on vocabulary related to Alpine pasture farming, in particular, milk processing. In the current phase (from November 2017 to November 2020) the project is concerned with the lexis of the domains fauna, flora, landscape formation and weather. The last stage (from December 2020 to December 2023) will focus on the vocabulary of modern Alpine life (ecology, tourism).

## 1.2 Data and Methodology<sup>2</sup>

VerbaAlpina gathers and analyzes linguistic data derived, on the one hand, from linguistic atlases and geo-referenced dictionaries from the past one hundred years. Figure 2 gives an overview of the large number of traditional atlases and lexica in which the relevant vocabulary of the Alpine region is recorded. Yet these resources lack a multilingual perspective, and only cover parts of the Alpine vocabulary. They were created at different times and document diverse concepts. This project's on-line crowdsourcing tool (c.f. Wiatr 2016; Krefeld/Lücke 2017b) thus helps to even out, complete and correct this inhomogeneous data stock ([www.lmu.de/verbaalpina](http://www.lmu.de/verbaalpina)).

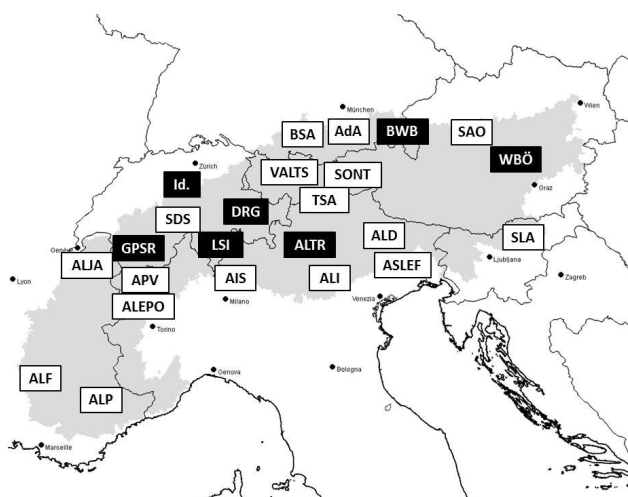


Figure 2: Overview of traditional atlases (white boxes) and dictionaries (black boxes) of the Alpine region (map created by VerbaAlpina)<sup>3</sup>.

<sup>2</sup> All relevant processes are documented in the menu item “methodology” on the project platform ([https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172)).

<sup>3</sup> See the reference section for the abbreviations of the different atlases and dictionaries.

VerbaAlpina thus combines and further develops three different approaches to digital geolinguistics: (1) atlases which are digitally published but whose data has been gathered through traditional methods (e.g., ALD); (2) atlases which document diverse languages and language families (e.g., WALS); and (3) web-based atlases which use the internet to create data sets (e.g., AdA). (c.f. Krefeld 2017f)

The project addresses the major challenge posed by the lack of uniformity of the data from the individual data sources (linguistic atlases, dictionaries, crowdsourcing) which are not structured in the same way. VerbaAlpina must first unify the different transcription systems of the atlases and dictionaries. For this purpose, the already existing data in both digital and analogue form undergo a process of systematic data processing to fit the unified structure of the relational database (MySQL) in which all project data is stored (for more details c.f. Oberholzer/Kunzmann in press). This process can be subdivided into three major steps: transcription, tokenization and typification. These sub-processes will be described in more detail in section 2.

### 1.3 Research Aims

VerbaAlpina seeks, first of all, to investigate the Alpine region which is characterized by numerous languages (cf. section 1.1.) and their corresponding dialects in its historico-cultural and historical linguistic unity. As such, VerbaAlpina also overcomes the traditional limitation of geolinguistic investigation to nation-states.

Of primary importance is the recognition of connections regarding the etymology of the individual dialectal words. In this way, differences and similarities between the individual language groups of the Alpine region can be found. Many words share a common etymology, even if this cannot be seen anymore at first sight (c.f. Krefeld 2017c). For example, the German word *Butter*, French *beurre* and Italian *burro* all go back to the Greco-Latin word *butyru(m)*. Two additional examples are: the Swiss-German *Staffel*, German *Stadel*, French *étable* and Italian *stabbio* are all based on the Latin word *stabulum*; the Swiss-German *Schotte(n)*, Italian *scotta* and Slovenian *skuta* all derive from Latin *\*excocta*.

The cooperation with other projects is fundamental for VerbaAlpina, as reflected by numerous cooperation agreements with international partners from the entire Alpine region. Each cooperation is based on a formal agreement which guarantees the project partners their own database to upload their data. Each project partner's data is then available to all partners in a structured form. The cooperation is not limited to data exchange, as all partners are also invited to use and further develop all the functional areas offered by VerbaAlpina (c.f. Krefeld 2017d).

## 2 Lexicographic Function

One of the core functions of the research environment of VerbaAlpina is its lexicographic function. This enables researchers to explore the data from two points of view: onomasiological and semasiological. The data from investigations from either perspective can be accessed through the interactive map on the project platform or via direct database query (see the detailed description in section 2.4.). For this purpose, the raw source data must be implemented and processed to fit the structure of the relational database. Since the sources of the input data are both analogue and digital, VerbaAlpina must adapt and handle them individually. Regardless of the type of source data there are three steps that always take place to process linguistic material: (1) transcription/transfer, (2) tokenization and (3) typification. This section provides an overview of these steps. Moreover, we will also discuss the challenges we encounter and how we deal with them in order to standardize the data to make it comparable (c.f. Lücke 2017c).

## 2.1 Transcription of the Input Data

### 2.1.1 Analogue Data

The main sources of our data stock are diverse language atlases from the Alpine region. The project's two biggest challenges regarding data processing are:

- the conception of the language atlases; and
- the transcription system used.

Most linguistic atlases differ in conception. This often depends on the type and objective of each atlas. In most cases one single map shows linguistic attestations for a single concept. It is also common for other kinds of information to be displayed: attestations, types, other concepts, and explorer's remarks. Furthermore, the linguistic material on the maps is not always presented in the same way. Whereas in the Romance (e.g. AIS, ALF) and Slavic (e.g. SLA) tradition the atlases show the actual expression that stands for a given concept in a given place, the Germanic atlases (e.g. VALTS) provide maps that show types (phonetical or morphological ones) as well as mixed forms, namely forms with types and expressions. Moreover, there are also atlases which show types referring to a certain area (e.g. ALP) (see Figure 3 below).

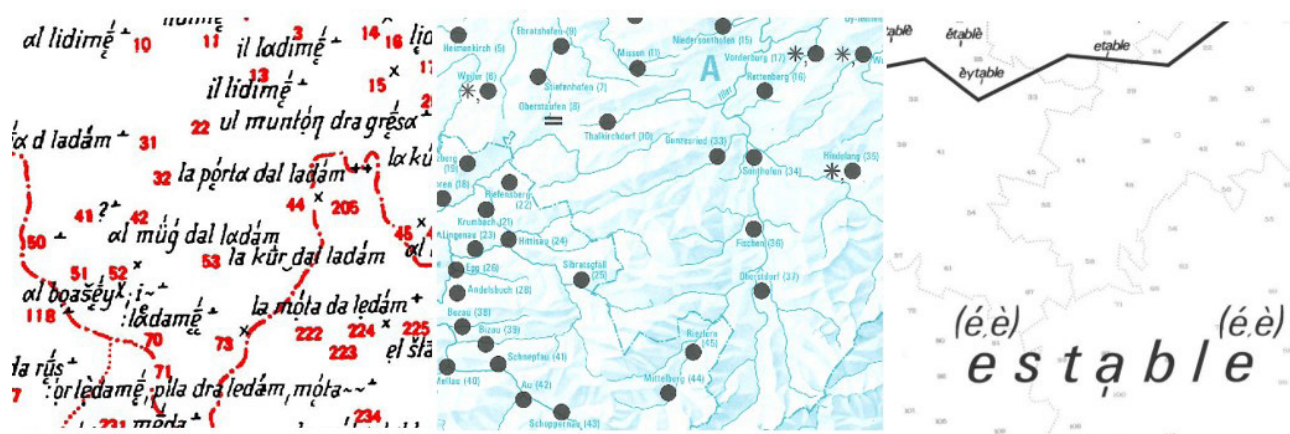


Figure 3: The picture illustrates the differences between language atlases regarding the presentation of linguistic material. On the left: AIS with linguistic attestations for each recording location; in the middle: VALTS with phonetical types; on the right: ALP showing a mixed form of attestations and phonetical types referring to a certain area.

Second, the transcription systems vary strongly depending on the individual scientific tradition. For example, AIS uses the transcription system Böhmer-Ascoli; ALF and ALJA use the Gilliéron-Edmont system; VALTS uses Teuthonista. When looking at the data in the context of a specific atlas and a specific region, the transcription problem may not be obvious at first. However, the use of different transcription systems turns out to be quite problematic when investigating language phenomena that cross national and linguistic borders. Since VerbaAlpina aims to examine the Alpine region from a transnational perspective, we have developed methods and tools to extract, handle and save all information (linguistic and non-linguistic one) to make the data structured and comparable. VerbaAlpina's transcription tool is essential when transferring analogue data from traditional atlases into digital data (c.f. Krefeld 2017g).

The main window of the tool shows the transcription object, in this case a single map. The attestations that have not yet been transcribed are suggested automatically. At this point, the transcriber must decide whether the individual attestation is defined as a real expression or as a phonetical or a morphological type. Before entering the transcription into the database, the corresponding concept

must be chosen. This is particularly helpful when dealing with maps containing multiple concepts, such as the maps of the AIS. We use Beta Code to transcribe the data from the different atlases in order to preserve all phonetic information and have access to the original transcription at any time. Beta Code is a system of unequivocal signs that are independent of any computer system and font type, and which can be used for the transcription of diverse phonetic systems. Since the transcription takes place on a graphematic and not a phonetic level, the Beta Code can be applied to every kind of atlas because it only symbolizes graphemes and not the phonetic values. Thus, only one code is necessary for the transcription of all data, allowing the transcription work to be more efficient and delegated to non-specialized staff. The Beta Code differentiates between basic signs (normal letters) and diacritics. First, the transcriber writes down the basic sign followed by diacritics from left to right and from bottom to top (see Figure 4) and then moves on to the next basic sign (c.f. Krefeld/Lücke 2017a).

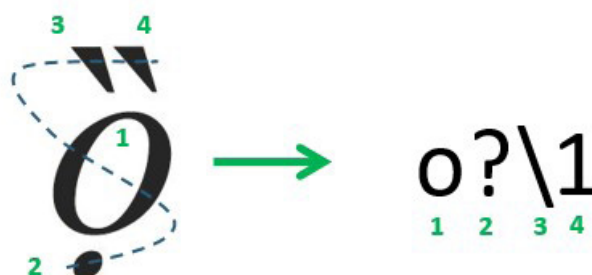



Figure 4: Use of Beta Code in the transcription tool of VerbaAlpina.

Unique combinations of Beta Code signs and the corresponding IPA symbols have been listed in a specific code page for every atlas. The transcribed linguistic material is converted into IPA without losing the original transcription at any stage of the process. Table 1 illustrates how the original transcription is transcribed with the help of the Beta Code and its conversion into IPA.

Table 1: Using Beta Code to transcribe different transcription systems.

Source attestation	Beta Code transcription		
	una1 mu:g/a1 da1 va(c)/		
IPA: [una mydʒa da v'atɕ]			
Source	Beta Code	IPA	
AIS	a1	ɑ	&#x0251
AIS	c)/	tɕ	&#x02a8
AIS	d	D	&#x0064
AIS	g/	ɟʒ	&#x02a4
AIS	m	M	&#x006D
AIS	n	N	&#x006E
AIS	u	U	&#x007
AIS	u:	y	&#x0079
AIS	v	V	&#x0076



### 2.1.2 Digital Data

Digital data sources present a different set of challenges compared with analogue data. To cope with these, we first create unique databases for every cooperation partner of VerbaAlpina in which the unaltered source data is stored. Although in most cases the data is provided as a database dump or Excel-data, we must adapt it to the structure of our relational database. All steps regarding the data processing of every specific source are documented in detail for consultation at a later stage.

The second and most challenging part of the implementation of digital data consists in the coding of signs or, more precisely, the font types used by the individual partner projects. In many cases the coding does not respect the standards of Unicode, which leads to problems when the data is transferred into a different system. The non-standard coding of signs is unproblematic as long as the coding is restricted to the project itself. It becomes problematic when two or more projects using different codings are brought together. For that reason, we replace the incorrect coding and add the corresponding IPA to the code page (c.f. Lücke 2017a).

## 2.2 Tokenization

All transcribed and transferred data is first stored in the table *aeusserungen* (utterances) in our database. Single words as well as entire utterances can be found there. Every single record is related to a geographic point of reference and to the corresponding concept and stimulus. During the process of tokenization we subdivide utterances into their single components and we separate the lexical information from the grammatical one, such as the article. At this point all data is uploaded into the table ‘tokens’. In case of single word attestations, the concept is inherited from the table *aeusserungen*. Multi-word lexical units additionally have to be split up into their single components (tokens) and then assigned to new concepts. This step is quite challenging, since it is not always obvious which concept is the correct one. However, it is always possible to reconstruct the data before the transformation. The conversion into IPA also takes place during this step.

Table 2: Illustration of the multi-word lexical unit and its components and the corresponding concepts.

Attestation in Beta Code	Attestation in IPA	Concept
una1 mu:g/a1 da1 va/c)/	una mydʒa da v <sup>l</sup> atɕ	HERD OF COWS
TOKENIZATION		
una1		ARTICLE
mu:g/a1	mydʒa	HERD
da1	da	PREPOSITION
va/c)/	v <sup>l</sup> atɕ	COW

## 2.3 Typification

The third process of the lexicographic function is the typification. In this all single tokens as well as whole utterances are assigned to a so-called morpho-lexical type. A morpho-lexical type is defined by the following categories: language family, part of speech (PoS), affix and gender. For example, for the tokens *barga*, *barg*, *margun*, and *bargun* we would have three different morpho-lexical types, as shown in Table 3 (c.f. Krefeld/Lücke 2017e).

Table 3: Creating morpho-lexical types for existing tokens.

	<i>barga</i>	<i>barg</i>	<i>margun</i>	<i>bargun</i>
<b>language family</b>	roa	roa	roa	roa
<b>PoS</b>	noun	noun	noun	noun
<b>affix</b>	-	-	+	+
<b>gender</b>	f	m	m	M
<b>morpho-lexical type</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>3</b>

With regard to orthography, we first search for an equivalent in a standard language and use it as an orthographic form. If standard language dictionaries do not provide an appropriate form, we search for one in regional and dialectal dictionaries. For example, for a series of single attestations one morpho-lexical type is created:

[l'atʃ], [l'atɕ], [l'etʃ], [l'etɕ], [l'atɕ], [l'atɕ], [l'atɕ], [l'at], [l'etʃ], [l'etʃ], [l'etʃ], [l'atɕ], [l'æjt], β 'lait/latte rom. m.'

Since we only work with information we have, it is also possible that morpho-lexical types without gender marking are created:

- (1) Butter (ger. m.)
- (2) Butter (ger. f.)
- (3) Butter (ger. n.)
- (4) Butter (ger.)

In the next step of the typification process the morpho-lexical types are assigned to a so-called basic type, which is only defined by the source language. The basic type is not to be confused with the etymon of a word, which can but is not necessarily be the same word as the basic type. As in the case of morpho-lexical types, we use dictionaries as a reference (c.f. Krefeld/Lücke 2017d).

## 2.4 Accessing the Processed Data in Two Ways: Via an Interactive Map and the Database Interface

Having completed the process of typification the data is made accessible through two different user interfaces. The first is the interactive map<sup>4</sup> which allows the user to choose elements from linguistic core data as well as data from the linguistic periphery, also called extra-linguistic data, such as toponyms, demographical and other information. For exploring the linguistic data, one can choose from concepts (semasiological point of view), morpho-lexical types and basic types (onomasiological point of view). In addition, different kinds of data can be easily grouped on a single map. The interactive legend which is updated dynamically according to the presentation of the data gives the user the possibility to switch on and off preselected data. (c.f. Krefeld 2017a) The interactive map offers two different views: the physical and the hexagonal one. The latter, due to its algebraic simplification of the geographic dimension to hexagons, is very useful when using the quantifying feature of the interactive map, enabling the user to quantify existing data in relation to a chosen territorial unit (language areas, NUTS-3) of the Alpine region (see Figure 5 below) (c.f. Lücke 2017e). Registered users have the possibility to save their selected data in a so-called synoptic map and to reuse or modify it later (c.f. Krefeld 2017e).

<sup>4</sup> To understand how the interactive map works and what possibilities it offers, please visit our website and experiment with it: [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=133&db=172](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133&db=172) [31.03.2018].

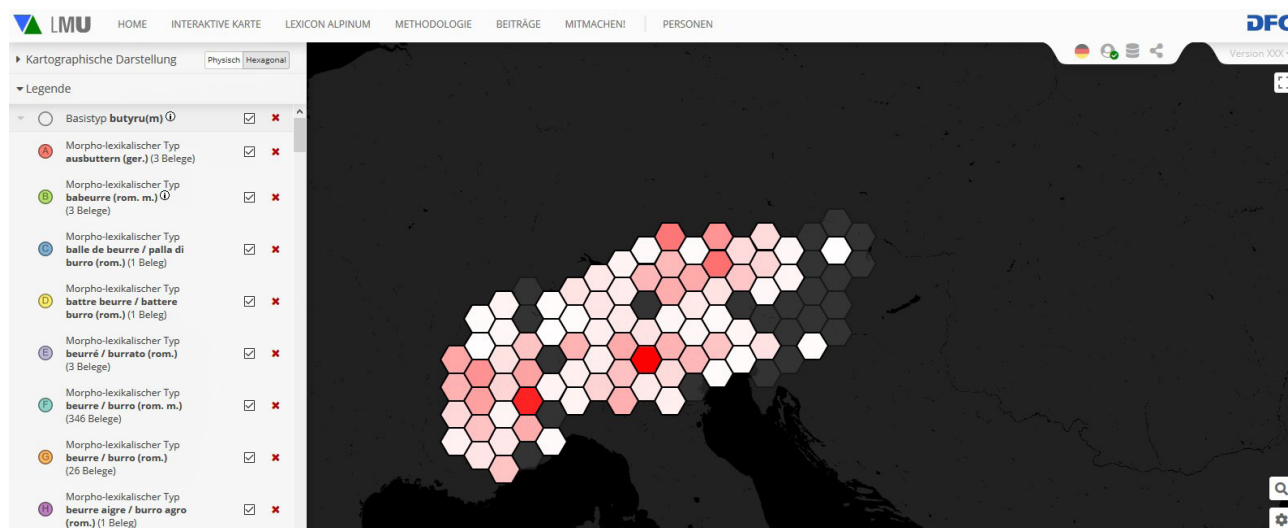


Figure 5: Demonstration of the quantification tool of VerbaAlpina in use (in the hexagonal mode). This screenshot illustrates the frequency of different morpho-lexical types related to the basic type *butyru(m)*. The redder the polygon the more tokens are present in this area. The legend on the left side of the screenshot shows the morpho-lexical types that matched with the basic type *butyrum*.

The other and more precise way to access the data is via the database interface. Whereas the interactive map is open access for everybody, access to the database of VerbaAlpina is currently granted to project partners only. It is, however, possible that other users are also granted access as soon as they register and are accepted by the VerbaAlpina team<sup>5</sup>. The database interface<sup>6</sup>, which is a MySQL view, presents the user with the data in a structured and comparable form. Provided they have some MySQL knowledge, the user can select, filter, group and organize the data in an unlimited number of ways. For example, one database query could ask for all tokens from AIS, ASLEF and SLA that have the basic type *butyrum* in common, that are all feminine and have no suffixes. It is necessary for users to have some knowledge of MySQL, in particular the structure of the relational databases in MySQL and of the query language. However, VerbaAlpina offers some options to download parts or all of the data and edit it with the help of spreadsheet software such as Microsoft Excel. Accessing data through our interfaces gives the user the possibility to work with recent data. Since the team of VerbaAlpina works continuously on the data, the access interface is updated every day (c.f. Lücke 2017b).

#### 2.4.1 Data from the Linguistic Periphery

VerbaAlpina does not only use and process linguistic data. In fact, every kind of data which is geo-referenced can be inserted into the structure of VerbaAlpina. The additional types of data help with the interpretation of the created maps. Thus, some linguistic phenomena can be understood better when combined with other data, such as demographic or infrastructural data. Currently we dispose of the following extra-linguistic and geo-referenced information: data about settlements, infrastructure, castles, Latin inscriptions, early medieval sites, churches and toponymy (c.f. Krefeld 2017b).

<sup>5</sup> You can register using the following link: <https://www.verba-alpina.gwi.uni-muenchen.de/wp-signup.php>.

<sup>6</sup> The database interface is accessible in all working languages of VerbaAlpina, at present in German, French, Italian, Slovenian and Rhaeto-Romance. We are also working on an English version to make it accessible to a broader audience.

### 3 VerbaAlpina as an Innovative and Sustainable Online Publishing Platform

One of the major and most debated problems of digitalization consists in the sustainability of digitized data (for further information see for example: Maron, Smith, Loy 2009, Bradley 2007, Krefeld/Lücke 2017c). To cope with this issue VerbaAlpina has developed an innovative approach of versioning and of citability.

#### 3.1 Versioning

VerbaAlpina consists of the following modules: VA\_DB, VA\_WEB, VA\_MT (see Figure 6 below). The module VA\_DB comprises the data stock which is stored in the project's MySQL database (va\_xxx). VA\_WEB encompasses the program code of the web interface of the project platform [www.verba-alpina.gwi.uni-muenchen.de](http://www.verba-alpina.gwi.uni-muenchen.de) including the corresponding WordPress database (va\_wp). The module VA\_MT contains media files (photos, videos, text and sound documents) which are stored in the media library of the web interface. All three modules build a consistent whole that is interrelated and interdependent, and therefore they cannot be separated from each other. During the duration of the project, the current status of the modules VA\_DB and VA\_WEB is “frozen” simultaneously in the form of an electronic copy every six months, on 15 June and 15 December each year. These frozen copies receive versioning numbers according to the scheme [year]/[sequence number] (e.g. 15/1). Every productive VA version is named XXX. Due to the large size of media files it is impossible to produce copies of the VA-media library (VA\_MT). For that reason, no copy of this module is produced in the course of a versioning process. Media files which have been stored there once cannot be removed from the VA media library as soon as just a single VA version relates to them. On the project platform it is possible to switch between the “productive” VA version—subject to constant changes—to the archived, “frozen” versions. Appropriate coloring of the background and of certain control elements makes clear whether the productive or an archived version of VerbaAlpina is displayed (c.f. Lücke 2017f).

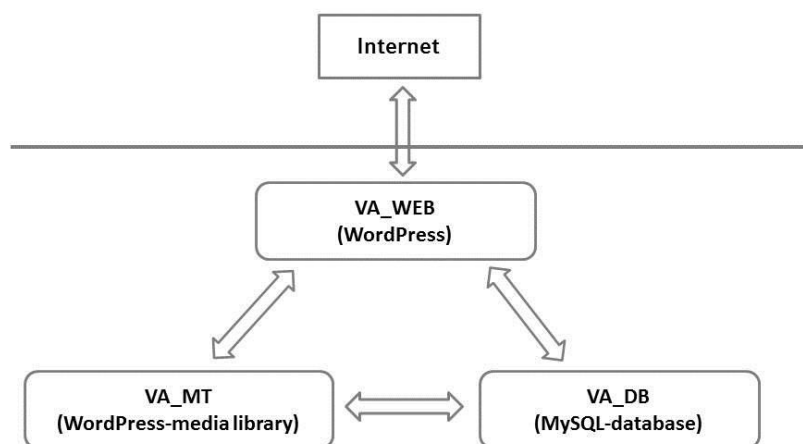


Figure 6: The modules of VerbaAlpina

#### 3.2 Citability

The versioning process also makes it possible for the contents of VerbaAlpina to be accurately cited. Unlike common online sources, the date of last access is not necessary since the cited versions (unlike the productive version XXX) are stable and will not be changed. Thus, VerbaAlpina as an Internet-based digitized source becomes viable for academic writing. As part of the bibliography VerbaAlpina can be cited in the following way:

VerbaAlpina (VA), [http://www.verba-alpina.gwi.uni-muenchen.de,\[version\]](http://www.verba-alpina.gwi.uni-muenchen.de,[version])  
 e.g.: VerbaAlpina (VA), <http://www.verba-alpina.gwi.uni-muenchen.de, 15/1>

For citations of contributions that belong to the menu “methodology” on the project platform, the following scheme is recommended:

[author/s]: s.v.\* “[lemma]”, in: VA-[language code according to ISO 639-1] [version], methodology, [URL]  
 e.g.: Krefeld, T. / Lücke, S.: s.v. “Tipizzazione”, in: VA-it 15/1, Metodologia, [http://www.verba-alpina.gwi.uni-muenchen.de/it/?page\\_id=21&letter=T#tipizzazione](http://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=21&letter=T#tipizzazione)

The authors, marked with “auct.”, at the end of every contribution, always need to be mentioned. Besides textual contents created by VerbaAlpina, graphic ones may also be cited. Every view of the interactive map can easily be shared and cited, since an individual URL is produced for every page and every pop-up window (c.f. Lücke/Oberholzer 2017).

### 3.3 Long-term Archiving

The project takes several steps to guarantee the sustainable use of all data. First, we take care to entrust several institutions with the long-term archiving of the data. Second, we document the data structuring, the logical relationships between data and data categories, and the applied character encoding. There are several different options to archive the project data by third parties. We plan to have multiple copies of the project data stored by several institutions. Currently, our data is saved on a regular basis by the IT-Gruppe Geisteswissenschaften of Munich University (ITG, IT group of humanities, <http://www.itg.uni-muenchen.de/index.html>) on backup servers of the Leibniz Rechenzentrum (LRZ, the Leibniz Computing Centre). At the same time that the data is archived, the different versions of VerbaAlpina are also created. At random intervals the module VA\_WEB is also stored in the Internet Archive (<https://archive.org>). In addition to the automatic archiving by archive.org through their wayback crawler, VerbaAlpina also actively archives the data (since 2018 on a regular basis in the course of the versioning every six months). We also intend to store further backup copies at other appropriate institutions like CLARIN-D. We hope in the long term to have the archiving conducted by the University Library of Munich so that project contents are also accessible via the electronic catalogues (c.f. Lücke 2017d).

## References

- AdA = Elspaß, S., Möller, R. (2003 ff.). *Atlas zur deutschen Alltagssprache*. Universität Salzburg.  
 AIS = Jaberg, K., Jud, J. (1928-1940). *Sprach- und Sachatlas Italiens und der Südschweiz*. Zofingen.  
 ALD-I = Goebel, H. (1998). *Atlant linguistich dl ladin dolomitich y di dialec vejins I* (sprechend: <http://ald.sbg.ac.at/ald/ald-i/index.php>). Wiesbaden: Reichert, vol. 1-7.  
 ALD-II = Goebel, H. (2012). *Atlant linguistich dl ladin dolomitich y di dialec vejins*. Editions de Linguistique et de Philologie, vol. 1-5.  
 ALJA = Martin, J.-B., Tuaillon, G. (1971, 1978, 1981). *Atlas linguistique et ethnographique du Jura et des Alpes du nord*. Paris: Éd. du Centre National de la Recherche Scientifique.  
 ALF = Gilliéron, J., Edmont, E. (1897-1900). *L'Atlas linguistique de la France*. Paris: Champion.  
 ALP = Bouvier, Jean-Claude (1975, 1979, 1986). *Atlas linguistique et ethnographique de la Provence*. Paris: Éd. du Centre National de la Recherche Scientifique, vol. 1, 2, 3.  
 Bradley, K. (2007). Defining digital sustainability. In *Library Trends*, (56)1, pp. 148-163.  
 Krefeld, T. (2017a). s.v. “Dokumentation”, in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=D#21](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=D#21).



- Krefeld, T. (2017b). s.v. "Ergänzende Daten", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=E#3](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=E#3).
- Krefeld, T. (2017c). s.v. "Interlinguale Geolinguistik", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=I#32](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=I#32).
- Krefeld, T. (2017d). s.v. "Kooperation", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=K#22](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=K#22).
- Krefeld, T. (2017e). s.v. "Synoptische Karte", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=S#56](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=S#56).
- Krefeld, T. (2017f). s.v. "Sprachatlant im Alpenraum", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=S#52](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=S#52).
- Krefeld, T. (2017g). s.v. "Transkription", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=T#57](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=T#57).
- Krefeld, T., Lücke, S. (2017a). s.v. "Betacode", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=B#7](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=B#7).
- Krefeld, T./Lücke, S. (2017b). s.v. "Crowdsourcing", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=C#12](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=C#12).
- Krefeld, T./Lücke, S. (2017c). Nachhaltigkeit – aus der Sicht virtueller Forschungsumgebungen. Korpus im Text. Version 7 (10.03.2017, 12:27). url: <http://www.kit.gwi.uni-muenchen.de/?p=5773&v=7>.
- Krefeld, T., Lücke, S. (2017d). s.v. "Referenzwörterbücher", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=R#51](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=R#51).
- Krefeld, T., Lücke, S. (2017e). s.v. "Typisierung", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=T#58](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=T#58).
- Krefeld, T./Lücke, S. (2014b). VerbaAlpina - Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit. In *Lad- inia*, XXXVIII, pp. 189-211.
- Lücke, S. (2017a). s.v. "Codepage", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=C#11](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=C#11).
- Lücke, S. (2017b). s.v. "Datenzugriffsschicht", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=D#70](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=D#70).
- Lücke, S. (2017c). s.v. "Digitalisierung", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=D#15](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=D#15).
- Lücke, S. (2017d). s.v. "Langzeitarchivierung", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=L#40](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=L#40).
- Lücke, S. (2017e). s.v. "Quantifizierende Darstellungen", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=Q#88](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=Q#88).
- Lücke, S. (2017f). s.v. "Versionierung", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=V#61](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=V#61).
- Lücke, S./Oberholzer, S. (2017). s.v. "Zitierweise", in: VA-de 17/2, Methodologie, [https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=493&db=172&letter=Z#64](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=172&letter=Z#64).
- Maron, N., Smith, K. K. & Loy, M. (2009). *Sustaining digital resources: an on-the-ground view of projects today: Ithaka case studies in sustainability*. Ithaka: S+R.
- Oberholzer, S., Kunzmann, M. (2017). Geolinguistic Documentation of Multilingual Areas – VerbaAlpina and the Challenges of Digital Humanities. In I. Buchstaller, B. Siebenhaar (eds.). *Language Variation - European Perspectives VI. Selected papers from the 8th International Conference on Language Variation in Europe (ICLaVE 8)*. Leipzig: Benjamins, pp. 199-214.
- SLA = Skofic, J. (eds.) (2011ff.). *Slovenski lingvistični atlas 1. Človek – telo, bolezni, družina*. Ljubljana: ZRC SAZU.
- VALTS = Gabriel, E. (1985-2004). *Vorarlberger Sprachatlas mit Einschluss des Fürstentums Liechtenstein, Westti- rols und des Allgäus, vol. 1-5*. Bregenz: Vorarlberger Landesbibliothek.
- WALS = Haspelmath, M. (2005ff.). *World Atlas of Language Structures*. Leipzig: Max Planck Institut für evolu- tionäre Anthropologie.
- Wiatr, A. (2016). Bedeutung und Funktion von Crowdsourcing im Projekt VerbaAlpina. In *Jezikovni zapiski*, (22)2, pp. 161-175.



## POSTER PRESENTATIONS



# Lexicographie et terminologie au XIX<sup>e</sup> siècle : *Vocabularu romano-francesu* [Vocabulaire roumain-français], de Ion Costinescu (1870)

**Maria Aldea**

*Université Babeş-Bolyai de Cluj-Napoca*

*E-mail: aldea\_maria@yahoo.com*

## Abstract

Dans cette étude, nous nous proposons d'analyser la manière dont un lexicographe roumain entend définir certains termes au sein d'un corpus choisi : il s'agit du *Vocabularu romano-francesu* [Vocabulaire roumain-français] élaboré par Ion Costinescu et paru en 1870 à Bucarest. Publié suite à une initiative privée quelques années après la fondation de l'Académie roumaine, ce dictionnaire trouve ses modèles déclarés dans la lexicographie française, surtout dans le Dictionnaire de l'Académie française et le Dictionnaire de Napoléon Landais. L'angle d'approche que nous privilégierons nous permettra, d'une part, de saisir la manière dont s'est déroulée l'une des étapes « pré-terminologiques » de la nouvelle discipline qui fait son apparition vers le milieu du XX<sup>e</sup> siècle, à savoir la science des termes et, d'autre part, de mesurer les termes y enregistrés à l'aune des principales tendances du développement culturel et scientifique de la société roumaine pendant la deuxième moitié du XIX<sup>e</sup> siècle.

**Keywords:** dictionnaire, terminologie, langue roumaine, langue française, modernisation, emprunt linguistique, Ion Costinescu

## 1 Introduction

Marquée par la création de la Société littéraire roumaine en 1866, celle qui deviendra un peu plus tard, en 1879, l'Académie roumaine, la Roumanie de la seconde moitié du XIX<sup>e</sup> siècle connaîtra à cette époque un tournant majeur concernant la production des dictionnaires et l'élaboration des outils lexicographiques. Bien que depuis sa création, l'Académie roumaine ait eu pour objectifs, entre autres, de fixer les normes de l'orthographe roumaine et d'élaborer un Dictionnaire trésor de la langue roumaine, projet qui a connu plusieurs tentatives dues aux attermolements des rédacteurs (voir le LM ; HEM ; DA, etc.), il existe des entreprises lexicographiques individuelles visant la rédaction à la fois de dictionnaires monolingues roumains, de dictionnaires spécifiques et de dictionnaires bilingues ou multilingues ayant pour langue de base le roumain ou une autre langue vivante.

Vers le milieu du XIX<sup>e</sup> siècle, la société roumaine de la Valachie et de la Moldavie se trouvait en plein processus de fondation du nouvel État national roumain. Ce processus s'accompagnait d'une série de réformes conçues dans le but de dynamiser le progrès culturel de cet espace ; projetées, tentées ou bien réalisées, ces réformes visaient des domaines aussi divers que l'enseignement, la culture, la justice, l'administration ou l'économie (voir Berindei 2003, VII / I : 329-870).<sup>1</sup> Quel que fût le domaine visé, les conséquences de ce processus de modernisation sont repérables aussi bien dans la dynamique de la langue que dans le vocabulaire, grâce à de nombreux emprunts faits par le roumain à plusieurs

1 Il convient de souligner que la Transylvanie connaissait une situation particulière à cette époque-là. Intégrée à l'Empire des Habsbourg, elle ne faisait pas encore partie de l'État nouvellement créé. Son union avec la Valachie et la Moldavie n'aura lieu qu'en 1918 (voir Pop, Năgler et Magyari 2008, III).



langues romanes et au latin médiéval. De tels emprunts ne trouvent pas toujours de référents dans les réalités de la société roumaine moderne, une société qui était d'ailleurs fortement tournée vers l'Occident (voir Niculescu 1978 ; Lupu 1999). Cette ouverture vers l'Occident est rendue visible tout d'abord par l'adoption de l'écriture en roumain à caractères latins (voir Aldea 2018). Dans ce contexte, il devient essentiel de créer de nouveaux outils pour surprendre et décrire ces réalités, de même que pour expliquer « les mots incompris qui envahissent massivement la langue par la presse et la chancellerie » (Costinescu 1870, I : [III] ; n. trad.). C'est ce qui explique l'apparition d'une longue série d'ouvrages lexicographiques.

## 2 La lexicographie roumaine et la terminologie: un bref survol historique

Les premières tentatives d'élaboration d'ouvrages lexicographiques dans l'espace culturel roumain se situent au XVI<sup>e</sup> siècle et elles concernent surtout des dictionnaires ou des projets de dictionnaires bilingues ou multilingues, ayant pour langue de base soit le roumain, soit le slavon.<sup>2</sup> À partir du XVIII<sup>e</sup> siècle, le slavon est remplacé par le latin et, par voie de conséquence, la typologie des dictionnaires commence à se diversifier de plus en plus. Ainsi, tout au long des XVIII<sup>e</sup>-XIX<sup>e</sup> siècles, vont paraître des dictionnaires bilingues, plurilingues, monolingues et surtout spécialisés (voir Seche 1966, I). L'examen de la nomenclature de ces dictionnaires nous permet de constater que les mots anciens et nouveaux qui y sont enregistrés appartiennent aussi bien à la langue commune qu'aux différents domaines d'activité scientifique, culturelle ou technique, ce qui nous oblige de placer les débuts de la « pré-terminologie » roumaine au XVIII<sup>e</sup> siècle.

Dans ce sens, on retient ici le *Lexicon italian-român* [Lexicon italien-roumain], rédigé vers 1700 par Constantin Cantacuzino, et tenu par les spécialistes pour « le premier dictionnaire spécialisé de termes scientifiques, plus exactement de termes géographiques » (cf. Seche 1966, I : 9) ; organisant l'information selon le critère de la matière, l'ouvrage manuscrit recense, à part des mots communs, des termes géographiques communs et des noms propres géographiques (cf. Seche 1966, I : 9).

Cette tentative est suivie par la parution en 1783, dans les pages de la revue *Magyar Könyvház*, du « premier glossaire de noms » de plantes en latin, hongrois et roumain, rédigé par Benkő József (cf. Seche 1966, I : 16-17). L'ouvrage recense 620 entrées, groupées d'après le critère de la classe de plantes.

Une autre tentative, restée elle aussi au stade de manuscrit et mettant en évidence la terminologie des sciences naturelles, est celle de Gh. Șincai. Intitulé *Vocabularium pertinens ad tria regna naturae* (datant des années 1808-1810), ce vocabulaire, rédigé en latin-roumain-hongrois-allemand et roumain-latin-hongrois-allemand, enregistre environ 427 termes désignant des noms de plantes, d'animaux et de minéraux (cf. Seche 1966, I : 24).

En 1822, paraît à Sibiu un *Vocabularium allemand et roumain* sous la plume de Ioan Molnar. L'ouvrage recense environ 8000 mots et il a le mérite d'introduire parmi les mots appartenant à la langue commune un grand nombre de termes médicaux (cf. Seche 1966, I : 28).

La production lexicographique roumaine sera « couronnée » en 1825 par la parution d'un dictionnaire en quatre langues : roumain, latin, hongrois et allemand. Connu sous le nom du *Lexicon de Buda* (abrégié LB<sup>e</sup>), celui-ci est tenu par les spécialistes pour un véritable repère dans la lexicographie roumaine moderne. Présentant environ 13000 entrées, le LB<sup>e</sup> enregistre de nombreux mots nouveaux appartenant au domaine technique et scientifique. Une analyse attentive de ces mots nous permet d'en

2 On désigne par « slavon » la variante littéraire tardive du vieux slave de l'église, qui pour l'espace cultural roumain, à cette époque, était encore la langue officielle de l'administration et de l'Église.

saisir qu'en général, les mots appartenant à différents domaines d'activité culturelle et scientifique sont traités comme des mots communs. Il y a peu de situations où l'entrée est accompagnée par une marque de domaine ou par une information qui apporte des précisions sur le registre et l'usage du mot (voir, par exemple, LB<sup>e</sup>, s.v. *caducitate* 'caducée', *çifra* 'chiffre', *multoratecu* 'pluriel' etc.).

Ainsi, tout au cours du XIX<sup>e</sup> siècle, paraissent dans l'espace roumain de nombreux ouvrages bilingues et plurilingues qui enregistrent dans leur nomenclature, au-delà des mots communs, non seulement des mots nouveaux, mais aussi des unités lexicales spécialisées.<sup>3</sup>

De cette série de dictionnaires, nous avons choisi de nous pencher dans la présente étude sur un seul : il s'agit de l'ouvrage lexicographique de Ion Costinescu, *Vocabularu romano-francesu* [Vocabulaire roumain-français] (désormais abrégé VRF).<sup>4</sup> Bien qu'il s'inscrive par son titre dans la série des dictionnaires bilingues, la lecture des entrées nous dévoile également un ouvrage lexicographique monolingue.<sup>5</sup>

### 3 Décrire les termes au XIX<sup>e</sup> siècle : *Vocabularu romano-francesu*

Puisant ses racines dans des ouvrages lexicographiques de référence de l'Europe occidentale, comme, par exemple, le *Dictionnaire* de Napoléon Landais, le *Dictionnaire* de l'Académie française et plusieurs dictionnaires italiens ou latins, le *Vocabulaire* de Ion Costinescu paraît en 1870 grâce au soutien financier et moral de l'évêque du diocèse de Buzău, Dionisie Romano (1806-1873). C'est à lui que Ion Costinescu (1810-1893) consacre explicitement son ouvrage.<sup>6</sup> Accueilli en 1850 au monastère de Băbeni, bénéficiant de la protection de l'évêque, Costinescu commence le travail à son vocabulaire animé par le désir de montrer l'état de « la culture littéraire » roumaine. La longue gestion du dictionnaire, dont la rédaction s'étend sur une période de vingt années même s'il est « sous presse » en 1857 (cf. VRF, s.v. *Y*), nous permet de reconstituer un état des lieux de l'évolution de la société, de la technique et des sciences à cette époque-là.

Costinescu commence par établir une distinction entre le *dictionnaire* qui est, selon lui, « un ouvrage classique qui ne peut pas être démolé par la critique, une œuvre devant laquelle s'inclinent à la fois le savant, l'ignorant et même l'envieux » (VRF, p. [III] ; n. trad.) et dont la particularité consiste dans le fait de contenir « tous les mots d'une langue, ordonnés alphabétiquement » (VRF, s.v. *dicționar* ; n. trad.), et le *vocabulaire* vu comme « une liste de mots roumains et romanisés classés selon un ordre alphabétique et accompagnés par une brève explication » (VRF, p. [III] ; n. trad.) : « recueil des mots les plus employés dans une langue accompagnés par une définition ou leur explication succincte. – Dictionnaire moins étendu. – Liste de mots appartenant particulièrement à une science, à un art, etc. » (VRF, s.v. *vocabular* ; n. trad.). Conscient des enjeux et des limites de sa tentative, Costinescu se propose ainsi d'élaborer et de mettre à la portée des lecteurs roumains un simple vocabulaire. À part les mots les plus courants et les mots anciens qui n'étaient plus employés que dans les conversations

3 Par exemple, sous la plume de S. Petri, paraîtra en 1861, à Sibiu, *Vocabularul portativ românesc-nemțesc* [Le Vocabulaire portatif roumain-allemand] qui, tout en imitant l'ouvrage de J. A. Vaillant, *Vocabularul purtător românesc-franțozesc și franțozesc-românesc* [Le Vocabulaire portatif roumain-français et français-roumain], paru en 1839, enregistre aussi des termes nouveaux provenant « des arts, des sciences et des métiers » (cf. Seche 1966, I : 42).

Un autre ouvrage, comptant environ 45000 mots (cf. Seche 1966, I : 45), très agréé à l'époque et qui a eu pour modèle le Dictionnaire de l'Académie française, est celui de P. Poienaru, F. Aaron et G. Hill. Intitulé *Vocabularul franțezo-românesc* [Vocabulaire français-roumain], il est paru à Bucarest en deux volumes, entre 1840-1841. La liste pourra bel et bien continuer.

4 L'ouvrage peut être consulté à la Bibliothèque Centrale Universitaire « Lucian Blaga » de Cluj-Napoca, sous les cotes 195405, 181704, 342326.

5 Ainsi, pour ce qui est de cette dernière catégorie, on mentionnera ici un manuscrit datant environ de 1832, intitulé *Condica limbii românești* [Un registre de la langue roumaine] et réalisé par Iordache Goleescu (cf. Seche 1966, I : 35).

6 Des informations sur la vie et l'œuvre de Ion Costinescu sont à retrouver dans les études de Cocora 1965 et Cocora 1977.

familiales, le VRF recense aussi « des termes techniques des sciences, des arts et des métiers », de même que « des termes mythologiques ayant un rapport avec les coutumes roumaines transmises par la tradition » (VRF, p. [III] ; n. trad.).

C'est sur de tels « mots techniques » ou appartenant à des « arts » que nous voudrions nous pencher dans ce qui suit. L'examen de l'article consacré au mot *termină* 'terme' nous permet de remarquer qu'il subit un traitement lexicographique polysémantique. À part son sens commun, le rédacteur enregistre aussi son emploi spécialisé : « en géométrie. Le point est le terme d'une ligne, la ligne est le terme d'une superficie, la superficie est le terme d'un corps solide. – en physique. Tout mouvement a deux termes, le terme de début et le terme d'arrêt. – en arts et sciences : termes d'architecture, de grammaire, de pratique, de physique, etc. les noms ou les expressions spécifiques à ces arts et sciences » (VRF, s.v. *termină* ; n. trad.). Avec ce dernier emploi spécialisé « en arts et sciences », nous pouvons affirmer que le mot *termină* acquiert une valeur propre ou spécifique qui le définira et le placera, presque un siècle plus tard, comme le noyau central d'un nouveau domaine, à savoir la terminologie. Si on analyse également les articles consacrés aux mots *tecnică* 'technique', *tehnologie* 'technologie', *știință* 'science', *artă* 'art' et *cultură* 'culture', on constate qu'ils bénéficient d'amples définitions explicatives. Par exemple, les articles consacrés aux unités lexicales *știință* 'science' et *cultură* 'culture' mettent en évidence le sens commun ou propre, tandis que les articles dédiés aux mots *tecnică* 'technique', *tehnologie* 'technologie' et *artă* 'art' apportent en plus des informations soit sur l'usage spécialisé du mot, soit sur des questions d'ordre encyclopédique. Si au sein de l'article consacré au mot *artă* 'art' (VRF, s.v. *artă*) on décèle la taxonomie explicite des arts (arts libéraux ; arts mécaniques ; beaux-arts), en ce qui concerne le mot *tecnică* 'technique', on constate qu'il apparaît sous la forme d'un doublet graphique (*tecnică*, *tehnică-ă*), étant placé dans des syntagmes qui éclairent son usage : « *șicere tehnice, termină tehnice* » 'mots techniques, termes techniques' (VRF, s.v. *tecnică* ; n. trad.). De plus, par le mot *tehnologie* 'technologie', le rédacteur désigne « la science des mots techniques, des mots qui appartiennent aux arts » (VRF, s.v. *tehnologie* ; n. trad.) : en fait, ce qu'on désigne aujourd'hui par la terminologie ou la science des termes (voir Cabré 1998 ; Mazière 1981-1982).

Le VRF compte environ 28000 articles. L'examen des entrées nous a permis d'observer que les unités lexicales enregistrées bénéficient en général d'un traitement lexicographique unitaire. Ainsi, le mot vedette (présent souvent dans une double ou triple entrée) est transcrit selon la norme orthographique spécifique au milieu du XIX<sup>e</sup> siècle, c'est-à-dire la norme imposée par la direction latiniste et étymologiste, qui envisageait un rapprochement très fort entre le mot et son étymon. En général, le mot vedette est rendu à la fin de l'article ou après son premier sens par son équivalent en français. Après le mot vedette, on indique en abrégé la classe grammaticale à laquelle le mot appartient (*s.* / *sus.* / *subs.* – substantif ; *v.s.* – verbe substantivé ; *adv.* – adverbe ; *adi.* – adjectif ; *prep.* – préposition ; *conj.* – conjonction ; *inter.* – interjection ; *pron.* – pronom) et la catégorie grammaticale du genre (*m.* – masculin ; *f.* – féminin ; *etr.* – hétérogène ou neutre) et/ou du nombre (*plr.* – pluriel). Dans le corps de l'article on trouve des définitions sous forme de commentaires amples, bien développés, ce qui permet d'encadrer l'ouvrage de Costinescu, malgré son titre, dans la série des dictionnaires explicatifs et universels de la langue roumaine. Dans la description du mot, le rédacteur emploie la présentation logique, tout en distinguant entre le sens fondamental et d'autres sens. L'explication du sens est souvent accompagnée par un exemple en roumain qui est rendu aussi en français. À l'exception des expressions latines, les locutions sont généralement traitées à la fois dans le corps de l'article et à travers des entrées distinctes. Il y a également une série de mots qui sont définis par des renvois. Une grande partie d'articles présentent dans leur corps des marques d'usage qui donnent des informations très précises sur la valeur d'emploi dans la société (« *șicere familiară* » 'mot familier', « *în stilă familiară* » 'dans le registre familier', « *în limba poporului* » 'dans la langue du peuple', « *în vorbirea ordinară* » 'dans le parler ordinaire') ou au long du temps (« *in tempi antichi* » 'dans les temps anciens',

« în vechime » ‘autrefois’, « la quei vechi » ‘chez les anciens’), sur le changement de sens (« în termeni metaforici » ‘en termes métaphoriques’), sur l’étymologie du mot (« ȳdicere latină » ‘mot latin’, « ȳdicere turcă » ‘mot turc’) ou le domaine de spécialisation, etc.

En ce qui concerne les domaines de spécialisation, l’examen de notre corpus nous a d’abord permis d’observer que le VRF présente dans le corps de certains articles (plus de 3000) soit l’indication du domaine, soit la marque « t<sup>er</sup>min<sup>u</sup> de... » ‘terme de...’, suivie de l’intitulé du domaine. Cette marque s’applique à la fois à une seule entrée et à une même entrée à laquelle correspondent deux ou plusieurs domaines différents.

Nous avons pu ensuite dresser une typologie de l’inventaire des unités lexicales à partir des marques de domaines enregistrées. Ainsi, grâce à ce marquage, nous avons identifié des termes appartenant à plusieurs domaines et sous-domaines, à savoir Mathématiques – Arithmétique, Algèbre, Géométrie etc. ; Physique et Optique ; Mécanique ; Chimie ; Astronomie ; Géographie ; Géologie et Minéralogie ; Histoire naturelle ; Hydraulique ; Marine ; Métallurgie ; Agriculture ; Botanique ; Médecine – Chirurgie, Anatomie ; Médecine vétérinaire ; Pharmacie ; Droit – Jurisprudence, Tribunal, Jugement, Procès etc. ; Sciences militaires et Art militaire, Guerre, Soldats, Artillerie, Fortifications ; Sciences économiques – Comptabilité, Commerce, Finances, Banque, Douane ; Sciences de l’éducation – Didactique ; Philologie – Grammaire, Calligraphie, Poétique, Littérature, Poésie, Prosodie ; Mythologie ; Imprimerie ; Librairie ; Philosophie – Logique, Dialectique, Philosophie ; Rhétorique ; Histoire – Histoire antique, Archéologie, Histoire des Romans, Antiquité, Blason, Féodalité, Chancellerie ; Théologie – Théologie, Dogmatique, Histoire ecclésiastique, Dévotions, Liturgie, Évangile, Morale etc. ; Art – Architecture, Peinture, Danse, Sculpture, Théâtre, Musique, etc. Comme on peut le constater, toutes ces « étiquettes » de domaines et de sous-domaines indiquent des branches fondamentales des sciences : sciences exactes, sciences de la terre et de l’atmosphère, sciences naturelles et médicales, sciences agricoles, sciences sociales, sciences humaines, arts etc. Mais ce qui frappe, ce sont les différentes notations pour un et même domaine, qui sont rendues en forme soit abrégée, soit complète, comme suit :

- Pour la chimie : « t. de him. », « t. de chim. », « him. », « în chimie », « în himie » ;
- la médecine : « t. de med. », « t. de medi. », « t. med. », « în t. de med. », « med. », « medi. », « în med. », « în medic. », « în medicină » ;
- la chirurgie : « t. de hir. », « t. de hirur. », « t. de chir. », « t. de chirur. », « hir. », « hirur. » ;
- l’anatomie : « t. de anat. », « t. de ant. », « t. d’anat. », « t. de anato. », « de anat. », « anat. » ;
- la botanique : « t. de bot. », « t. de b. », « t. b. », « bot. », « botan. », « în bot. », « în botanică » ;
- la musique : « t. de mus. », « t. de music. », « t. de musi. », « t. de m. », « t. de musică », « mus. », « music. », « în musică » ;
- la grammaire : « t. de gram. », « t. de grm. », « t. de gramat. », « fig. de gramatică », « de gram. », « gram. », « în gramat. », « în gram » ;
- la philosophie : « t. de fil. », « t. de filos. », « în filos. », « în filosofie » ;
- la rhétorique : « t. de retor. », « fig. de retorică », « fig. de retor. », « figură de retorică », « retor. », « retorică », « în retorică » ;
- la marine : « t. de mar. », « mar. », « în marină » ;
- l’astronomie : « t. de ast. », « t. astr. », « t. de astr. », « t. d’astr. », « t. de astro. », « t. de astron. », « astr. », « astronomii », « în astronomie » ;
- les mathématiques et les sous-domaines adjacents : « mat. », « matem. », « t. de mat. », « t. mat. », « în mat. », « arit. », « aritm. », « aritmetică », « t. de arit. », « t. de aritm. », « t. de ari. », « algeb. », « t. de alg. », « t. de algeb. », « în algebră », « în t. de algeb. », « geom. », « geomet. », « t. de geom. », « t. de geo. », « t. de geomet. », « în geom. », « în geometr. », « în geometrie », etc.

La liste ci-dessus est loin d’être exhaustive.



Toutes ces variations graphiques synonymiques au niveau de la notation de l'intitulé du domaine, marquant explicitement l'appartenance à une spécialisation donnée, expriment de ce point de vue un manque de cohérence de la part du rédacteur du à la distance temporelle installée entre la parution de la tranche alphabétique de la lettre A en 1857, du premier volume en 1859 et du seconde volume en 1870 ; pour des raisons typographiques, tous les deux volumes ont été imprimés à nouveaux en 1870.

En dernier temps, nous nous proposons d'analyser la manière dont ces termes sont décrits.

Ainsi, la lecture de notre corpus nous a permis d'identifier plusieurs manières de présentation de ceux-ci :

(a) la définition de type commentaire explicatif est précédée soit par la forme abrégée de l'indication du domaine, soit par la formule « terme de + nom du domaine » délimitée graphiquement soit par un « point », soit par un « deux-points » à valeur d'un « point » :

- (1) Zorille. *s.f. ist. nat.* unŭ animalŭ quatrupedŭ, de felulŭ jderilor, quare locuesce pe la capulŭ de buna speranŭia ŭn Africa. Zorille. (VRF, *s.v. Zorille*) [Zorille. *s.f. hist. nat.* un animal quadrupède, semblable à la martre, qui habite dans la région du cap de Bonne-Espérance, en Afrique. (n. trad.)]
- (2) Eclimetru. *s.m. t. noŭ de geom.* Unŭ felŭ de grafometru de mesuratŭ inclinaŭiunea unuŭ tŕimŭ. Eclimètre. (VRF, *s.v. Eclimetru*) [Éclimètre. *s.m. t. nouveau de géom.* Un type de graphomètre pour mesurer la pente d'un terrain. (n. trad.)]
- (3) Anelide. *s.f. plr : t. de ist : nat :* Nume allu unei clase de vieŭitŕorie quare coprinde vermii queŭ cu ŭnge roŭiu, ŭi allu quârŕoru corpu este inelatu ŭn curmedziŭ. *Annelides.* (VRF, *s.v. Anelide*) [Annélides. *s.f. pl. t. d'hist. nat.* Nom d'une classe d'animaux qui comprend les vers à sang rouge et dont le corps est divisé en anneaux. (n. trad.)]
- (4) Animŭŭti. *s.m. plr : t. de filos :* Se ŭice asfelu aquellora quarŭi attribuescu suffletului tŕte fenomenile economiei animale. *Animistes.* (VRF, *s.v. Animŭŭti*) [Animistes. *s.m. pl. t. de philos.* On dit ainsi à propos des gens qui rapportent à l'âme tous les phénomènes de l'économie animale. (Landais 1834, *s.v. Animistes*)]
- (5) Irreductibilŭ-ă. *adi. t.chim. ŭi de chirur.* Quare nu se maŭ pŕte reduce la starea dintŭu. – *t. de algeb.* Quare nu se maŭ pŕte reduce la uă altă formă maŭ simplă. *Irréductible.* (VRF, *s.v. Irreductibilŭ-ă*) [Irréductible. *adj. t. chim. et de chirur.* Qu'on ne peut réduire à l'état initial. – *t. d'algèb.* Qu'on ne peut réduire à une forme plus simple. (n. trad.)]

En examinant les exemples ci-dessus, on note que le premier, entrée *Zorille*, ne soulève pas des problèmes ; par contre, pour le deuxième, entrée *Eclimetru*, le rédacteur attire l'attention sur la nouveauté de l'unité lexicale tout en incluant la mention « nouveau » dans la formule indiquant le domaine (cet adjectif est absent dans le dictionnaire de Landais). En même temps, il introduit la définition par des expressions telles que : « un type de... », « nom d'une... », « on le dit... » (voir les exemples 2, 3 et 4, entrées *Eclimetru* ; *Anelide* ; *Animŭŭti*). Quant à l'exemple 5, entrée *Irreductibilŭ-ă*, on remarque qu'au sein du même article l'unité lexicale est définie comme appartenant à plusieurs domaines de spécialité soit dans le même sens, soit dans des sens distincts.

(b) le domaine de spécialisation est indiqué au sein de la définition en position soit initiale soit médiane ou finale, en présence ou non d'un verbe métalinguistique du type « se ŭice » 'on (le) dit' ou « se numesce » 'on appelle', ou du verbe « Se ŭea ŭn... » 'il est employé...' :

- (6) Oxidŭ. *s.etr.* In himia modernă, substanŭia combinată cu oxigenŭ, nu ŭnsă pŭnă la starea de acidŭ. *Oxidele metalice* sunt quea que ŭn vechia doctrină se numia fŕte greŭitŭ, vărurŭ



metalice. – *Oxidă sticlosă*, Sticlă metalică. *Oxyde*. (VRF, s.v. *Oxidă*) [*Oxyde*. s.m. Dans la chimie moderne, substance combinée avec l'oxygène, mais non jusqu'au point d'être ramenée à l'état d'acide. *Les oxydes métalliques* sont ce que dans l'ancienne doctrine on appelait très improprement chaux métalliques. – *Oxyde vitreux*, Verre métallique. (Landais 1834, s.v. *Oxyde*)]

- (7) *Facere*. v.s. [...] Se ăia asemenea cu mai multe înțelesuri în termină de marină, de guerră, de medicină, de botanică, de totă felulă de sciinție și de arte. [...]. (VRF, s.v. *Facere*) [*Faire*. v. [...] Il est employé aussi avec plusieurs sens dans des termes de marine, de guerre, de médecine, de botanique, de toutes sortes de sciences et d'arts. [...]. (n.trad.)]
- (8) *Banco*. adi. Dăcere întrebunțată de comercianți spre a deosbebi printr'însa cursulă bani-loră la bancă din cursulă loră în dar-averi. *Banco*. (VRF, s.v. *Banco*) [*Banco*. adj. Mot employé par les commerçants pour distinguer entre le cours de l'argent à la banque et le cours de l'argent entre plusieurs individus. (n.trad.)]
- (9) *Circulantă-ă*. adi. Dăcere forte întrebunțată în comerțu, quare circulă, quare este în circulațiune. *Circulant-e*. (VRF, s.v. *Circulantă-ă*) [*Circulant-e*. adj. Mot fort en usage dans le commerce, qui circule, qui est en circulation. (Landais 1834, s.v. *Circulant*)]
- (10) *Acrotère*. s.plr : Ună felă de ornamente în arhitectură. Ună mică piedestală subtă frontispiciu pe quare staă vase, statuie. – t. de mar. Capă saă promontoriu. *Acrotère*. (VRF, s.v. *Acrotère*) [*Acrotère*. s.m. sg. Un type d'ornements en architecture. Petits piédestaux au-dessus d'un frontispice pour servir de support à des vases ou à des statues. – t. de mar. Cap ou promontoire.]
- (11) *Aerofonă*. s.etr : Ună noă instrumentă de musică de vîntă, cu clape. *Aérophone*. (VRF, s.v. *Aerofonă*) [*Aérophone*. s.m. Un nouvel instrument de musique à vent ou à clavier.]

L'analyse de l'entrée *Oxidă* (voir l'exemple 6) nous permet d'observer surtout la dimension encyclopédique de l'information contenue dans la définition. L'exemple 7, entrée *Facere*, trait sous une valeur secondaire une précision relative au domaine de spécialisation dans lequel on l'emploie. En ce qui concerne les exemples 8 et 9 (entrées *Banco* ; *Circulantă-ă*), on observe que la définition commence par une expression du type « dăcere întrebunțată » 'mot employé' (où l'adjectif « employé » est marqué ou non par un degré d'intensité), étant suivie soit par une référence au cercle des représentants d'un domaine donné (exemple 8, entrée *Banco*), soit par l'intitulé du domaine (exemple 9, entrée *Circulantă-ă*). Pour les exemples 10 et 11, entrées *Acrotère* ; *Aerofonă*, on note que le domaine de spécialisation est présent au sein de la définition, étant exprimé par des structures telles que : « un type d'ornements en... », « ornement de... », « un instrument de... » etc.

(c) le domaine de spécialisation ressort de la définition par une référence directe soit au nom d'un spécialiste, soit à un cercle de spécialistes d'un domaine :

- (12) *Apăducă*, *Apeducă*. s.etr. Canală pentru aducerea apeă dintr'ună locă într'altulă. Analogicamente, în limbăgiulă anatomistiloră : ore quare locuri alle corpului pe unde intră saă trece apa. *Aqueduc*. (VRF, s.v. *Apăducă*) [*Aqueduc*. s.m. Canal pour conduire les eaux d'une place à l'autre. Analogiquement, en langage des anatomistes, certaines parties du corps par où l'eau pénètre ou court. (Landais 1834, s.v. *Aqueduc*)]
- (13) *Factură* și *Fatură*. s.f. [...] – In arte, fasonulă dupe quare este făcută ună lucru. In t. organiștiloră, qualitatea, lărgimea și lungimea țieveloră. *Facture*. (VRF, s.v. *Factură*) [*Facture*. s.f. [...] – Dans les arts, la façon dont une chose est faite. Dans les termes des organistes, la qualité, la largesse et la longueur des tuyaux.]

- (14) Sectă. *s.f.* [...] – În materie de religie, opiniune eretică și eronată. – *Secte*. (VRF, *s.v.* Sectă) [Secte. *s.f.* [...] – En matière de religion, opinion hérétique ou erronée. (Landais 1834, *s.v.* Secte)]
- (15) Teoremă. *s.f.* între matematici, propozițiunea unei verități speculative que se pôte demunstra. Differă de problemă într'atâta quâ aquêsta este uă propozițiune de veritate practică. *Théorème*. (VRF, *s.v.* Teoremă) [Théorème. *s.m.* chez les mathématiciens, propositions d'une vérité spéculative qu'on peut démontrer. Il diffère de problème, en ce que celui-ci est une proposition de vérité pratique. (Landais 1834, *s.v.* Théorème)]
- (16) Oxigenū. *s.etr.* Nume datū de himiștiū modernī principiului acidificantū său generatorū allū aciduluī. Oxigenulū este basa aeruluī vitalū, numitū altă-dată aerū diflogisticatū. Topitū în caloricū și lumină, forméză gazulū oxigenū, său aerulū vitalū atmosfericū, și amestecatū în aquêstă stare cu trei părți aprôpe de gazū azotū (în proporțiunea lui 27 către sută), constitue aerulū atmosfericū. Combinatū cu diferite base, oxigenulū formésă oxidele și acidele. *Oxygène*. (VRF, *s.v.* Oxigenū) [Oxygène. *s.m.* Nom donné par les chimistes modernes au principe acidifiant ou générateur de l'acide. L'oxygène est la base de l'air vital, appelé autrefois air déphlogistique. Fondu dans le calorique et la lumière, il forme le gaz oxygène ou l'air vital atmosphérique, et mêlé dans cet état avec trois parties environ de gaz azote (dans la proportion de vingt-sept à cent), il constitue l'air atmosphérique. Combiné avec différentes bases, il forme les oxydes et les acides. (Landais 1834, *s.v.* Oxygène)]
- (17) Cartesianismū. *s. etr.* Filosofia lui Descartū. *Cartésianisme*. (VRF, *s.v.* Cartesianismū) [Cartésianisme. *s.m.* Philosophie de Descartes. (Landais 1834, *s.v.* Cartésianisme)]

En examinant les exemples de 12 à 16, entrées *Apâducū*, *Apeducū* ; *Facturâ* și *Faturâ* ; *Sectă* ; *Teoremă* ; *Oxigenū*, on observe que les définitions et les sens spécialisés commencent par des syntagmes ou des structures du type « în limbagiulū... » 'dans le langage de...', « în t'ermeniiū... » 'dans les termes de...', « în materie de... » 'en matière de...', « între... » 'chez...', « nume dat de... » 'nom donné par...' etc., qui sont précisés par une référence aux spécialistes d'un domaine, tandis que dans l'exemple 17, entrée *Cartesianismū*, on note une désignation directe au nom d'un spécialiste.

Nous n'avons mentionné que quelques exemples, mais la liste pourra bel et bien continuer.

## 4 En guise de conclusion

Par la présente étude nous avons voulu retracer une des étapes de la « pré-terminologie » roumaine, d'une part, et d'autre part, observer dans quelle mesure les termes enregistrés dans un ouvrage lexicographique particulier reflètent les tendances du développement culturel et scientifique de la société roumaine pendant la seconde moitié du XIX<sup>e</sup> siècle. Cela faisant, nous avons pu remarquer la place occupée par le VRF dans l'histoire de la lexicographie roumaine.

Par rapport à d'autres dictionnaires roumains de l'époque parus antérieurement, le VRF enregistre à la fois des mots communs et usuels, et des unités lexicales spécifiques à plusieurs domaines de spécialisation. La lecture de notre corpus nous a permis de découvrir l'architecture et l'organisation interne du VRF. Suite à l'analyse des unités lexicales sélectionnées, nous avons pu constater qu'au niveau de la structure de l'article, le VRF reprend parfois telles quelles des définitions présentes dans l'ouvrage de Landais (voir les exemples 4, 6, 9, 12, 14, 15, 16, 17). Dans d'autres cas, le rédacteur apporte des ajouts ou des suppressions à l'explication puisée chez Landais (voir les exemples 10, 11, 13) ou, effectivement, il donne ses propres définitions (voir les exemples 1, 2, 3, 5, 7, 8). Dans tous les exemples puisés on observe que le rédacteur accorde aux mots roumains l'encadrement correct

du point de vue morphologique, parfois différent par rapport à la langue française. L'influence forte de l'ouvrage de Landais s'explique par la diffusion des idées des Lumières françaises dans les Principautés roumaines de même qu'en Transylvanie, ainsi que par un retour symbolique souhaité par les Roumains aussi bien vers « l'espace romane que vers les sources culturelles profondes de la langue roumaine », retour « réalisé concrètement par l'accueil dans la langue roumaine, tout au cours du XIX<sup>e</sup> siècle, de nouveaux éléments latins et néolatins. » (Aldea 2017 : 18). Néanmoins, Costinescu a le mérite d'avoir mis à la portée d'un large public un excellent instrument de travail contenant des explications amples, détaillées qui facilitaient la compréhension. Son effort intellectuel de trouver les « bons mots roumains » pour transmettre des sens et des réalités techniques ou culturelles parfois absentes dans la langue cible, inscrit sa tentative dans la direction théorique qu'on appelle de nos jours « sémantique désignationnelle ».

Nous concluons cet exposé en déclarant que nous l'avons pensé comme un préambule pour un futur projet de recherche. En nous appuyant sur un autre projet que nous avons dirigé et consacré à l'édition électronique du *Lexicon de Buda*, le premier dictionnaire roumain ancien qui ait été digitalisé (voir Leucuța et alii 2012, Aldea 2016, LB<sup>e</sup>), nous souhaitons à l'avenir réaliser une édition informatisée du VRF, qui mettra en évidence toutes ces unités lexicales appartenant à différents domaines de spécialité et qui permettra, par ces équivalents français enregistrés dans le corps de l'article, d'établir des rapports virtuels avec d'autres dictionnaires français informatisés.

## Bibliographie

- Aldea, M. (2018). L'enjeu de l'orthographe dans le processus d'affirmation de la langue roumaine. In *Actes du XXVIII<sup>e</sup> Congrès international de Linguistique et de Philologie Romanes (Rome, les 18-23 juillet 2016)*. (sous presse).
- Aldea, M. (2017). Reromanizarea limbii române în viziunea lui Sextil Pușcariu. In *Caietele Sextil Pușcariu*, III, pp. 15-20.
- Aldea, M. (2016). Un projet accompli : le *Lexicon de Buda* (1825) en édition électronique. In *Proceedings of the 17th EURALEX International Congress 6-10 September 2016, Tbilisi*, Edited by Tinatin Margalitadze, George Meladze. Ivane Javakhishvili Tbilisi University Press, pp. 856-862.
- Berindei, D. (éd.) (2003). *Istoria românilor*, vol. VII, tom I. București : Editura Enciclopedică.
- Cabré, M.T. (1998). *La terminologie : théorie, méthode et applications*. Ottawa : Armand Colin-Les Presses de l'Université d'Ottawa.
- Cocora, Gabriel (1965). Știri despre viața și opera lexicografului Ion Costinescu. In *Limba română*, nr. 1, pp. 167-173.
- Cocora, Gabriel (1977). Cum s-a tipărit primul dicționar explicativ în limba română. In Idem, *Tipar și cărturari*, Cu o prefață de Dan Zamfirescu. București : Litera.
- Costinescu, I. (1870). *Vocabularu romano-francesu*, lucratu dupe Dicționarulu Academiei Francese dupe alu lui Napoleone Landais și alte Dicționare latine, italiene, etc. Vol. I-II. Bucuresci : Tipographia Naționala Antreprenor C.N. Rădulescu.
- DA. (Academia Română) *Dicționarul limbii române*. Sub conducerea lui Sextil Pușcariu. Tomul I. Partea I : A-B, 1913 ș.u. București : Librăriile Socec & Comp. și C. Sfetea.
- HEM. Bogdan Petriceicu Hasdeu, *Etymologicum Magnum Romaniae. Dicționarul limbei istorice și poporane a românilor*. Tom. I-II, 1887, tom. III, 1893. București : Stabilimentul grafic Socec și Teclu.
- Landais, N. (1834). *Dictionnaire général et grammatical des dictionnaires français*, vol. I-II. Paris : Imprimerie d'Éverat. Consulté sur : <http://gallica.bnf.fr> [14/09/2017].
- LB<sup>e</sup>. Pour l'édition électronique de Aldea, Maria (coord.), 2013. *Lesicon romanescu-latinescu-ungurescu-nemtescu quare de mai mulți autori, in cursul a trideci, si mai multoru ani s'au lucrat. Seu Lexicon valachico-latino-hungarico-germanicum quod a pluribus auctoribus decursu triginta et amplius annorum elaboratum est*. Budae, Typis et Sumtibus Typographiae Regiae Universitatis Hungaricae, 1825. Ediție electronică de Maria Aldea, Daniel Corneliu Leucuța, Lilla-Marta Vremir, Vasilica Eugenia Cristea, Adrian Aurel Podaru. Cluj-Napoca. Consulté sur : <https://doi.org/10.26424/lexiconuldelabuda> [10/12/2017].

- Leucuța, D.C., Harhăță, B., Vremir, L.M. & Aldea, M. (2012). The Romanian-Latin-Hungarian-German Lexicon - The Lexicon of Buda (1825). Informatics Challenges for an Emended and On-Line Ready Edition. In *Proceedings of the 15th EURALEX International Congress 7-11 August 2012, Oslo*, Edited by Ruth Vatvedt and Julie Matilde Torjusen. Reprosentralen, UIO, pp. 903-909.
- LM. A.T. Laurian și I.C. Massim, *Dicționarul limbei romane*. Tomu I (A-H), 1871 (1873); tomu II, 1876; tom. III : Glossariu, sau care cuprinde vorbele din limba română străine prin originea și forma lor, cum și cele de origine înduioasă, 1871 (1877). București : Noua tipografie a laboratorilor români.
- Lupu, C. (1999). *Lexicografia românească în procesul de occidentalizare latino-romanică a limbii române moderne (1780-1860)*. București : Logos.
- Mazière, F. (1981-1982). Le dictionnaire et les termes. In *Cahiers de lexicologie*, no 39, pp. 81-104.
- Niculescu, Al. (1978). Occidentalizarea romanică a limbii și culturii românești moderne. In Idem, *Individualitatea limbii române între limbile romanice. 2. Contribuții socioculturale*. București : Editura Științifică și Enciclopedică, pp. 55-98.
- Pop, I.-A., Năgler, Th. & Magyari, A. (éd.) (2008). *Istoria Transilvaniei*, Vol. III (De la 1711 până la 1918). Cluj-Napoca : Academia Română – Centrul de Studii Transilvane.
- Seche, M. (1966). *Schiță de istorie a lexicografiei române*, Vol. I. De la origini până la 1880. București : Editura Științifică.
- VRF. Pour l'ouvrage de Costinescu (1870).

# Developing a Russian Database of Regular Semantic Relations Based on Word Embeddings

**Ekaterina Enikeeva, Andrey Popov**

*Saint Petersburg State University*

*E-mail: protoev@yandex.ru, hedgeonline@gmail.com*

## Abstract

Recent computational semantic models yield high-quality results with regard to semantic relations extraction tasks, and thus may be applied as a baseline for semantic lexicon construction. Moreover, the stochastic information about lexical compatibility is useful for reducing ambiguity and detecting anomalies during syntactic parsing. We prove that this approach is reasonable and describe a Russian semantic lexical database, acquired in an unsupervised manner and employed as a semantic component of a syntactic parser and a fact extraction system.

**Keywords:** distributional semantics, word vector representations, semantic lexicon, Meaning  $\leftrightarrow$  Text model

## 1 Introduction

Semantic components have usually been seen as crucial in natural language processing and understanding systems. A semantic lexicon may be constructed as a lexical database, such as WordNet<sup>1</sup>, which maps concepts into lexemes, groups lexemes into synonymic sets, describes a number of relations between them and provides short definitions and usage examples. A more complicated variation, thus applicable in a wider range of NLP tasks, is an extensive dictionary of a language recording all possible information about lexical units. Such descriptions are rather laborious, however, since they require a comprehensive analysis of each lexeme. Therefore, there are only a few known completed examples of this approach. As regards Russian language resources, we should mention Explanatory Combinatorial Dictionary (Mel'čuk, Zholkovsky 1984) based on the Meaning  $\leftrightarrow$  Text linguistic model (Mel'čuk 1974/1999). For each described lexeme, an entry comprises its subcategorization frame, lexical co-occurrence data and a set of examples supplied with morphological description.

As noted above, this kind of lexical resource requires lots of manual work, including corpus studies and a thorough lexicographic description. However, recent studies have introduced a variety of computational semantic models that may transform this process into at least a semi-supervised one. These models are based on a distributional hypothesis (cf. the review in Sahlgren (2008)), which claims that a meaning of a word may be derived from its context; machine learning approaches are then applied to produce a computationally effective representation of words (word embeddings) that incorporates contextual information (Mikolov et al. 2013a). A “semantic space” built in such a way appears to reflect regular relations between lexical units; one can perform simple vector operations to infer examples of paradigmatic relations; a widely-cited example being the one of *king* – *man* + *woman* = *queen*.

Semantic word embeddings have numerous advantages: they are learnt from raw corpora and require virtually no external linguistic resources. Recently developed models are easily interpretable and may be applied to a number of tasks, including paradigmatic relations extraction, predicting semantic

<sup>1</sup> <https://wordnet.princeton.edu/>



analogy and describing selectional restrictions (cf. SemEval task on semantic comparison for words and texts, RUSSE contest, etc.).

We propose an integrated system for building a semantic lexicon for Russian from raw text corpora using few linguistic resources. Our lexical database is designed to reinforce the syntactic parser and fact extraction system working with the Russian language. We aim at overcoming the syntactic ambiguity issue and reducing the number of possible parse trees for a sentence (Popov & Enikeeva 2017), so in this paper we focus mainly on the description of syntagmatic relations. Instead of just listing selectional restrictions for each lexeme, the lexicon includes collocational probability even for those combinations that do not occur in training corpus.

The paper is structured as follows: firstly, we describe the related work, including lexicographical applications of word embeddings. Then we briefly outline our computational approach to describing syntagmatic relations within the lexicon, and provide an overview of corpora and other data we use. In the Evaluation section we report the system quality in terms of widely-used simple metrics (precision of collocates ranking), and discuss the necessity of more elaborated annotation. Finally, an automated lexical description for several lexemes is presented.

## 2 Related Work

Compositional distributional semantics is successfully applied to a number of semantics-related tasks in NLP. As far as evidence for Russian is concerned, a number of semantic vector models were evaluated during the RUSSE workshop (Panchenko et al. 2015). However, they focused on paradigmatic relations between lexical units: the annotated data includes human judgements on semantic relatedness (synonyms, hypernyms and hyponyms) and free association (collected during a large-scale psychological associative experiment). This remains a gold standard annotation and is used by more recently developed tools for Russian distributional modeling: RusVectōrēs (Kutuzov & Kuzmenko 2017), AdaGram (Bartunov et al. 2015) and RDT (Panchenko et al. 2016).

Selectional preference extraction is not so obviously captured by distributional models, and the research in this direction is quite sparse. Jauhar and Hovy (2017) develop a minimally supervised frame lexicon induction method based on a predictive embedding model and an Indian Buffet Process posterior regularizer. Interestingly, they show that the model yields some regularities in frame realizations in addition to hand-crafted data. In Rodríguez-Fernández et al. (2016) a distributional baseline metric is introduced: collocates are evaluated against the difference between an example headword and collocate added to the test headword. The main method proposed in the same paper is based on a linear transformation between a headword and collocate space. The approach is tested on manually classified samples drawn from Macmillan Collocations Dictionary (Rundell 2010). A neural network architecture for selectional preference modeling (also based on distributional hypothesis) is described in Van der Cruys (2014).

Experiments described in Bukia et al. (2016) and Kutuzov et al. (2017) lay the foundations of selectional preferences extraction from Russian corpora, and propose alternative solutions to the problem, but there is much to be done in this field. Bukia et al. (2016) compares two distributional approaches to selectional preference modeling. The first implies semantic similarity calculation based on cosine distance, while the second relies on Mikolov's (2013b) assumption about linguistic regularities captured by distributed word vector models. The clustering of attributive collocations in Kutuzov et al. (2017) is also worth mentioning: in this case the authors employ two-step clustering to group collocations with body parts names into semantic classes. Our study follows the line of previous research in the field generalizing the results obtained on specific test sets.

### 3 Computational Model

#### 3.1 Semantic Relation as a Linear Transformation

As mentioned above, Mikolov and colleagues (Mikolov et al. 2013b) show that regular linguistic relations between two word spaces may be described as a linear transformation on them. The syntagmatic relations may be classified and described in numerous ways (for example, as a lexical function or just semantic collocation class), but in this section we will refer to the relation to be modeled in general.

Our task is to predict possible values of a particular relation for a given target word (base) using training exemplars of this relation. Following Rodríguez-Fernández et al. (2016), we define a base space  $B$  and a collocate space  $C$  produced by a word embedding model. Let  $T$  be a set of collocations  $t_i$  comprising base – collocate pairs  $(b_{t_i}, c_{t_i})$  that represent a given relation  $L$ . Argument matrix  $B_T = [b_{t_1}, \dots, b_{t_n}]$  and collocate matrix  $C_T = [c_{t_1}, \dots, c_{t_n}]$  are made up of corresponding word vectors. Then, given the examples of a particular relation (e.g., a lexical function MAGN: *тяжёлая болезнь* ‘serious illness’, *сильный акцент* ‘heavy accent’, etc.), we should find a transformation which converts a base vector to a collocate vector, for instance, predicts a collocate *бурный* ‘wild’ (MAGN value) for a base *аплодисменты* ‘applause’.

A linear transformation matrix  $\Psi \in R^{B \times C}$  learnt from training set  $T$  satisfies the following:

$$A_T \Psi_T = C_T.$$

Therefore,  $\Psi$  can be approximated using a singular value decomposition to minimize the sum:

$$\sum_{i=1}^{|T|} \|\Psi_T a_{t_i} - c_{t_i}\|^2.$$

Thus, we obtain a transformation matrix for a given relation. Applying it (multiplying it by the base embedding) we obtain a ranked list of potential collocates for given target word and relation.

Rodríguez-Fernández et al. (2016) prove their assumption that base and collocate embeddings should be trained on different corpora. In their work base vectors are obtained from a small corpus containing primarily literal usage (Wikipedia), while collocate vectors are trained on a large corpus full of various figurative meanings. The performance of this model was evaluated on Russian lexical functions data in Enikeeva & Mitrofanova (2017). The authors conduct experiments on the 10 most frequent lexical functions from the Russian National Corpus, fitting linear transformation for each LF and applying it to rank potential collocates. The final scores after heuristic filtering are quite promising (reaching 0.9 in precision).

#### 3.2 Collocation Clustering

We have already noticed that the amount of properly classified syntagmatic relations examples is usually low, hence we need a technique to induce the semantic classes automatically. Following Kutuzov et al. (2017), we can apply clustering algorithms to collocations from corpus represented by the corresponding embeddings. The K-means algorithm is a simple clustering technique producing a known number of classes, and has been successfully applied to various NLP tasks (cf. Berry, Kogan (2010)). This approach may be helpful to fit an existing classification, but in natural language this is not usually the case. Consider, for example, attributive adjective-noun combinations that reflect various relations between an object (noun) and its description (for a detailed study of adjectives categorization see Heyvaert (2010)): the attributive relation in ‘*a sad book*’ is not the same as in ‘*a sad girl*’. Thus, an Affinity Propagation algorithm is used here to infer unknown number of groups. We perform

clustering on stacked base and collocate embeddings and the results of such a simple approach will be discussed in the Evaluation section.

### 3.3 Lexicon Structure

Each lexical entry includes the following information about the semantic compatibility of a headword:

- its regular syntagmatic relations;
- paradigmatic relations: as mentioned above, we provide lists of collocates corresponding to a particular relation instead of strict selectional restrictions;
- idioms;
- peculiarities of word meanings.

Semantic relations are represented as a web of lexical units connected by marked links of several types (synonymy, hyponymy, etc.) and probability counts are assigned to each link. This structure may also be useful to define inherited semantic relations. Consider several nouns belonging to a particular semantic class. “A lion”, “a cat”, “a squirrel” are instances of “an animal”, and some properties of the hypernym may be inherited by a specimen of the class: e.g., the probability of combining “an animal” with a particular set of motion verbs (“to run”, “to jump”) is rather high, and this information may be transferred to the next level (“a lion” also tends to appear with the verbs mentioned above). As opposed to the paradigmatic relations weighting described above, syntagmatic relations extraction is a straightforward application of word embeddings. Our model produces synonyms, hypernyms and hyponym lists by means of cosine similarity applied to word vectors; the results are filtered by heuristics. The semantic network may be enhanced by hierarchical clustering of word representations, taking into account syntactic annotation and representing word context as its syntactic neighbors.

## 4 Data Sources

To the best of our knowledge, the only source of selectional restrictions information for Russian of considerable size is the Framebank project.<sup>2</sup> The main focus of this project is verbal subcategorization frames, so the annotation is verb-oriented. Moreover, the examples of the frame realization may include syntactic constructions and even clauses, which is not of primary interest in our case.

Another option is to use lexical functions (LF) formalism developed within the Meaning↔Text theory (Mel'čuk 1998). At present the inventory of LFs comprises 116 varieties of standard and nonstandard LFs (Apresjan et al. 2007). Russian collocations revealing LF relations are thoroughly described in the *Explanatory Combinatorial Dictionary of Modern Russian* (Zholkovsky & Melchuk 1984). The machine-readable resources containing LF markup for Russian language are quite limited. In our experiments we use SynTagRus Treebank<sup>3</sup> and a verbal combinatory dictionary of Russian abstract nouns.<sup>4</sup>

SynTagRus (Boguslavsky 2014) is a treebank subset of the Russian National Corpus. It consists of more than 28,000 sentences annotated with a dependency parse tree as well as some semantic information including a list of LFs in Meaning↔Text notation and word sense disambiguation. The verbal combinatory dictionary uses its own markup scheme based on LF inventory. About 10,000 collocations are classified in terms of ‘regular abstract meanings’, such as necessity, existence, and action, with additional labels such as phase (start, finish) or semantic class (cognition, perception, etc.).

<sup>2</sup> <https://github.com/olesar/framebank>

<sup>3</sup> <http://ruscorpora.ru/en/corpora-structure.html>

<sup>4</sup> [http://dict.ruslang.ru/abstr\\_noun.php](http://dict.ruslang.ru/abstr_noun.php)

Collocation description within the framework of Meaning $\leftrightarrow$ Text theory rests on the idea that collocations are expected to reveal both the syntagmatic unity and lexical correlation of its parts. Consequently, the majority of LF examples are not free word combinations and reflect a bound usage. Our task of syntactic parser refinement implies that free word combinations should be captured on equal terms with idioms; and we even notice that real texts (especially colloquial speech) show much less restricted usage: word combinations that seem to be abnormal without their real context are in fact acceptable in texts. Finally, some LF attested in SynTagRus annotation are quite rare, so we select only the 20 most frequent types (ignoring special tags marking the action phase: Incep, Fin etc.)

Table 1: Training data sources and size.

Source	Type	Number of classes	Size
SynTagRus treebank	LF examples	20	4958
Verbal combinatory dictionary	LF examples	5 (+ 6 syntactic types)	9729
SynTagRus treebank	syntactically linked collocations	20	75142

In order to train the model on more diverse examples, we extracted word pairs connected by a syntactic link from the SynTagRus treebank. These collocations are classified generally by corresponding syntactic relations, so they were further clustered into more specific semantic classes. The resulting data sources and their sizes are presented in Table 1. Several sources of distributional word representations are freely available for Russian. We have chosen RusVectōrēs<sup>5</sup> (Kutuzov & Kuzmenko 2017) as a primary source: this provide 300-dimensional vectors pre-trained by continuous skip-gram architecture using the word2vec toolkit (Mikolov et al. 2013a) on corpora tagged with Universal Dependencies<sup>6</sup> morphological tags. Word embeddings learnt from Russian Wikipedia were used for modeling headword sense, while the vectors learnt from the Russian National Corpus were applied to possible collocates in order to capture figurative and metaphorical usage. The vocabulary size is 19,5071 and 384,746 lexemes, respectively. Each vector in this model corresponds to a combination of lemma and its part-of-speech tag. More precise results are expected with the AdaGram model (Bartunov et al. 2015), which provides multiple vectors for a word to reflect polysemy; though by now ‘one vector per word’ models seem to capture semantic ambiguity. The linear transformation model and clustering were trained by means of scikit-learn toolkit<sup>7</sup> (v0.19.1).

## 5 Evaluation

### 5.1 Collocate Weighting

Firstly, we describe the collocate ranking process and the evaluation of the results. The training data is heterogenous, and therefore we do not merge it and train the whole model to predict an uninterpretable ‘generalized’ probability, but process and evaluate different datasets separately. Given a list of

<sup>5</sup> <http://rusvectors.org/en/>

<sup>6</sup> <http://universaldependencies.org>

<sup>7</sup> <http://scikit-learn.org/stable/>

examples of a specific relation, a corresponding linear transformation is learnt from it and the result is applied to a set of collocation hypotheses to be weighted. We define several morphosyntactic types of collocations, such as ‘*attributive adjective + noun*’ (1) or ‘*verb + noun as a direct object*’ (2) and the training sets are divided in the same manner (for instance, the examples of lexical functions Magn, AntiMagn, Bon, and AntiBon are seen as corresponding to collocation type (1)).

The lexical function example lists are used as is with the morphosyntactic annotation from SynTagRus. The collocates from corpus are annotated by the pymorphy2<sup>8</sup> morphological analyzer and clustered by means of the K-means algorithm as a simple baseline. This appears to perform quite well on the task of collocation clustering: for example, the attributive noun+adjective collocations are clustered into 20 groups, which can be easily recognized as containing collocations of a particular semantic class: quantitative description (*длинная дорога* ‘long road’), relation to a particular object (*химический опыт* ‘chemical experiment’), characteristic feature description (*верный друг* ‘true friend’), etc. Some groups belong to the same semantic class, and we merged these manually: for example, the ‘*relation to a particular object*’ collocations were found in three automatically clustered groups.

The performance of ranking collocates for a particular headword is then evaluated using precision and mean reciprocal rank (MRR) on this list of top N collocates, as annotated by experts:

$$precision = \frac{tp}{tp + fp}$$

where *tp* is the number of correct collocates on the retrieved list, *fp* is the number of false collocates on the list;

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where *Q* is the top-N list and *rank<sub>i</sub>* is a rank of the first correct collocate. Gold standard annotation is performed by two experts, and the final test set includes only the cases where the annotators agree (about 85%).

Table 2 shows the top-20 evaluation results on several collocation types for 10 headwords using five-fold cross-validation.

A brief comment on LF notation should be made. We used the following lexical functions corresponding to attributive construction:

- *Magn* means ‘very’, ‘to a (very) high degree’, ‘intense(ly)’: *Magn(проблема* ‘problem’) = *серьезная* ‘serious’;
- *AntiMagn* – vice versa.

The following lexical functions are usually represented by verb + direct object in Russian:

- *Oper1* introduces a support verb meaning ‘do’, ‘perform’ something, expressed by noun: *Oper1(поддержка* ‘support’) = *оказывать* ‘(to) lend’;
- *Func1* means that its argument belongs to the subject of corresponding verb: *Func1(власть* ‘authority’) = *принадлежать* ‘belong’;
- In the verbal combinatory dictionary the collocations are annotated with base classes (action, state, etc.) and relation between headword and its argument (subject, direct object, etc.). Here we use only actions with a direct object tag.

8 <https://github.com/kmike/pymorphy2>



Table 2: Top-20 evaluation results on several collocation types for 10 headwords using five-fold cross-validation.

Collocation type	Training data	Precision	MRR
attributive adjective + noun	LF examples (Magn, AntiMagn)	0.84	0.9
attributive adjective + noun	corpus collocation (clustered as 'characteristic feature')	0.89	0.92
attributive adjective + noun	corpus collocations (clustered as 'relation to object')	0.9	0.94
verb + direct object (noun)	LF examples (Oper1, Func1)	0.64	0.73
verb + direct object (noun)	LF examples from verbal combinatory dictionary	0.6	0.66
verb + subject (noun)	LF examples (Func0)	0.42	0.89

A specific relation between verb and subject is represented by the Func0 lexical function: thus means that an event described by a headword takes place: Func0(*снег* 'snow') = *идёт* 'falls'.

## 5.2 Lexical Description

To the best of our knowledge, there are no semantic lexicon examples for Russian to compare with the results. Instead, we compile several entries automatically and assess them manually. For each syntactic type we learn from multiple data sources and then assign each collocate hypothesis the probability from the model in which it is scored the highest. The examples for frequent Russian lexemes are presented in Figures 1 – 3.

The rows in figures represent collocates grouped by the source of training data and possible syntactic construction. The erroneous predictions of the distributional model are marked in red color, green shades correspond to collocational probability: the more intense the color, the higher the score.

We include examples of errors in entries to discuss the matter in detail. Figure 1 represents collocational examples from the 'syntagmatic' part of the entry for a lexeme *проблема* 'problem/issue'. It shows various degrees of collocational strength for attributives with the words *серьезный, резкий, неожиданный* 'serious, sharp, unexpected' being more probable collocates and the words *разный, фактический* 'different, actual' being less probable ones. As for verb + direct object relation, we list acceptable verb examples: *ставить, создавать, поставить, поднимать* 'raise (an issue), cause, raise, raise' along with the erroneously predicted with a high probability collocates *заниматься, интересоваться* 'deal with, be interested in'. In fact, these verbs are usually

	серьезный		
	резкий	SynTagRus LF	
	неожиданный		
	особый		attributive adj+noun
	разный	corpus collocations	
	фактический		
проблема	заниматься		
	интересовать	corpus collocations	
	ставить		verb + direct object
	создавать		
	поставить	SynTagRus LF	
	поднимать		

Figure 1: Part of lexicon entry for a noun проблема ‘problem’.

collocated with the word *проблема* in Russian, but also appear in other types of construction, as in the following examples<sup>9</sup> (1-2):

Мировой опыт показывает, что строители вообще не должны *заниматься проблемами* (Verb + Noun instrumental) подключения к ресурсам. ‘The global experience shows, that the builders should not deal with resource connection problems.’ (1)

Но его *интересовали проблемы* (Verb + Noun nominative), не имевшие отношения к науке вообще. ‘He was interested in issues unrelated to science.’ (2)

Figure 2 briefly shows several fillers of an argument structure of a verb *стоять* ‘stand’. We would like to emphasize the ability of the model to describe ambiguous examples: consider the upper part of the figure representing collocates with a subject role. In Russian, one of the frequent figurative usages of the verb *стоять* ‘stand’ is related to time periods, especially seasons: *Стояла зима* ‘It was winter’; a similar one is attested in collocation with abstract nouns such as *проблема* ‘issue’: *стоящая перед нами проблема* ‘the issue we are facing’. On the other hand, the literal collocations with nouns denoting physical objects are also scored quite high: *дом* ‘house’ is colored with a bright shade in Figure 2.

Figure 3 illustrates an adjective *широкий* ‘wide’ as it reveals the peculiarities of different training sources within one syntactic type. Possible collocates obtained during learning from SynTagRus lexical functions *кругозор*, *возможность*, *развитие* ‘horizon/outlook, possibility, development’ are abstract nouns and typical illustrations of Magn LF. The distributional model also assigns high scores to other nouns with abstract meanings such as *влияние* ‘influence’ and even to *прыжок* ‘jump’, which has a figurative abstract meaning. However, the level of generalization is too high, because some unacceptable collocates (bearing abstract meaning) are ranked in top, as it is the case with *длина*

<sup>9</sup> The examples are taken from Russian National Corpus: <http://ruscorpora.ru/en/search-main.html>

	зима		
	задача	SynTagRus LF	
	дом		
	проблема		verb + subject
	весы	corpus collocations	
	сила		
стоять	очередь		
	намерение	SynTagRus LF	verb + direct object
	час		
	спина		
	лицо	corpus collocations	verb + indirect object
	деньги		

Figure 2: Part of lexicon entry for a verb *стоять* 'stand'.

	кругозор		
	возможность		
	развитие		
	длина	SynTagRus LF	
	влияние		
	прыжок		
широкий	уклон		attributive adj+noun
	аудитория		
	горизонт		
	путь	corpus collocations	
	колея		
	свод		

Figure 3: Part of lexicon entry for an adjective *широкий* 'wide'.

'length'. On the other hand, learning from corpus collocations yields collocates with more literal meaning: аудитория, горизонт, путь, колея 'audience, horizon, way, track' and even less frequent as свод 'vault', as well as several abstract nouns such as уклон 'tendency/direction'.

The described procedure is applied to frequent Russian words belonging to the main parts of speech – 693 nouns, 282 verbs, 221 adjectives and 109 adverbs.

## 6 Conclusion

In the paper we introduce an approach to automatically constructing a semantic lexicon for the Russian language based on distributional word representations. The lexicon is constructed in order to simplify syntactic disambiguation and the fact extraction process. The issues of evaluation are discussed and several examples of retrieved lexicon entries are presented. We hope that the search interface to the lexicon and some visualization features will be made available online soon.

## References

- Apresjan Ju.D. (1995) Selected works. Vol. 1. Lexical Semantics: Synonymic Means of Language. [Izbrannyje trudy. T.1. Leksicheskaja semantika: sinonimicheskiye sredstva jazyka]. Moscow.
- Bartunov S., Kondrashkin D., Osokin A., Vetrov D. (2015) Breaking Sticks and Ambiguities with Adaptive Skip-gram. <https://arxiv.org/abs/1502.07257>
- Berry M. W., Kogan J. (2010) Text Mining: Applications and Theory. Wiley.
- Boguslavsky I. (2014). SynTagRus – a Deeply Annotated Corpus of Russian. In Blumenthal, P., Novakova, I., and Siepmann, D., editors, *Les émotions dans le discours-Emotions in Discourse*, pages 367–380, Peter Lang, Frankfurt am Main, Germany.
- Bojanowski P., Grave E., Joulin A., Mikolov T. (2017) Enriching Word Vectors with Subword Information. In: *Transactions of the Association for Computational Linguistics*. Vol. 5. Pp. 135–146.
- Bukia G. T., Protopopova E. V., Panicheva P. V., Mitrofanova O. A. (2016) Estimating Syntagmatic Association Strength Using Distributional Word Representations. In: *Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference “Dialogue’16”*. Vol. 15. pp. 112–122.
- Enikeeva E., Mitrofanova O. (2017) Russian Collocation Extraction based on Word Embeddings. In: *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference «Dialogue 2017»*. Issue 16. Vol. 1. Moscow. Pp. 52–64.
- Gak V.G. (1998) Language transformations [Jazykovyje preobrazovanija.] Moscow.
- Heyvart F. An outline for a semantic categorisation of adjectives. In Anne Dykstra & Tanneke Schoonheim (eds.). 2010 Proceedings of the XIV EURALEX International Congress. 6-10 July 2010. Leeuwarden/Ljouwert: Fryskke Akademy.
- Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) *Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science*, vol 661. Springer.
- Kutuzov A., Kuzmenko E., Pivovarov L. Clustering of Russian Adjective-Noun Constructions Using Word Embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*.
- Jauher S. K., Hovy E. Embedded Semantic Lexicon Induction with Joint Global and Local Optimization. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (\* SEM 2017)*.
- Mel’čuk I. (1974/1999) Experience in Theories of «Meaning ↔ Text» Linguistic Models [Opyt teorii lingvisticheskikh modelej «Smysl ↔ Tekst»]. Moscow.
- Mel’čuk I., Zholkovskij A. (1984) Explanatory Combinatorial Dictionary of Modern Russian [Tolkovo-kombinatornyj slovar russkogo jazyka]. Vienna.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a) Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of Workshop at ICLR*.
- Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013b) Distributed Representations of Words and Phrases and their Compositionality. In: *Advances in neural information processing systems*. Pp. 3111–3119.

- Panchenko A., Loukachevitch N., Ustalov D., Paperno D., Meyer C., Konstantinova N. (2015) RUSSE: The first workshop on Russian semantic similarity. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference “Dialogue”. Pp. 89-105.
- Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N. and Biemann C. (2016): Human and Machine Judgements about Russian Semantic Relatedness. In Proceedings of the 5th Conference on Analysis of Images, Social Networks, and Texts (AIST’2016). Communications in Computer and Information Science (CCIS). Springer-Verlag Berlin Heidelberg.
- Popov A., Enikeeva E. (2017) Template Search Algorithm for Multiple Syntactic Parses. In: IMS’17, June 21–23, 2017, St. Petersburg, Russia. DOI: <http://dx.doi.org/10.1145/12345.67890> (In print)
- Rodríguez-Fernández S., Anke L., Carlini R., Wanner L. (2016) Semantics-driven recognition of collocations using word embeddings, Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany.
- Rundell, M. (2010) Macmillan Collocations Dictionary, Macmillan.
- Sahlgren M. (2008) The Distributional Hypothesis. From context to meaning. In: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), *Rivista di Linguistica*. Vol. 20. №1. Pp. 33–53.

## Acknowledgements

The reported study is supported by the Russian Fund of Basic Research (RFBR) grants 16-06-00529 “Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques” and 17-29-09159 “Quantitative grammar of Russian prepositional constructions”. We are deeply grateful to SynTagRus team, especially to Leonid L. Iomdin and colleagues from IPPI RAS, for providing access to the Russian treebank and lexical functions annotation. We would also like to thank Olga Mitrofanova and anonymous reviewers for their valuable comments.





# Semantic Classification of Tatar Verbs: Selecting Relevant Parameters

*Alfia Galieva<sup>1</sup>, Ayrat Gatiatullin<sup>1</sup>, Zhanna Vavilova<sup>2</sup>*

*<sup>1</sup>Tatarstan Academy of Sciences, <sup>2</sup>Kazan State Power Engineering University*

*E-mail: amgalieva@gmail.com, agat1972@mail.ru, zhannavavilova@mail.ru*

## Abstract

This paper describes the methodology and current results of the ongoing classification of the Tatar lexicon in the process of developing databases of semantic classes of verbs. Our previous work included a semantic classification of Tatar verbs according to their basic meaning and thematic class. As a result, there have distinguished 59 basic semantic classes, with a semantic tag, or a set of tags, attributed to each of 3,200 verbs.

If the thematic classification is universal and may be applied to any language, the currently developed classification described in this paper is based on the parametric principle and includes a set of morphological, syntactic, semantic, and derivational characteristics that are relevant for Tatar grammatical and semantic systems. In a sense, the work is aimed at creating a Tatar analogue of B. Levin's verb classes, taking into account language-specific features.

In the database, each semantic class, or subclass, is supposed to be provided with a set of admissible diathesis alternations and syntactic descriptions, depicting the verb valency, thematic roles of the arguments and semantic restrictions on them. By now we have created a detailed classification of speech, behavior, sound emissions, weather, emotions, mental states and actions verbs; when selecting the pertinent parameters and verifying their relevance, verbs of other classes were also considered.

**Keywords:** Tatar verb, semantics, corpus, semantic classes

## 1 Introduction

In recent decades, the emergence of search engines and annotated linguistic corpora has significantly expanded the toolkit of linguistic studies. Computer dictionaries, in particular, are of crucial importance in natural language processing which is aimed at data interpretation. Such lexicographic resources that provide researchers with examples from existing texts are being developed to help us obtain the latest information on semantics and the distribution of words of different parts of speech, on the models of lexical government, on variations of lexical units of different classes and on other important aspects.

Developing a semantic classification of the vocabulary of any low-resource language is a difficult and time-consuming task, when the main challenge is a deficiency of appropriate lexicographic resources, which results in the necessity to process “raw” linguistic material. This paper describes an approach to developing databases of semantic classes of Tatar verbs. Tatar, which falls into this category of low-resource languages, belongs to the Turkic family and has a rich agglutinative morphological system. Among other parts of speech, the Tatar verb is one of the most complicated, semantically intricate and grammatically sophisticated categories, which is distinguished by a multiplicity of senses, forms and requirements to the argument structure.

Generally speaking, verbs constitute a nucleus of lexical and grammatical systems of any language. The semantic structure of any verb is usually a complex of ontological and relational meaning components, which may find a formal expression on different levels of the linguistic structure. Their complicated semantic organization requires an integrated approach to their classification, a work that acquired a new élan with the rise of computational linguistics in the recent decades.

The classification which is currently being developed by the authors of the paper, relies upon the results of our previous work of selecting primary classes of Tatar verbs according to their basic meaning and thematic properties. That preliminary classification was carried out with the purpose of developing a corpus semantic dictionary, basing on the available explanatory dictionaries of the Tatar language (1977-1981; 2005), bilingual Russian-Tatar dictionaries, and the data from the Tatar National Corpus (2018). As a result, 59 basic semantic (ontological) classes (such as movement verbs, speech verbs, etc.) have been distinguished, marked by two types of tags: constructional (categorical) and semantic (thematic). Constructional components of meaning are identical for all semantic classes and subclasses. The semantic annotation tag system is currently being developed for the Tatar National Corpus, and 3,200 Tatar verbs have already been streamed into semantic classes. The peculiarity of this work is that semantic classes were defined exclusively on a thematic basis, without taking into account individual grammatical behavior or syntactic alternations of verbs, whereupon the resulting classes include items with different structural and syntactic properties (Galieva & Nevzorova 2016).

Thus our next step is to achieve syntactic and semantic coherence among members of classes by refining the lexical material and separating individual subclasses within basic classes. A crucial point in this work is determining a set of relevant grammatical and semantic language-specific parameters to refine the previously defined classes and to distribute verbs with similar syntactic behavior into subclasses. In a way the project is aimed at creating a Tatar analogue of B. Levin's verb classes (Levin 1993), which would focus on examining the distribution of syntactic frames of a verb in order to establish its class – in our case taking into account the language-specific features of Tatar verbs.

## 2 Related Work

Due to corpus studies, it can be argued that present-day English is rather well provided with semantic verb classifications, the most famous being WordNet (Miller 1995; Fellbaum 1998; Vossen 2002), VerbNet (Kipper et al. 2006; Kipper et al. 2008), FrameNet (Fillmore et al. 2004; Boas 2009) and some others. However, even in the English-speaking domain, experiments in search for the best features for verb classifications proceed. For instance, a research also based on Levin-style verb classification (Li & Brew 2008) is aimed at discovering the optimum combination of syntactic and lexical features for verb classification, a method which has yet to be elaborated.

The methodology of semantic classification of verbs is thoroughly examined in a case study on Italian verb features (Lenci 2014). The author juxtaposes two basic approaches towards classifying verbs: the ontology-based paradigm (like FrameNet) which considers the extra-linguistic situation where the verb's meaning unfolds, and the distribution-based approach (like the above-mentioned Levin's verb classification) which is focused on the verb's linguistic behavior and thus provides a more objective and linguistically relevant methodological framework. The distributional perspective offers a powerful instrument for studying the verb's behavior; taking this perspective, the methods of computational linguistics applied to the study of large-scale corpora are invaluable, as they allow linguists to obtain a plethora of evidence about verbal distributions (Lenci 2014).

Automatic acquisition of information is extremely helpful in compiling a distributional profile of a verb, which would embrace a set of its distributional properties. However, unlike English and other high-resource languages, this is not often the case with languages that are spoken by minor communities within states with a different dominant language. Thus the available corpora of the Turkic languages spoken on the territory of the Russian Federation are not yet provided with any system of semantic annotation (the Tatar National Corpus (2018), the Crimean Tatar Corpus (2018), the Bashkir Corpus (2018), the Tuvan Corpus (2018), the Yakut Corpus (2018), etc.). Its development for a number of corpora is currently underway: the electronic Khakass-Russian lexical database is being provided with an inventory of semantic tags based both on paradigmatic and syntagmatic characteristics of the word forms (Dybo et al. 2015); the corpus of the Tuvan language is also being equipped with the system of semantic annotation (Oorzhak & Khertek 2015).

In the framework of our project, an electronic database of Tatar verbs is being compiled to obtain the distributional profiles of verbal classes. It is supposed to be used in the system of semantic annotation of the Tatar National Corpus, as well as for textual analysis, information retrieval, or machine translation. The following section will be devoted to the approach towards classifying verbs for this project.

### 3 Selecting Relevant Parameters for Classification

The currently developed classification is based on the parametric principle and includes a set of morphological, syntactic, semantic and derivational characteristics which were considered relevant for Tatar grammatical and semantic systems and which allow distinguishing various semantic groups of verbs. The new classification is based on the following parameters of verbal lexemes:

- thematic features, linked with the verb's thematic class, which allows us to mark up the verb's denotation sphere;
- derivational features, related to the verb's derivation pattern (grammatical class of the stem, derivational meaning of the verb forming affix);
- grammatical features, linked with the valency changing operations of voice affixes (possibility of producing grammatical voice derivatives and particular meanings of voice forms);
- syntactic features, related to the allowable predicate-argument structure and thematic roles of arguments.

Tatar is an agglutinative language characterized by a regular morphology, and the derivational structure of a verb (taking into account its stem's grammatical and semantic class and the verb-forming affix) predicts in many respects the verb's basic semantic and grammatical properties. For example, stems referring to tools join the regular verb-forming affix *-la/-lä* and produce transitive verbs with the basic derivational meaning 'to operate with an instrument named by the noun stem' (Example 1). Adjectives may join the *-lan /-län* affix and produce intransitive inchoative verbs (Example 2). So in many cases verbs of the same derivational structure are characterized by similar grammatical properties and basic syntactic behavior.

- (1) *Pıçkı* 'saw' – *pıçkılaw* 'to saw'; *boraw* 'drill' – *borawlaw* 'to drill'; *ütük* 'iron' – *ütüklaw* 'to iron'.
- (2) *Matur* 'beautiful' – *maturlanu* 'to become beautiful'; *yäşel* 'green' – *yäşellänü* 'to become green'; *turı* 'straight' – *turılanu* 'to become straight, to straighten'.

Another significant feature is valency changing operations of voice affixes – the possibility to produce grammatical voice derivatives and particular meanings of voice forms. The Tatar verb has five grammatical voices (basic, passive, causative, reflexive, and reciprocal). Voice affixes are joined in

a strict order and modify the verb's meaning in a direction which depends on its semantic structure. For example, the reciprocal affix is ambiguous and may accept cooperative (associative), assistive or reciprocal meaning; when joined to verbs of different types, it modifies their meaning in different ways. Thus, with labor verbs the reciprocal affix actualizes the assistive meaning, with emotion verbs – the cooperative (associative) meaning, etc. (Example 3). So joining voice affixes is restricted by the verb's semantics, and the fact of producing particular meanings, in combination with other characteristics, serve as a marker for distinguishing verb classes and subclasses.

- (3) *Kuanu* 'to rejoice' – *kuanıſu* 'to rejoice together' (the reciprocal affix -*ſu* actualizing the cooperative meaning of the emotion verb);  
*kazu* 'to dig' – *kazıſu* 'to help somebody dig' (the reciprocal affix -*ſu* actualizing the assistive meaning of the labor verb);  
*tayanu* 'to lean' – *tayanıſu* 'to lean on each other' (the reciprocal affix -*ſu* actualizing the reciprocal meaning of the contact verb).

The predicate-argument structure and its surface realization is a very important classification parameter. The forms of verb control and the thematic roles of arguments are to a large extent determined by peculiarities of verb meaning. So mapping the verb's argument structure is an important step in distinguishing verbs of certain semantic types.

The combination of parameters may be represented in the form of a grid (see Table 1). For example, Tatar fear verbs (*kurku* 'to fear, to be afraid', *örkü* 'to experience a sudden strong fear', *şölläw* 'to experience a slight fear', *şürläw* 'to experience a slight fear') are all non-derivative and intransitive, and may join the reciprocal affix (with the cooperative meaning) and the causative affix (with the factitive causative meaning), but cannot join the passive affix. The source (causer) of fear is marked with the ablative affix (Examples 4-6).

Table 1. Basic distribution parameters for Tatar fear verbs

Examples of verbs	Semantic class	Derivational structure	Grammatical forms of main arguments		Joining voice affixes		
			Agent	Causer	Passive	Reciprocal	Causative
<i>kurku</i>	Emotion, fear	Non-derivative	NOM	ABL	-	+	+
<i>örkü</i>	Emotion, fear	Non-derivative	NOM	ABL	-	+	+
<i>şölläw</i>	Emotion, fear	Non-derivative	NOM	ABL	-	+	+
<i>şürläw</i>	Emotion, fear	Non-derivative	NOM	ABL	-	+	+

- (4) Bala büredän kurka.  
 Child wolf-ABL to be afraid of-PRES  
 'The child is afraid of the wolf'.
- (5) Büre balanı kurkıta.  
 Wolf child-ACC be afraid of-CAUS, PRES  
 'The wolf scares the child' (the verb of fear with the causative affix).
- (6) Balalar bik kurkıſalar.  
 Child-PL very be afraid of-COOP, PRES  
 'Children are scared (together)' (the verb of fear with the cooperative affix).

Such basic semantic and formal characteristics of verbs denoting fear are sufficient grounds for creating a separate subclass within the thematic class of emotions. The developed parameters of classification are used to represent Tatar verbs in a special database which is presented in the next paragraph.



## 4 Database of Tatar Verbs

The database of Tatar verbs is implemented by means of the Microsoft SQL Server database management system and is filled manually after carrying out a careful analysis of linguistic data. Figure 1 illustrates the structure of the database, which consists of eight interconnected tables, including

- a list of verbs' semantic classes (subclasses) which consists of basic forms of verbs provided with semantic tags;
- lists of possible voice derivatives (causative, cooperative, reflexive verbs) provided with semantic tags;
- expertly selected examples of verbs' usage provided with descriptions of possible surface realizations of the argument structure (the grammatical forms and thematic roles of the arguments required, and the semantic restrictions on them);
- a brief description of the typical derivational structure of each semantic subclass member with the indication of the possibility of producing voice derivatives;
- a list of semantic tags.

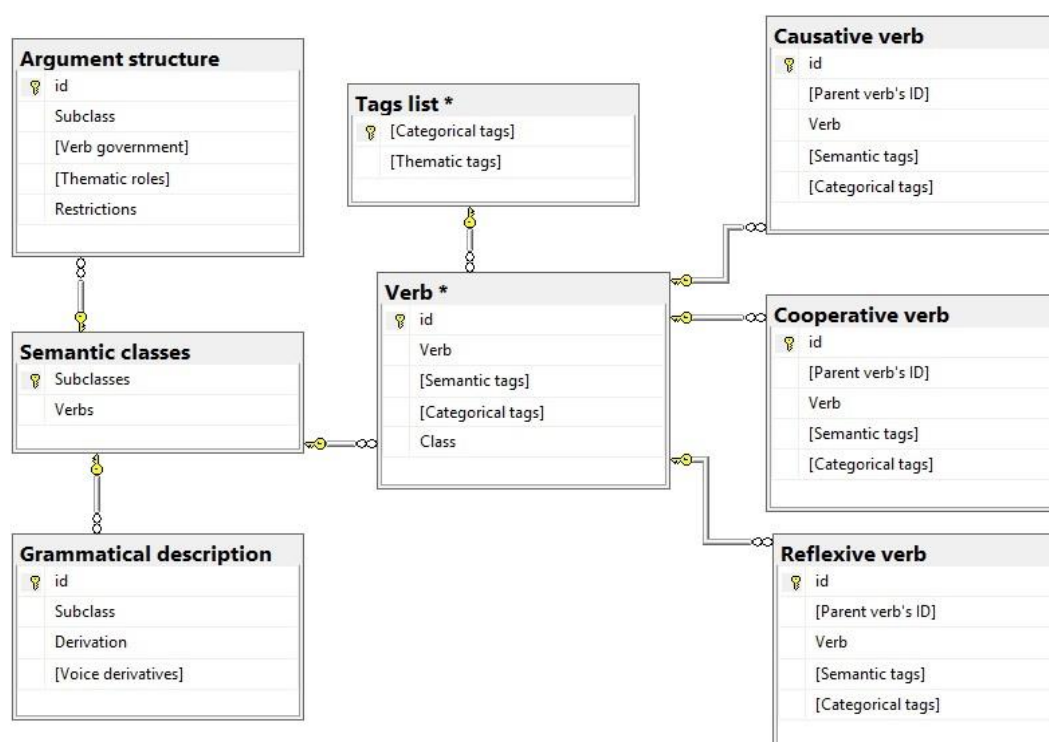


Figure 1: The structure of the database of Tatar verbs.

The database search system provides search for individual verbs by applying semantic tags, represents their semantic class and subclass, as well as enables users to retrieve the list of verbs related to the same subclass. Besides, the database stores information about derivation patterns and possible voice forms. For each subclass there has been chosen a prototypical representative which is provided with typical examples of its usage, information on its syntactic environment and thematic classes of arguments. The information about the typical arguments is taken from contexts of the Tatar National Corpus (2018). Table 2 presents a part of the database table describing the syntactic environment of Tatar fear verbs (whose list is given in Table 1), with the verb *kurku* 'to fear, to be afraid of' as the prototypical verb of the whole fear verbs subclass.

Table 2. Syntactic environment of the verb *kurku* ‘to fear, to be afraid of’

1	<i>Et büredän kurka.</i> ‘The dog is afraid of the wolf’.		
	<b>Verb government</b>	<b>Thematic role of the argument</b>	<b>Semantic constraints to the argument</b>
	N(NOM)	subject of the emotional state	Living being / creature
	N(ABL)	source/cause of the emotional state	
2	<i>Bala uramga ıığarga kurka.</i> ‘The child is afraid of going outside’.		
	<b>Verb government</b>	<b>Thematic role of the argument</b>	<b>Semantic constraints to the argument</b>
	N(NOM)	subject of the emotional state	Living being / creature
	INF_1	cause of the emotional state	
3	<i>Xatın-kız üz balası öçen kurka.</i> ‘The woman fears for her child’.		
	<b>Verb government</b>	<b>Thematic role of the argument</b>	<b>Semantic constraints to the argument</b>
	N(NOM)	subject of the emotional state	Living being / creature
	N(NOM) + PSP <i>öçen</i> ‘for’	cause of the emotional state	

Therefore each subclass contains a list of related verbs and is provided with a set of linguistic descriptions depicting the verb’s derivational structure, possible voice derivatives and admissible surface realizations of the argument structure. The items of the same thematic class with a similar meaning, derivation pattern and syntactic behavior fall into the same subclass, while synonyms with different argument structures fall into different subclasses.

In the current version of the database of Tatar verbs words denoting speech, behaviour, sound emissions, weather, mental states and actions, are presented (see Table 3). When selecting the pertinent parameters and verifying their relevance, verbs of other semantic classes were also considered. The study of verbs of other semantic classes on the basis of the developed description model is in prospect, which will considerably extend the scrutinized lexical material.

Table 3. Distribution of verbs and subclasses within thematic classes.

Thematic class	Number of verbs	Number of subclasses	Semantic tag	Examples
Emotion verbs	234	31	t:psych:emot	<i>yılaw</i> ‘to cry’ <i>moıayı</i> ‘to sorrow’
Speech verbs	157	17	t:speech	<i>söyläw</i> ‘to tell’ <i>maktaw</i> ‘to praise’
Behaviour verbs	233	9	t:behav	<i>aldaw</i> ‘to deceive’ <i>maymillany</i> ‘to ape’
Mental verbs	119	11	t:ment	<i>aılaw</i> ‘to understand’ <i>kartsınu</i> ‘to consider somebody too old’
Sound emission verbs	230	2	t:sound	<i>ulaw</i> ‘to howl’ <i>miyawlaw</i> ‘to meow’
Weather verbs	22	3	t:weather	<i>buranlaw</i> ‘to storm’ (of the weather) <i>bolıtlaw</i> ‘to cloud’ (of the sky)
Total	995	73		

## 5 Conclusion

The presented classification of Tatar verbs is based on the parametric principle, involving a set of semantic, morphological and syntactic characteristics which were considered relevant for Tatar grammatical and semantic systems. This allows us to mark up the denotation sphere of the verb and to consider the grammatical class of the stem and meaning of the verb forming affix, the possibility of producing voice derivatives and particular meanings of voice forms, as well as the verb's valency and the thematic roles of its arguments.

These criteria for classifying verbs are still being revised. By now we have discovered that verbs of the same subclass share some other properties; for example, they may (or may not) be used in certain types of grammaticalised converb constructions. So when adding new items and classes (subclasses) we are planning to update the database with their descriptions considering these new parameters.

The database of Tatar verbs is developed for both professional linguists and Tatar language learners; it can also be used in text processing systems. At present it has a local intranet version; in the future we are planning to develop an online database under a CC-BY-SA copyright license. Currently the database uses a Russian interface and provides linguistic descriptions in Russian. The issue of translating it into English is to be addressed in the nearest future; in particular, with the purpose of extending the field of scientific cooperation in linguistics, it is being planned to develop an English interface and to provide linguistic information on verbs and semantic classes in English.

## 6 Abbreviations

ABL – Ablative, ACC - Accusative, CAUS - Causative, COOP - Cooperative, INF\_1 – Infinitive 1, N – noun, NOM – Nominative, PL - Plural, PRES - Present, PSP – Postposition.

## References

- Bibliographical Bashkir Corpus*. Accessed at: <http://mfbl.ru/bashkorp/korpus> [29/03/2018].
- Boas, H. C. (ed.) (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Walter de Gruyter.
- Crimean Tatar Corpus*. Accessed at: <http://korpus.juls.savba.sk/QIRIM/#id9> [29/03/2018].
- Dybo, A., Sheymovich, A. & Krylov, S. (2015). Some Possibilities of Semantic and Etymological Tagging of Corpora for Turkic Languages. In *Proceedings of the International Conference "Turkic Languages Processing" (TurkLang-2015)*, 17-19 September 2015. Kazan, pp. 304-327.
- Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. Cambridge, Mass: MIT Press.
- Fillmore, C. J., Baker, C. F. & Sato, H. (2004). FrameNet as a "Net". In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC2004)*, 26-28 May 2004. Lisbon, vol. 4, pp. 1091-1094.
- Galieva, A. & Nevzorova, O. (2016). Semantic Annotation of Verbs for the Tatar Corpus. In *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. 6-10 September, 2016. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 340-347.
- Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, 24-26 May 2006. Genoa, Italy, pp. 1027-1032.
- Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A Large-Scale Classification of English Verbs. In *Language Resources and Evaluation*, 42(1), pp. 21-40.
- Lenci, A. (2014) Carving Verb Classes from Corpora in Word Classes: Nature, Typology and Representations. In R. Simone, F. Masini (eds.) *Word Classes: Nature, typology and representations (Current Issues in Linguistic Theory 332)*. John Benjamins Publishing, pp. 17-36.

- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Li, J. & Brew, C. (2008). Which Are the Best Features for Automatic Verb Classification. In *Computational Linguistics: Human Language Technologies. Proceedings of the Conference*. 15-20 June 2008. Columbus: The Ohio State University, pp. 434-442.
- Miller, G. A. (1995). WordNet: A Lexical Database for English. In *Communications of the ACM*, 38 (11), pp. 39-41.
- Oorzhak, B. & Khertek, A. (2015). Development of Semantic Markup for the Corpus of Tuvan Language. In *Proceedings of the International Conference "Turkic Languages Processing" (TurkLang-2015)*, 17-19 September 2015. Kazan, pp. 351 – 373.
- Tatar National Corpus*. Accessed at: <http://tugantel.tatar/?lang=en> [29/03/2018].
- Tatar Explanatory Dictionary*. In 3 volumes (1977-1981). Kazan (In Tatar).
- Tatar Explanatory Dictionary*. In 1 volume (2005). Kazan (In Tatar).
- Tuvan Corpus*. Accessed at: <http://www.tuvancorpus.ru> [29/03/2018].
- Vossen, P. (ed.) (2002). *EuroWordNet General Document*. Version 3. Accessed at: <http://vossen.info/docs/2002/EWNGeneral.pdf> [29/03/2018].
- Yakut Corpus*. Accessed at: <http://adictsakha.nsu.ru/corpora/corp> [29/03/2018].

# Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings

*Nicolai Hartvig Sørensen, Sanni Nimb*

*Society for Danish Language and Literature*

*E-mail: nhs@dsl.dk, sn@dsl.dk*

## Abstract

We describe the use of a tool that assists lexicographers with extending the lexical coverage of an online Danish dictionary. The tool is based on a word embedding model (word2vec) trained on a large Danish corpus, and it presents semantically related lemmas already included in the dictionary and, importantly, their definitions. Furthermore, lemma candidates, i.e. words from the corpus which are *not* included in the dictionary, are presented in the tool, supplemented by information on corpus frequency. The tool thereby facilitates the lemma selection as well as the process of writing consistent definitions across synonyms and near synonyms. We discuss the shortcomings of the tool and the semantic model when it comes to identifying words similar in meaning from different genres and registers. We also look closer into whether it does in fact benefit the dictionary-making process or not by studying a number of previously edited words, including their synonyms, and comparing them with the output data from the tool.

**Keywords:** lemma selection, word embedding, word2vec, semantic similarity, dictionary-making process

## 1 Background

When compiling a new dictionary, it is a well-known and well-proven strategy to edit the lemmas in a semantic order rather than strictly alphabetically in order to ensure consistency of definitions among semantically similar words (cf. e.g. Lorentzen 2004). This strategy was adopted when the printed version of the dictionary we are working on, *Den Danske Ordbog* (*The Danish Dictionary*, henceforth DDO), was compiled 1992-2005. However, when the task consists of augmenting the lemmata of an existing dictionary by adding either completely new or formerly neglected lemmas, as is the case with the current DDO project, it is less obvious how to carry out the process. How do you in a fast and consistent way compare new lemma candidates to already described lemmas within the same semantic field in order to ensure the consistency of the definitions? And additionally, how do you obtain the extra advantages of such a semantically driven workflow in order to identify other lemma candidates in the same field? In this paper, we describe the use of a lexicographic tool that we have developed based on a word embedding model in order to present a number of words that are most semantically related to the lemma that the lexicographer is describing, see Figure 1.

Using distributional methods for suggesting semantically similar words is not new in the field of lexicography. Sketch Engine (Kilgariff & Tugwell 2001), for instance, includes an “automatic thesaurus” for several languages (e.g. for Slovenian (Krek & Kilgariff 2006)). But to the best of our knowledge, this is the first time it is being used specifically to assist in the editing process to compare the entry with previously written entries.



## 2 The Word2Dict Tool

“Word embeddings” is a group of techniques used to investigate the semantic similarity of words by mapping the context of the words in a large corpus into multidimensional vectors. In these models, each type in the corpus is represented by a vector in vector space, and the similarity of the words is based solely on the similarity of the context in which they occur. No morphological or syntactical information is used. Word2vec (Mikolov et al. 2013a; Mikolov et al. 2013b) is a very efficient word algorithm that produces word embeddings.

We used the version of the word2vec algorithm implemented in the Gensim Python package (cf. Řehůřek 2017; Řehůřek & Sojka 2010) to train a model based on the Danish corpus used by the lexicographers of DDO (cf. Asmussen 2017). The corpus included at the time of the training roughly 920 million running words, mainly newswire, but also, material from magazines, transcripts from the Danish Parliament, and some fiction, among others, spanning the years 1982 to 2017. We trained the model with 500 features, a window size of five, a minimum occurrence of five for all types, and used the default choice of CBOW instead of skip-gram. The corpus included 6.3 million types, five million of which occurred less than five times. The training took roughly 18 hours on a 2017 MacBook Pro.

### Word2Dict

Search

**Most similar words:**

LIMEJUICE	-	75	NOT IN DICTIONARY
APPELSINJUICE	-	569	
APPELSINSAFT	-	628	
ANGOSTURA	-	64	
VERMOUTH	-	395	
ÆBLEJUICE	-	309	
ÆBLECIDER	-	123	
KOKOSMÆLK	-	562	
CACHACA	-	92	NOT IN DICTIONARY
FRUGTSAFT	-	231	

SHOW MORE

**anasasjuice [Not in dictionary]**

**limejuice (0.75) [Not in dictionary]**

**appelsinjuice (0.72) [kommenteret]**

- 1 juice fremstillet af pressede appelsiner
- 2 karton, flaske e.l. med denne type juice

**appelsinsaft (0.72) [publiceret]**

- 1 ren, orangegul saft fra appelsiner
- 2 drik fremstillet af koncentreret appelsinsaft, evt. med tilsætningsstoffer

**angostura (0.72) [publiceret]**

- 1 bitter fremstillet på ekstrakt fra barken af visse sydamerikanske træer

**vermouth (0.72) [publiceret]**

- 1 sød el. tør hedvin fremstillet af vin der er krydret med aromatiske urter,

**æblejuice (0.72) [kommenteret]**

- 1 juice fremstillet af (koncentreret saft fra) pressede æbler

**æblecider (0.71) [kommenteret]**

- 1 cider af gæret æblesaft

Figure 1: A search for *anasasjuice* (‘pineapple juice’). To the left (the top-half of the interface) we see the most similar words according to the context in which they appear in a corpus. Frequency counts for each word and whether or not the word is included in DDO is also displayed. The frequency counts are color-coded for quicker visual decoding: the darker the color, the higher the frequency. To the right (the bottom-half of the interface), definitions of the words already in the dictionary are shown, as well as their editorial status (e.g. “publiceret” (‘published’)) and the similarity score from the model (e.g. “0.75”, “0.71” – 1.0 equals identical).

We then implemented the Word2Dict tool as a CherryPy web app, by modifying the open-source Gensim HTTP service (Řehůřek 2014) and included calls to a simple API for accessing DDO data. This means that the lexicographic tool combines the word2vec model with on the fly lookups in DDO.

This combination makes it possible to immediately determine whether or not the words returned by the word2vec model are already included in the dictionary – words which are not included are marked by “not in dictionary”, as shown in Figure 1 (left). The higher up the word is, the more similar to the query word (in this case *ananasjuice*). For the words that are included, their definitions are looked up and displayed immediately in the bottom half of the interface, see Figure 1 (right). This allows the lexicographer to be inspired by – and steal from – existing definitions of semantically similar words, and thus ensure greater consistency in the descriptions within semantic domains. The second most similar word to *ananasjuice* in the model is *appelsinjuice* (‘orange juice’), and DDO describes this word with two main senses (the juice itself, and a container with that juice). In this case, the editor may conclude that the two definitions of *appelsinjuice* might serve as an excellent starting point when editing the new word *ananasjuice*.

The Word2Dict tool also makes it easy to consider the words not yet included in the dictionary as new lemma candidates. Figure 2 demonstrates a search for *mediaopmærksomhed* (‘media attention’) with several similar words not yet included in the dictionary.

The DDO project uses corpus frequency as an important indicator of relevance. Therefore, information on corpus frequency is also presented in the tool, supplying the editor with important knowledge when deciding whether to accept or discard the lemma candidate. In this case the tool reveals that *mediabevågenhed* (‘media attention’) and *medieinteresse* (‘media attention’) are rather frequent words, while *pressebevågenhed* (‘attention from the press’) and *presseopmærksomhed* (‘attention from the press’) are less so. In addition to this, the similarity score suggests that both *mediabevågenhed* (0.82) and *medieinteresse* (0.76) might be considered as good synonyms of the query word, which they indeed are – as indicated by the English glosses above.

## Word2Dict

**Most similar words:**

MEDIEBEVÅGENHED	-	1051	NOT IN DICTIONARY
MEDIEINTERESSE	-	252	NOT IN DICTIONARY
MEDIEOMTALE	-	1471	
MEDIEDÆKNING	-	1143	
PRESSEDÆKNING	-	524	
PRESSEBEVÅGENHED	-	104	NOT IN DICTIONARY
OPMÆRKSOMHED	-	51323	
VIRAK	-	737	
PRESSEOPMÆRKSOMHED	-	39	NOT IN DICTIONARY
PRESSEOMTALE	-	1003	

SHOW MORE

**mediaopmærksomhed [Not in dictionary]**

**mediebevågenhed (0.82) [Not in dictionary]**

**medieinteresse (0.76) [Not in dictionary]**

**medieomtale (0.73) [publiceret]**  
1 omtale i aviser, radio, tv m.m.

**mediedækning (0.72) [publiceret]**  
1 (omfanget af) pressens behandling af en begivenhed, en sag e.l.

**pressedækning (0.70) [publiceret]**  
1 (omfanget af) pressens behandling af en begivenhed, en sag e.l.

**pressebevågenhed (0.69) [Not in dictionary]**

**opmærksomhed (0.68) [publiceret]**  
1 det at have sine sanser el. sin opfattelsesevne rettet mod én bestemt ting  
1.1 det at vise interesse og omsorg for nogen  
1.2 gave som gives som tegn på interesse el. påskønnelse

**virak (0.68) [publiceret]**  
1 stor hyldest; overstrømmende, næsten overdreven ros  
2 uro; ballade; postyr

**presseopmærksomhed (0.64) [Not in dictionary]**

**presseomtale (0.64) [publiceret]**  
1 omtale i aviser, radio, tv m.m.

Figure 2: The most similar words to *mediaopmærksomhed* (‘media attention’), top-half of the interface to the left, bottom-half to the right.

We believe that the Word2Dict have the potential of speeding up the editing process without sacrificing the quality of the work. It has, however, proven difficult to reliably measure the speed-up, as the Word2Dict tool also reveals hidden inconsistencies in the existing entries that need to be fixed in order to decide on the semantic structure of the new word. It seems unjust to blame the tool for earlier shortcomings. However, we make an informal estimate of the benefits of editing the dictionary when also including relevant lemma candidates revealed by Word2Dict in section 3.3 below.

A more basic question therefore becomes whether or not the model returns relevant words at all. Bearing in mind that the word2vec model knows nothing about the morphology of the words in the corpus or syntax of the sentences, only the word forms in the context, does it return words that are relevant to lexicographic work? Does it reflect the mind of a lexicographer sufficiently well? This will be tested in section 3.

### 3 Testing Word2Dict

Since 2015 the DDO dictionary has been extended with more than 8,400 lemmas. Many of these are compounds (especially noun compounds) which were already occurring with a mid-to-low frequency in Danish corpora texts from the 1980s and 1990s. However, the sense descriptions were left out of the dictionary due to space limits when the first, printed edition of DDO was compiled. Other lemmas that have been included in recent years are new words in the Danish language since 1990, e.g. English loanwords. All the newly included lemmas, most of which are nouns, are represented with a mid-to-low corpus frequency in the large corpus of approximately one billion words mainly consisting of newswire, which is today used for lemma selection and corpus investigations in the dictionary-making process.

Almost 4,000 of the 8,400 lemmas contain information on “related words”, i.e. on synonyms, antonyms and “see also” words (presented to the user with the label *Se også* (‘see also’) – a large part of these are near-synonyms) in separate fields. These have been manually selected by the editor without the use of any tool, e.g. based on subjective judgment and, maybe more importantly, to a high degree by consulting the lexical description of the approximately 65,000 lemmas which were already part of the printed edition of DDO (some of them being synonyms of these). In order to find out whether the Word2Dict tool supplies the editor with the same semantically related words that he or she would include manually, based on a subjective judgment, we compared the output data of the tool with a randomly selected number of the newly included DDO lemmas containing related words.

We studied 100 randomly chosen lemmas of 1,332 lemmas with related words which were edited and included in DDO in 2016. Of the 1,332 lemmas, 1,160 (87%) are nouns, 132 (10%) are adjectives, 36 (3%) verbs and only three are adverbs, and the study we present therefore mainly focuses on nouns. We wanted to see whether the related word (or words) included in the dictionary entry by the editor was in fact also part of the output of the Word2Dict tool. We also wanted to see whether it happened to be no. 1 or 2 on the list, considering this to confirm the high quality of the Word2Dict output. Furthermore, we studied whether the output contained highly relevant words that are *not* already mentioned in the dictionary entries, but in our opinion ought to be included and maybe even replace the related words selected by the editor without the use of the tool. First, we exemplify the investigation method by presenting two quite different results:

1. When we studied the noun lemma *fejlinformation* (‘erroneous information’) in DDO, we found the noun synonym *misinformation* (‘misinformation’). Word2Dict gives as output of *fejlinformation* the following list of nouns: *vildledning* (‘deception’), *misinformation* (‘misinformation’), *forhaling* (‘delay’), *fordrejning* (‘misrepresentation’), *fejlfortolkning* (‘misrepresentation’), *obstruction*

(‘obstruction’), *ignoring* (‘ignoring’), *mistænkeliggørelse* (‘casting suspicion on’), *manipulation* (‘manipulation’), *fortielse* (‘concealment’). In this case the Word2Dict output matches the manual selection very well, since the dictionary synonym *misinformation* is the second word on the list. Furthermore, the list supplies us with several other very relevant nouns: *vildledning*, *fordrejning*, *manipulation*, *fortielse*, of which *fordrejning* is not yet included in DDO. In this case the tool would have been very useful in the editing process of the lemma, having facilitated a faster and more consistent dictionary editing flow, since the editing of several near-synonymous noun lemmas could have been carried out at the same time, assuring relevant references between the words.

2. In contrast to the example above, in the case of the noun *mediebranche* (‘media business’) Word2Dict does not supply us with any words of interest. The DDO editor has chosen to present the noun synonym *medieverden* (‘the media world’), but we got as output from the Word2Dict tool only a list of words denoting other types of working branches (i.e. co-hyponyms of the noun *branche* (‘line of business’)), none of which we found relevant to include in the DDO entry.

In Table 1 we present the results of the study.

In 60% of the cases, the output includes the editor’s manually inserted related word, and in 43% it is either no. 1 or 2 in the output list (see Table 1, case 3), and in only 18% of the studied cases did the tool not supply us with any lexicographically relevant related words (see Table 1, case 4). This, we believe, strongly indicates that the Word2Dict tool sufficiently well reflects the mind of a lexicographer.

### 3.1 Cases Where the Tool Did Not Find the Editor’s Related Word From the DDO Entries

In 23% of the 100 studied cases, we find that the output does contain relevant words, however the editor’s choice of synonym is *not* among the words found by the Word2Dict tool (case 2 in Table 1). In most cases this is due to a too low frequency of the synonym in our corpus, which is based on newswire and does not represent spoken language, more specific domains or older language, nor much literary language. In these cases, the DDO synonym is of course very relevant. For example, in the cases of the two old-fashioned, poetic nouns *hjertereven* (‘bosom friend’) described by the editor as a synonym in the entry *bedsteven* (‘best friend’) and *kvindekær* (‘fond of women’) described as a synonym in the entry *pigeglad* (‘fond of girls’), these are not represented in the Word2Dict output. Nor is the specific language noun *veterinær* (‘veterinary’) on the output list of the common noun *dyrlæge* (‘vet’) in DDO, but in the dictionary the two nouns are presented as synonyms of one another. In these latter cases, the Word2Dict tool is not able to supply the editor with relevant words.

### 3.2 Cases Where the Tool Finds Relevant Words Not Discovered by the Editor

In far the most cases (82%, case 4 in Table 1) the list supplies us with more relevant synonyms, near-synonyms and antonyms than the ones that were manually selected by the editor for presentation in the DDO entry. In other words, the tool reveals relevant related words that seem to have been left out by the editors, either on purpose or because they were not aware of them. Most of the related word candidates on our output lists appear at a first glance to be more common words compared to the ones mentioned in DDO. The words have more or less the same meaning but are simply more frequent in the modern DDO-corpus based on newswire of today. We therefore guess that the difference between DDO and Word2Dict in this case is caused by the fact that the editors of the newly included lemmas in DDO first of all have based the sense descriptions and related words on the already described DDO-lemmas from the first, printed edition which was based on older Danish dictionaries and words and their meanings in corpus texts from 1982 to 1992. In this case, the introduction of Word2Dict as a supplementary editorial tool would make sure that newer related words from modern corpora are also presented in the entries.



Table 1: The study of the Word2Dict output when compared to already edited lemmas with related words (i.e. synonyms, antonyms and ‘see also’ words) in DDO

	Comment on output from Word2Dict	100% = 100 lemmas studied	Examples
(1)	No useful output (often due to low corpus frequency of the input lemma)	5%	<i>aktieobligation</i> (‘special kind of stock option’), <i>anvendelsesmulighed</i> (‘application’), <i>normalpris</i> (‘normal price’), <i>originaleksemplar</i> (‘original copy’)
(2)	Output does contain relevant words, however it does not contain the editor’s manually selected related word	23%	<i>amatørcykelrytter</i> (‘amateur racing cyclist’) without the synonym <i>amatørrytter</i> <i>aktualitetsstof</i> (‘news’) without the synonym <i>nyhedsstof</i> <i>bedsteven</i> (‘best friend’) without <i>hjertereven</i> (‘bosom friend’) <i>diktaturstat</i> (‘dictatorship’) without the synonym <i>diktaturland</i> <i>kødfri</i> (‘without meat’) without <i>vegetarisk</i> (‘vegetarian’) <i>letsindighed</i> (‘carelessness’) without the synonym <i>skødesløshed</i>
(3)	Output includes the editor’s manually selected related word within the first 10 words	60% (28% as no. 1 on list, 15% as no. 2)	<i>afterparty</i> (‘afterparty’) contains the synonym <i>efterfest</i> as no. 1 on the list <i>auktionsfirma</i> (‘auction house’) contains the synonym <i>auktionshus</i> as no. 1 on the list <i>blomstringsperiode</i> (‘flowering period’) contains <i>blomstringstid</i> (‘time of flowering’) as no. 2 on the list <i>actionkomedie</i> (‘action comedy’) contains <i>krimikomedi</i> (‘crime comedy’) as no. 2 on the list <i>borgmesterstol</i> (‘mayoralty’) contains the synonym <i>borgmesterpost</i> as no. 1 on the list <i>bundkamp</i> (‘match between relegation candidates’) contains the synonym <i>bundopgør</i> as no. 8 on the list
(4)	Output includes a number of relevant related words which were not (yet) presented by the editor as a related word in the DDO entry	82%	<i>actionkomedie</i> (‘action comedy’): the synonyms <i>thrillerkomedie</i> , <i>dramakomedie</i> , <i>komediedrama</i> <i>bundkamp</i> (‘match between relegation candidates’): the synonyms <i>bundgyser</i> , <i>bundbrag</i> , the antonym <i>topkamp</i> (‘top-of-thr-table clash’) <i>annekering</i> (‘annexation’): the synonyms <i>indlemmelse</i> , <i>invasion</i> , <i>erobring</i> , <i>okkupation</i>
(5)	Output contains lemma candidates among the related words (= words not yet included as a lemma in the dictionary)	32%	<i>actionkomedie</i> (‘action comedy’): the near-synonym <i>thrillerkomedie</i> <i>bundkamp</i> (‘match between relegation candidates’): the synonyms <i>bundgyser</i> , <i>bundbrag</i> <i>amatørcykelrytter</i> (‘amateur racing cyclist’): the synonym <i>motionscyklist</i> <i>annekering</i> (‘annexation’): the synonym <i>indlemmelse</i> <i>bilværksted</i> (‘car repair shop’): the synonym <i>mekanikerværksted</i> <i>diktaturstat</i> (‘dictatorship’): the near-synonym <i>etpartistat</i> (‘one party state’) <i>dystopisk</i> (‘dystopian’): the near-synonym <i>mareridsagtig</i>



In a few cases we even found the output from Word2Dict to be more semantically precise than the synonyms selected by the editors. This is for example the case for *dystopisk* ('dystopian') with the related word *utopisk* ('utopian') in DDO. Word2Dict lists *apokalyptisk* ('apocalyptic'), *mareridsagtig* ('nightmarish'), and *futuristisk* ('futuristic'), all of which we consider more appropriate to describe in the entry of *dystopisk*. Another case is *eliteidrætsudøver* ('elite athlete') with the related word *elitegymnast* ('elite gymnast') in DDO, which is just one of a large selection of more specific hyponyms. We would suggest it to be replaced by synonyms like *topatlet* ('elite athlete'), *topidrætsmand* ('elite athlete'), *elitesportsmand* ('elite athlete'), and *eliteatlet* ('elite athlete'). Also in these cases we find that the dictionary-making process would benefit to a high degree from the Word2Dict tool, not only to ensure consistency but also to avoid aggravating omissions.

### 3.3 Word2Dict Used to Identify Semantically Related Lemma Candidates

In 32% of the studied word examples (case 5 in Table 1) we find lemma candidates among the groups of words which we find semantically related to the headword. This means that in 1/3 of the cases where Word2Dict gives an output, we identify words which not only mean almost the same as the input word, but which are furthermore very good candidates (with a relevant frequency in the corpus) to be described in the dictionary.

By editing words which are semantically related all at once, we are able to reuse and adapt definitions. We estimate that the use of the tool in such cases will speed up the dictionary-making process by at least 10-20% on average, once the editor is confident with using it. It should be taken into consideration that the time benefits vary quite a lot depending on the character of the semantic domain. Those with many relevant lemma candidates (e.g. semantic areas which have not yet been fully described in DDO) allow the editor to greatly speed up the work in contrast to domains where the tool reveals fewer or no relevant lemma candidates.

We also expect the quality of lexical work to improve. By comparing a group of words with related, however not completely identical senses, the editor becomes aware of the subtle differences between the words and is thereby in each case able to formulate a more precise definition. Furthermore, the method facilitates the creation of references between the words.

## 4 Conclusions and Future Work

While others have investigated the use of word embedding techniques for unsupervised identification of synonyms with some success (cf. e.g. Nguyen et al. 2015, Leeuwenberg et al. 2016), our focus is instead on presenting a set of semantically similar words used as input to *manual* lexicographic inspection. Investigating near-synonyms, co-hyponyms and hyponyms, as well as synonyms, might offer the lexicographer a better overview of how the semantic domain previously has been described in the dictionary.

By comparing the manually selected synonyms in the DDO with the output from the word2vec model, our small-scale study has proved that the model is able to supply us with a number of semantically related words from the same register and genre as the query word. Secondly, we have shown that the model to a great extent suggests words to be included in DDO as synonyms, antonyms and "see also" words, which the lexicographer would otherwise easily neglect. This strongly indicates that the already established definitions of DDO words being part of the output list constitute a very good starting point when it comes to defining the new lemmas to be included in the dictionary.

It is worth noting that the model is not able to identify semantically similar words belonging to very different registers, due to the fact that the textual context of such words differs greatly – this is an inherent feature of word embeddings. For this reason, the lexicographer should not rely solely on the Word2Dict tool when selecting synonyms.

In further work we plan to compare the word2vec with other models for Danish (e.g. the models recently made public from the Sketch Engine group, cf. “Models for download”) and to fine-tune the training parameters. We want to study different aspects in order to improve the model, i.e. in order to be able to select the best window size when it comes to use the model for lexicographic work. The work will be carried out in collaboration with the Centre for Language Technology, University of Copenhagen. We also plan to add more features to the tool, for instance making it possible to see not only the definitions but also the synonyms of the already described DDO lemmas returned from the model. This feature would allow the dictionary editor to carry out a consistency check of the synonyms, as well as the definitions of similar words in a very efficient way. Finally, we plan to use the tool in the editing process of a Danish thesaurus published in print 2014 (Nimb et al. 2014), which is currently being extended with more words and expressions and which we plan to publish online in the coming years, depending on funding.

## References

- Asmussen, J. (2017). Hvor kommer ordene fra? In Nyhedsbrev fra Det Danske Sprog- og Litteraturselskab nr. 5, pp. 6-7.
- DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. Accessed at: <http://ordnet.dk/ddo> [28/03/2018].
- Kilgarriff, A. & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proc. ACL workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*, pp. 32-38.
- Krek, S. & Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec, J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*. Ljubljana, Slovenia.
- Leeuwenberg, A., Velab, M., Dehdaribc, J. & van Genabith, J (2016). A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. In *The Prague Bulletin of Mathematical Linguistics, Number 105*, pp. 111–142.
- Lorentzen, H. (2004): The Danish Dictionary at large: Presentation, Problems and Perspectives. In G. Williams & S. Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress*, pp. 285-294.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*. eprint arXiv:1301.3781
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proc. Advances in Neural Information Processing Systems 26*, pp. 3111–3119.
- Models for download. Accessed at: <https://embeddings.sketchengine.co.uk/static/index.html> [01/04/2018].
- Nguyen, N.T.H., Miwa, M., Tsuruoka, Y. & Tojo, S. (2015). Identifying synonymy between relational phrases using word embeddings. In *Journal of Biomedical Informatics*, Volume 56, pp. 94-102.
- Nimb, S., Trap-Jensen, L. & Lorentzen, H. (2014). The Danish Thesaurus: Problems and Perspectives. In *Proceedings of the 16th EURALEX International Congress*, EURAC research, Bolzano, Italy, pp. 191-199.
- Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, pp. 46–50.
- Řehůřek, R. (2013). *Word2vec in Python, Part Two: Optimizing*. Accessed at <https://rare-technologies.com/word2vec-in-python-part-two-optimizing/> [24/11/2017].
- Řehůřek, R. (2014). *Word2vec as an HTTP service*. Accessed at [https://github.com/RaRe-Technologies/w2v\\_server\\_googlenews](https://github.com/RaRe-Technologies/w2v_server_googlenews) [05/10/2017].

# Building a Lexico-Semantic Resource Collaboratively

*Mercedes Huertas-Migueláñez<sup>1</sup>, Natascia Leonardi<sup>2</sup>, Fausto Giunchiglia<sup>1</sup>*

<sup>1</sup>University of Trento, <sup>2</sup>University of Macerata

E-mail: [mdlm.huertas@unitn.it](mailto:mdlm.huertas@unitn.it), [natascia.leonardi@unimc.it](mailto:natascia.leonardi@unimc.it), [fausto@disi.unitn.it](mailto:fausto@disi.unitn.it)

## Abstract

Multilingual lexico-semantic resources are used in different semantic services, such as meaning extraction or data integration and linking, which are essential for the development of real-world applications. However, their use is hampered by the lack of maintenance and quality control mechanisms over their content. The Universal Knowledge Core (UKC) is a multilingual lexico-semantic resource designed as a multi-layered ontology that has a language-independent semantic layer, the concept core, and a language-specific lexico-semantic layer, the natural language core. In this paper, we focus on expert-based, collaborative workflow for building and maintaining our resource through lexicalization and evaluation of language elements via a dedicated User Interface (UI). We have run a three-month study to analyze the feasibility of the proposed solution. We interviewed participants to obtain a comprehensive vision with respect to different aspects related to the way they interacted with the UI and how the content presented through it was perceived. We concluded that this collaborative experience fostered not only the implementation of a resource, but also an improvement of its functionalities, and, above all, it represented an example of effective knowledge sharing which opened up the way to a network of collaborative intelligence.

**Keywords:** multilingual resource, collaboration, knowledge sharing, user study

## 1 Introduction

Lexico-semantic resources, such as English WordNet (Miller 1995) and the corresponding parallel projects such as GlobalWordNet<sup>2</sup>, are important to guarantee the presence of a language in our information society; for machine understanding related tasks, such as natural language processing (NLP) and machine translation (MT); and for people to learn and understand the lexico-semantic relations among language elements. However, the majority of these resources have some unresolved issues, such as content quality, i.e. typos or wrong translations (Zhang, Ojha & Giunchiglia 2017), or license restrictions, which can hamper their use and maintenance (Bond & Foster 2013).

Various people's contributions have been used to build and maintain linguistic resources. Wiktionary adopted crowdsourcing to build and maintain its content (Meyer & Gurevych 2012), and this allows the collection of data in a fast and cheap manner. However, the quality of the work produced by this method might be undermined by workers who are interested in the number of tasks completed rather than in the quality of the results (Eickhoff & de Vries 2013). Nevertheless, according to Morita and Ishida (2009), collaborative translation produces high-quality results. In order to successfully employ the metaphor of collaboration, we need to design systems that facilitate communication between people and organize them in teams with a range of expertise (Kittur et al. 2013). Furthermore, people should identify themselves with the group they collaborate with and believe that their effort is important for the community (Rashid et al. 2006; Munro 2010).

1 The present study is the result of a close collaboration among the three authors. However, Mercedes Huertas-Migueláñez wrote Sections 1, 4 and 5. Natascia Leonardi wrote Section 3. Fausto Giunchiglia wrote Section 2. The Abstract and conclusions were a collaborative effort of the three authors.

2 <http://globalwordnet.org/> [last accessed 31-3-2018].

Whereas our long-term goal can be found elsewhere (Giunchiglia et al. 2015), in this paper we focus on the evaluation of the preliminary version of a tool to co-construct a high-quality multilingual lexico-semantic resource, the UKC (Giunchiglia, Batsuren & Bella 2017). Initially, we import freely available resources automatically. However, due to the complexity of the vocabularies, we involve experts to refine and maintain what we import. We selected the Italian language as our case study. The results of this preliminary study will be used to improve the current design of a dedicated UI and the collaboration pipeline.

The paper is organized as follows. Section 2 describes the UKC. In Section 3 we describe the Italian LKC. Section 4 presents the design of the UI. Section 5 reports on the study we conducted, and Section 6 concludes the paper.

## 2 The Universal Knowledge Core

The *Universal Knowledge Core* (UKC) is a knowledge base developed at the University of Trento. Just like in WordNet (Miller 1995), a vocabulary consists of *synsets*, *lemmas*, *word forms*, *senses*, and *examples*, which are representations of the sense in use. However, the UKC is different from WordNet, and the parallel projects, in that it features a language independent layer called the *concept core*. The *concept core* includes the lexico-semantic relations and provides mappings of common lexical elements from different languages, contained in the *language core*, to formal concepts (Giunchiglia, Batsuren & Freihat 2018). Every vocabulary is stored in a *Language Knowledge Core* (LKC). An LKC is a working copy of the UKC's concept core restricted to two vocabularies: English and another one. In our case study, we have chosen Italian. In Figure 1, we provide an example to illustrate how the UKC is organized. The English word *bike* has two meanings, as a verb and as a noun. They are represented by two single word synsets and are connected to the corresponding Italian words through their reference concepts. However, in Italian there is no lemma for the verb *to bike*, and therefore it will be represented as a *lexical gap*, which denotes missing lexicalization in a given language.

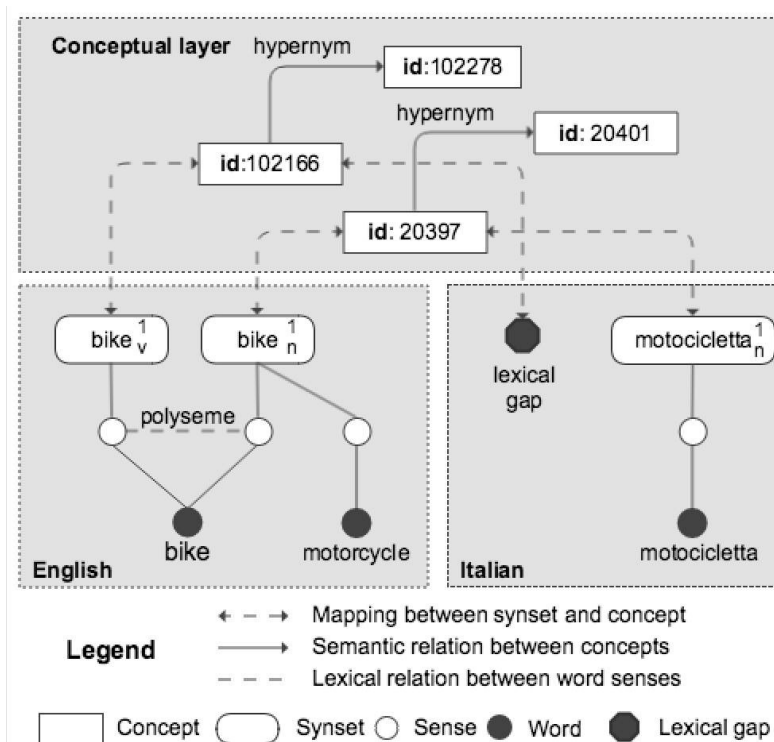


Figure 1: The UKC structure

The UKC is constantly growing by importing freely available resources that have been previously evaluated to keep our high-quality standards. Moreover, people are required to further populate the UKC, as we propose in this paper. Currently the UKC contains 335 languages, 1,333,869 words, 2,066,843 senses, and more than 120,000 concepts.

### 3 The Italian LKC

The development of the Italian LKC has been an experiment via collaboration between the University of Trento and the University of Macerata, which brought about the interdisciplinary merger of practices and methodological approaches of two distinct knowledge domains – namely Linguistics and Computer Science. A first experiment took place in the didactic context of a university seminar, where the students were the developers of the LKC and the professor was the instructor and final validator of their contributions. The local team was composed of five postgraduate students in Modern Languages for International Communication and Cooperation and an assistant professor in Computational Linguistics, whose roles were LKC developers and LKC validator/instructor, respectively. The students, whose mother-tongue was Italian, had a good knowledge of English; therefore, they correctly responded to the requirements of the UKC activity of equivalence compilation as required in their role. The professor, with an evaluator and trainer role, had a higher degree of expertise than the students regarding the background theory and procedures of the project, and was an expert in the fields of Linguistics, Terminology and Languages for Special Purposes (LSP), while the students could be considered semi-experts in these areas and in LSP Translation.

The inter-linguistic perspective and the analysis and definition techniques of the lexicon represented the common ground between the language developers' background knowledge and the requirements for the development of a lexico-semantic resource.

Linguistics and terminographic techniques (Wright & Budin 2001; Kockaert & Steurs 2015) were the principal skills required to accomplish the tasks oriented to the production of the bilingual lexico-semantic resource. Moreover, the students' expertise in terminology and terminography was applied both to the conceptual-semantic analysis of the lexicon and the elaboration of *intensional definitions* (Löckinger, Kockaert & Budin 2015). Indeed, the perspective adopted by the local team in their contribution to the LKC coincides with that used in the compilation of conceptually oriented lexical-semantic descriptions for the Italian equivalents of the English lemmas.

The shared knowledge between the two domains – i.e. terminology and LKC compilation – was used as a starting point for training in the current method, which entailed an adjustment of the developers' theoretical and practical approach to the activities of equivalence identification and definition writing. Therefore, the students' experience on (LSP) translation and terminological analysis turned into the ability to compile a computational lexico-semantic resource.

### 4 The User Interface

We present a preliminary version of the UI to facilitate contributions to build and maintain the UKC. In the UI presented in Figure 2, users can lexicalize English WordNet concepts into another language. In this preliminary approach, each contributing user will have a set of English WordNet concepts to lexicalize. The UI is divided into two parts:



- the top part contains the concept in the source language, Figure 2A, English in the example;
- the bottom part contains the corresponding empty fields to provide a lexicalization in the target language, Figure 2B.

If the concept to be lexicalized does not have a lexical equivalent in the target language, it can be marked by clicking on ‘Signal as GAP’, as in Figure 2B1. However, if there exists a lexicalization for the current concept, the user can complete the lexicalization by 1) adding the gloss and 2) selecting the corresponding POS from the pull-down menu, as in Figure 2B1; 3) adding a lexical equivalence for the lemma and an exceptional word form, that is irregular plural forms, irregular superlatives or irregular verb conjugations, when available as in Figure 2B2; and 4) adding an example as in Figure 2B3. When all the fields are completed, the user has to save the lexicalization first, ‘Save’ button, and then submit for evaluation, ‘Submit for validation’ button as seen in Figure 2B4. By clicking on the button ‘Translate Next’ a new English WordNet concept to be lexicalized will be available.

**Figure 2: UI design to complete a lexicalization.**

**Section A: Source Language (English)**

Reference Language: English (dropdown)

ConceptId: 45039

Gloss: a geographical area politically controlled by a distant country

POS: NOUN

Senses	Rank	Lemma	Exceptional forms
	1	colony	
	2	dependency	

Examples: Missing Example

**Section B: Target Language (Italian)**

Target Language: Italian (dropdown)

Synset: B1

Is GAP? ☐ Signal as GAP

Gloss:

POS:

Senses: B2

Rank	Lemma	Exceptional word forms
<input type="text" value="Add sense (write lemma for auto-completion from already present)"/>		<input type="text" value="Exceptional forms for the lemma, comma separated"/>

Examples: B3

**Section B4: Buttons**

Figure 2: UI design to complete a lexicalization. A corresponds to the source language part. B corresponds to the target language part. B1 corresponds to the gloss and POS of the word. B2 corresponds to the lexicalization of the word. B3 corresponds to the example of the word use. B4 buttons to save or submit the lexicalization.

In Figure 3 we present the UI to evaluate language elements. Again the screen is divided into two different areas:

- the left-hand side contains the concept in the source language, Figure 3A, English in this example;
- the right-hand side contains the lexicalized concept to evaluate, Figure 3B.

When a lexicalized concept is in the validation phase, it can be accepted by clicking on the ‘Submit for UKC validation’ or ‘Save’ if the concept needs further revision. By clicking the button ‘Validate Next’, a new concept to be evaluated will be shown. If a lexical element is marked as wrong the concept will be sent back to the lexicalization phase so that it can be revised.

**Reference Language** (English) **Provenance**

ConceptId: 21240

Gloss: a farm where pigs are raised or kept

POS: NOUN

Senses

Rank	Lemma	Exceptional forms
1	piggery	
2	pig farm	

Examples: Missing Example

**Target Language: Italian** **LOG** **Validate Next**

Synset **B1**

Gloss: fattoria dove si allevano i maiali

POS: NOUN

Senses **B2**

Rank	Word	Exceptional Forms
1	allevamento di maiali	

**B3** **Save** **Submit to UKC Validation**

Figure 3: UI design to complete an evaluation. A corresponds to the source language part. B corresponds to the target language. B1 corresponds to the synset. B2 correspond to the senses and B3 are the buttons to save or submit the evaluation over a concept.

## 5 Study Design

We run a study on the Italian LKC to obtain a comprehensive vision with respect to 1) different aspects related to the evaluation of the UI; 2) how its content is perceived from the participants' point of view, and; 3) the feasibility to build a lexico-semantic resource collaboratively. As far as we are concerned, only YARN (Braslavski, Ustalo & Mukhin 2014) involved users in a pilot study. However, the methods used to capture participants' opinions were not elaborated. In our study, we collected data using four methods to help us clarify contradictions in case any inconsistencies might be found. The approaches selected were: 1) think aloud (McDonald 2012) to understand and observe how they completed a translation task; 2) semi-structured interviews (Galletta 2013) to get a deeper insight on participants' views and opinions on the UI and the content included in it; 3) desktop video-recording while interacting with the UI; and 4) a background questionnaire to obtain the demographics of the participants.

### 5.1 Evaluation

We granted access to the UI to the participants, and they decided how to organize the tasks they were asked to complete. We assigned them different sub-trees of the location domain related to region, geographical area, line, space, and point. Each of them contained between 75 and 127 nodes that corresponded to different concepts. After a period of three months, we met individually with the participants to interview them. All of them agreed to be voice-recorded and allowed the use of the

resulting data for further analysis. The study was divided into three parts. Initially, we collected demographic information using a questionnaire. After that, we asked them to complete two translations using the UI while verbalizing what they were doing. Finally, we used semi-structured interviews to understand aspects related to the UI and how users perceived the content. Interviews were transcribed and thematically analyzed (Braun 2006).

## 5.2 Results

After three months, a total of 127 concepts were translated, among which 24 were classified as *lexical GAPS*. In spite of their other academic activities and the failures in the server where the UI was hosted, some of the students translated around half of the concepts they were assigned.

### 5.2.1 Interface Layout

With regard to the current implementation of the UI, the participants suggested that the design of the UI to lexicalize could be improved so that they always have at hand what needs to be lexicalized. Participant 1: “Sometimes I had to read the gloss several times, so I had to go up and down on the screen. I think it would be better to have everything on one screen”. Their observation is corroborated after analyzing the video of their interactions, as they had to scroll up and down the page. Some of them pointed out that the visualization of the set of concepts they were assigned would have helped them to understand the relation among the words they had to lexicalize and, as a consequence, they could produce glosses accordingly. Participant 2: “Once I found the word ‘colony’ meaning colony of the United States and after, I found again ‘colony’ with a more general definition. Initially, I didn’t know that I would find a second one, so I gave a more general definition in the first place. However, when I found it for the second time, I had to return and change the first definition.” Some others felt that the way the tasks were presented was rather disorganized, making them feeling disoriented. Participant 4: “The fact that I could only see the current word instead of all the words I was assigned, made me feel that it was disorganized”, Participant 6: “when I log in as a validator I get random entries”.

### 5.2.2 Task Perception

The participants were enthusiastic about the tasks and the research process to find lexical equivalents, as it allowed them to learn nuances in the meanings of the words. Participant 4: “It helped to enrich my vocabulary and it is very useful to understand the language”. In general, they felt that the experience of using a system like this was enriching and challenging. Participant 1: “it is very demanding because it needs a lot of research. It required a lot of time, but it was never boring. It was a very enriching task”. As observed, the participants were very precise when completing their lexicalizations, as they were checking different monolingual dictionaries (LSP) corpora, in English and Italian, as well as trusted websites and images. They would only complete the translation when they really had a clear idea on how to add the lexical equivalent for the given concept.

### 5.2.3 Collaboration

The participants shared their experiences and doubts when their tasks were similar. Participant 4: “After finishing a task we compared what one has done with the others... contrasting always helps”. When finding difficult concepts they asked the professor what steps to follow or what lexicalization for a specific object would be better. Participant 2: “with respect to conflictive cases, in order to create the gloss I would ask the professor what is a better option”. The professor taught them how to produce a good lexicalization that would not be a literal translation of the English concept. She evaluated the lexicalizations produced, as well as replied to the different enquiries from the students so that there

was a constant flow of information and feedback. Participant 6: “My students wondered whether they had to identify equivalents of the synsets whose definition might belong to another domain. I said ‘no’ because these concepts are not related to the domain of space/location”. Most of these communications were done face to face or via email.

## 6 Conclusions and Future Work

In this paper, we have presented a preliminary study to evaluate the design of a UI to maintain a multilingual lexico-semantic resource, the UKC, and whether collaboration among people is a feasible way to build and maintain a resource of these characteristics. We conducted a user study for three months in which six participants, five students and one professor of Linguistics, were involved. Although the total number of concepts lexicalized can be considered as low, mainly due to the failure of the system that forced the participants to access it in a discontinuous manner, this study helped us to obtain various improvements that could be introduced in the design of the UI, such as the inclusion of communication facilities and a of the redesign layout. We thus believe that it is possible to build a lexico-semantic resource based on collaboration. As shown here, the students were collaborating with each other while lexicalizing, as well as with the professor who was providing feedback and corrections to their work. This arrangement, where the professor is evaluating the semantic equivalences produced by the students, can be seen as the most basic configuration. However, this can also be our baseline to understand if future collaboration settings, such as peer-to-peer, where students are lexicalizing and evaluating each other, could improve the results. In the future, we plan to import more freely available resources and involve contributors from different countries. We already have ongoing collaborations with groups in China, India, Mongolia, Romania, South Africa, and the United Kingdom (for Gaelic). The approach seems to be scaling without difficulty, at least from a technological point of view. The real difficulty is organizational: how to find and coordinate people from so many different countries working in parallel. The approach we are following is to build a community and a non-profit organization that will collaboratively manage the evolution of this resource.

## References

- Bond, F., Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *51st Annual Meeting of the Association for Computational Linguistics. 4- 9 August 2013*. Sofia, Bulgaria.
- Braslavski, P., Ustaloc, D., & Mukhin, M. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In *Proceedings of the Demsontrations at the 14th Conference od the European Chapter of the Association for Computational Linguistics, ecac12014, 26-30 April 2014*. Gothenburgh, Sweden.
- Eickhoff, C., de Vries, A.P. (2013). Increasing cheat robustness of crowdsourcing tasks. In *Information Retrieval*, 16(2), pp 121-137.
- Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication*. New York/London: NYU Press.
- Giunchiglia, F., Jovanovic, M., Huertas-Migueláñez, M., & Batsuren, K. (2015). Crowdsourcing a large scale multilingual lexico-semantic resource. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP2015, 8-11 November 2015*. San Diego, USA.
- Giunchiglia, F., Batsuren, K., & Bella, G. (2017). Understanding and Exploiting Language Diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI2017, 19-25 August 2017*. Melbourne, Australia.
- Giunchiglia, F., Batsuren, K., & Freihat, A.A. (2018). One World – Seven Thousand Languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018*. Hanoi, Vietnam.

- Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. & Horton, J. (2013). The Future of Crowd Work. In *Proceedings of the 2013 conference on Computer supported cooperative work, CSCW 2013, 23-27 February 2013*. San Antonio, TX, USA.
- Kockaert, H.J., Steurs, F. (2015) (eds.). *Handbook of Terminology*. Vol. 1. Amsterdam/Philadelphia: John Benjamins.
- Löckinger, G., Kockaert, H.J., & Budin, G. (2015). Intensional Definitions. In H.J. Kockaert, F. Steurs (2015), pp. 60-81.
- McDonald, S., Edwards, H.M. & Zhao, T. (2012). Exploring Think-Alouds in Usability Testing: An International Survey. In *IEEE Transactions on Professional Communication*, 55(1), pp. 2-19
- Meyer, C.M., Gurevych, I. (2013). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. Chapter 13 in S. Gragner, M. Paquot. (eds.) *Electronic Lexicography*. November 2012. Oxford: Oxford University Press, pp 259-251.
- Miller, G. (1995). WordNet: a Lexical Database for English. In *Communications of the ACM*, 38(11), pp 39-41.
- Morita, D., Ishida, T. (2009). Designing Protocols for Collaborative Translation. In *International Conference on Principles and Practice of Multi-Agent Systems, PRIMA2009, 14-16 Nov 2009*. Nagoya, Japan.
- Munro, R. (2010). Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Translation Crowdsourcing for Translation. 31 October 2010, Denver, USA*.
- Rashid, A.M., Ling, K., Tassone, R.D., Resnick, P., Kraut, R. & Riedl, J. (2006). Motivating Participation by Displaying the Value of Contribution. In *Proceedings of the SIGCHI conference on Human Factors in computing systems, CHI2006, 22-27 April 2006*. Montreal, Canada.
- Wright, S.E., Budin, G. (2001). *Handbook of Terminology Management: Application-Oriented Terminology Management*. Vol. 2. Amsterdam/Philadelphia: John Benjamins.
- Zhang, H., Ojha, S.R. & Giunchiglia, F. (2017). Finding errors in a Chinese lexico-semantic resource using GWAP. In *Proceedings of the IEEE Eleventh International Conference on Semantic Computing, ICSC2017, 30 January-1 February 2017*. San Diego, USA.



# The CPLP Corpus: A Pluricentric Corpus for the *Common Portuguese Spelling Dictionary (VOC)*

**Maarten Janssen<sup>1</sup>, Tanara Zingano Kuhn<sup>1</sup>, José Pedro Ferreira<sup>1</sup>, Margarita Correia<sup>1,2</sup>**

<sup>1</sup>CELGA-ILTEC, Universidade de Coimbra, <sup>2</sup>FLUL, Universidade de Lisboa

E-mail: maartenjanssen@uc.pt, tanarazingano@uc.pt, jpf@uc.pt, margaritacorreia@uc.pt

## Abstract

The Pluricentric Corpus of the Portuguese Language (CPLP Corpus) aims to provide comparable corpora for the national varieties of the countries where Portuguese is an official language, making it possible to undertake corpus-based comparisons among the varieties of these countries. It is intended as a publicly available corpus for comparative linguistics and language resource development, but furthermore constitutes one of the pillars of the *Vocabulário Ortográfico Comum da Língua Portuguesa (VOC)*, the official spelling dictionary for Portuguese. The headword list in *VOC* is partly derived from lexicographic tradition, which is to date based almost exclusively on the European and Brazilian varieties, and partly made up of words retrieved from the CPLP corpus, many of them included for the first time in official language resources for Portuguese. This double inclusion route aims at presenting an integral (i.e., non-contrastive) and increasingly balanced perspective on all the varieties. This paper describes the general design of the corpus, the challenges faced in its development, as well as the way it was used in the compilation of *VOC*.

**Keywords:** corpus, pluricentric languages, Portuguese, spelling dictionaries

## 1 Introduction

Portuguese is spoken by over 260 million people (Reto et al. 2016) in Africa, Asia, Europe and South America. It is an official language of Angola, Brazil, Cape Verde, Guinea-Bissau, Equatorial Guinea, Mozambique, Portugal, São Tomé and Príncipe, and Timor-Leste. Despite this geographical heterogeneity, attention has been given mostly to the varieties of Brazil and Portugal, meaning that language resources for the others are scarce (Branco et al 2012). One of the resources that are missing is a corpus of Portuguese in which these under-represented varieties are given the same status as the varieties from Brazil and Portugal. While some multi-varietal corpora have been built, most notably the Corpus Africa (Nascimento et al. 2008), which covers African varieties, their size and coverage are limited. The corpus described in this article, called the *Corpus Pluricêntrico da Língua Portuguesa* – CPLP Corpus (‘Pluricentric Corpus of the Portuguese Language’) aims to provide a corpus that represents all varieties equally, rather than one focused on Brazilian and European Portuguese only.

This paper consists of two main parts. The first introduces the CPLP Corpus, which is comprised of sub-corpora from different countries, describing the compilation process, from data extraction to the methods used for cleaning, tagging, and balancing the corpus. Moreover, we will give an overview of the particular challenges concerning the creation of a pluricentric corpus of this type. A general description of the characteristics of the corpus will also be provided.

The second part demonstrates how the CPLP Corpus is integrated into *VOC*, in which it plays a double role. On the one hand, the CPLP corpus constitutes one of the two pillars used to establish the headword lists of the official national spelling dictionaries, which are available as marked-up sub-selections of the entire dictionary. And on the other hand, the online interface of the dictionary uses the

corpus to indicate for each word in *VOC* whether it has a high, low, or medium frequency in a given country. Before moving on to those two main parts, the next section will provide some background of the way the corpus came to be, and the rationale behind it.

## 2 Background

Portuguese spelling is determined in legally-binding, state-sanctioned documents, which means that in the countries where Portuguese is an official language, official documents written in Portuguese are obliged to follow the official spelling. Until recently, there were two official spelling norms for Portuguese: the *Formulário Ortográfico* ('Orthographic Guidelines') from 1943, which was followed in Brazil, and the *Acordo Ortográfico da Língua Portuguesa* ('Portuguese Language Orthographic Agreement') published in 1945, which was the norm in Portugal, Angola, Cape Verde, Guinea-Bissau, Mozambique, and São Tomé and Príncipe. In 1990, the Portuguese-speaking countries signed an orthographic agreement treaty (*Acordo Ortográfico da Língua Portuguesa - AOLP1990*, 'The Portuguese Language Orthographic Agreement'), with the objective of unifying the official spelling rules in different countries. The actual application of the spelling rules for Portuguese is traditionally made clear through *vocabulários* ('spelling dictionaries'), which up until AOLP1990 were typically national-level and were only published for Brazil and Portugal. In order to support the implementation of the AOLP1990, the countries that signed the treaty projected the creation of a common spelling dictionary for all countries, which is called the *Vocabulário Ortográfico Comum da Língua Portuguesa* (*VOC* – 'Common Spelling Dictionary of the Portuguese Language', Ferreira, Correia, & Almeida (Orgs.) 2017).

The agreement was only officially implemented in 2009, and the development of *VOC* started in the following year, under the supervision of the International Institute of the Portuguese Language (IILP), a multilateral bureau for language policy of the Community of Portuguese-Speaking Countries (CPLP – *Comunidade dos Países de Língua Portuguesa*). During a transitional period in which both the old and new spelling norms were simultaneously accepted, *VON* – *Vocabulário Ortográfico Nacional* ('national-level spelling dictionaries') were published in Brazil and Portugal following the new spelling rules of the AOLP1990. In the context of the development of *VOC*, these spelling dictionaries were made compatible and integrated into a single framework called OSLIN (Janssen 2005). At the same time, new spelling dictionaries were developed from scratch for the first time for the other countries (*VON* for Cape Verde, Mozambique, Timor-Leste and São Tomé and Príncipe are already available, with development still on-going for Angola and Guinea-Bissau). *VOC* is a free-access spelling dictionary that aims to represent the contemporary lexicon of Portuguese as a whole, in a framework and set-up that is common to all countries in CPLP (Ferreira et al. 2012).

The collection of lexical data for the development of *VOC* involved, among other methods, corpus-based acquisition of lexical entries. It was decided that headword lists of at least 30,000 entries per country would be extracted from lexicographic and para-lexicographic sources and from corpora provided by each country, intending to represent the way Portuguese is used in written contexts at a national level. An evaluation of existing corpora indicated that they were too small to attain this objective, and a decision was thus made to develop new corpora for the compilation of the national spelling dictionaries. The project to create those corpora set common design criteria, such as corpus size, textual genre types, and balance, to guarantee equal representativity among different national varieties of Portuguese (Almeida et al. 2013). However, due to a number of reasons, the compilation of some of these corpora was never finished: in some instances there were political reasons (e.g., Guinea-Bissau), while in others (e.g., Timor-Leste) it was the lack of available digital sources that prevented attainment of the initial objectives. The closest to completion were the corpora for Cape

Verde and Mozambique, which reached over 90% of corpus compilation goals, enough to fully enable their primary intended role in VOC.

Given the unquestionable importance of the existence of corpora for these currently under-studied and under-resourced linguistic varieties of Portuguese, and the fact that a great part of the compilation work and source acquisition had already been started, a decision was made to fully develop these corpora as an independent project, the *Corpus Pluricêntrico da Língua Portuguesa* (CPLP corpus). The CPLP corpus treats Portuguese as a pluricentric language in the sense of Clyne (1992:1), who defines such as language as one with several interacting centers, each providing a national variety with its own norms. Baxter (1992) affirms that Portuguese is a pluricentric language with two standards, the Brazilian and European varieties, and that “[t]he two standards differ from each other in phonology, morphology, syntax, lexicon, spelling and pragmatics” (Baxter 1992:35). Although in official discourse Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Príncipe, and Timor-Leste adopt the European variety as their standard language, it has been shown that some of these countries have emerging autonomous norms (Gonçalves 2010). Nevertheless, the varieties of those countries are not yet fully described and codified. Thus, one of the purposes of the CPLP corpus is to contribute to the study of these under-represented varieties of Portuguese, and serve as a base for the development of representative language resources for them.

### 3 The CPLP Corpus

The pluricentric nature of the CPLP Corpus is twofold. On the hand, the CPLP Corpus aims to encompass comparable sub-corpora that are representative of the written Portuguese variety in each of the countries of the CPLP. On the other hand, while all tasks related to computational processing, homogeneity of formats, and balancing are the responsibility of a core team, the sources and data for each country are vetted by national-level teams who in the future should manage their own representative corpora. This means multiple centers build and manage the CPLP Corpus.

Furthermore, the CPLP corpus is balanced. On the one hand, country-specific sub-corpora have the same approximate size; on the other, each sub-corpus has the same internal make up, containing comparable representativity of the same textual genres. The balancing over the different genres is given in Table 1.

Table 1: Distribution by genre of the *VON* corpora.

<b>Journalistic</b>	25%	Texts taken mostly from newspapers, but also including magazine texts, taken from online versions where available, or provided by local teams from written printed sources.
<b>Parliamentary</b>	25%	Transcriptions of the parliamentary meetings, partly representing the local formal spoken variety of the language.
<b>Literary</b>	20%	Consisting of prose and provided by local publishers or directly by the authors.
<b>Academic</b>	25%	Academic texts written at the universities of the various countries, typically available through open repositories, and focusing as much as possible on common specific domains (Health, Education, Sea and Environment, Agriculture and Energy, Law).
<b>Other</b>	5%	Privately-produced texts taken from websites within each country, typically blogs.

The differences in available data for different countries is huge, meaning that in order to create a balanced corpus with equal sizes for all countries the varieties with less resources create a bottleneck for

the entire corpus. To make sure that a sizable corpus could nevertheless be provided for the larger varieties, several sub-corpora were defined. The sub-corpora sizes were planned in terms of thresholds, the less-represented countries being the bottleneck for the minimum threshold planned for all sub-corpora (Guinea-Bissau, São Tomé and Príncipe, Timor-Leste): a corpus counting material for all the countries, but with a modest scale of only three million tokens per variety. A second corpus of 30 million tokens per country is the second threshold, for those countries for which this is an attainable goal: Portugal, Brazil, Mozambique, Angola, and Cape Verde. For those benefiting from a greater wealth of data, which is to say Brazil and Portugal, a further threshold was planned for data acquisition in tandem.

An additional challenge that we face derives from the fact that most African newspapers written in Portuguese publish a substantial amount of material from news agencies based elsewhere, especially in Portugal, as extensive analyses have shown. This makes the creation of a clean pluricentric corpus for Portuguese, and likely for any pluricentric language, problematic: there is no easy and reliable way to verify through the data itself whether a text published in, for example, Mozambique, is indeed representative of the variety of Mozambique, or rather copied from or written by authors from different varieties. Moreover, the fact that, contrary to Brazil, Portuguese orthography was already the same before AOLP1990 in Portugal and African countries, means that discards using orthographic variation as a cue for variety distinction is a method that can reliably be used to tell apart Brazilian and European Portuguese texts (see Kuhn et al. 2017). Therefore, in order to keep the sub-corpora as representative of each variety as possible, which is to say, to reduce the number of texts not actually belonging to the variety in question minimum, a decision was made that the CPLP corpus for the African varieties would reject a large number of texts acquired from potentially problematic sources, in some instances giving preference to offline and less easy to process sources, which more reliably represent the variety of the country they were published in.

The CPLP Corpus will be made publicly available for consultation through TEITOK (Janssen 2016), a web-based platform for visualizing, searching, and editing corpora with both rich textual markup and linguistic annotation. A sub-selection of the corpus will be provided through the same platform for download, containing a balanced selection of texts that can be freely distributed.

## 4 *VOC* Integration

The main initial motivation behind the creation of the CPLP Corpus was the need for a reference corpus for *VOC*, which could provide source material for those countries with no lexicographic resources, along with frequency data for each of its constituting *VON*. To see how the CPLP Corpus was used for this purpose, it is important to highlight some of the structural decisions behind the design of the lexicon, as well as their motivations.

*VOC* is a reference resource for the implementation of the spelling rules defined by AOLP1990. Previous attempts at an orthographic agreement, namely in 1931, failed at the implementation level: different, national-level wordlists had divergent interpretations of a common legal text or explicitly introduced unilateral changes to the text itself. The contention points are historically rooted in incompatible views on the ideal character of the orthography: conserving traditionalized features or more transparently conveying phonemic information; forcing unique forms for all countries or allowing for country-level, phonemic-induced variation.

While retaining some non-phonemic features, AOLP1990 recognizes the fact that the spelling of Portuguese has a phonemic base, with country-level pronunciation motivating country-specific variants in some contexts. The most visible of these differences is that, much like French, the Portuguese spelling encodes the quality of vowels with diacritics, which motivates variation, since the vowel quality



can be different in Portugal and Brazil in a large class of contexts. Along with this, consonant clusters such as *ct* are written as they are pronounced, and the pronunciation again differs per country. As a result, AOLP1990 defines a number of cases in which the recommended spelling differs per country, as in the case of the vowel quality, where in Portugal ‘anonymous’ is written as *anónimo*, whereas in Brazil it is written as *anônimo*. For consonant clusters, in Brazil you write ‘fact’ as *fato*, whereas in Portugal you write *facto*. Yet, in purely legal terms, all spellings that are legally correct in one country are legally correct in all countries, allowing for common legal documents. So, in AOLP1990 there is a fundamental difference between a recommendable spelling (which is specific to each country), and a legally acceptable spelling (which is common to every country). As such, the online interface of *VOC* can display the entire accepted lexicon, showing all words for all countries, but by default will display the *VON* for the country from which it is being consulted.

Since these country-specific spellings have to reflect the pronunciation and, to some degree, the orthographic tradition within a country, it would not only be legally incorrect, but also impractical to have a central team define the spelling in each country. Therefore, a guiding principle in *VOC* is that each country is responsible for its own *VON*, both in terms of which words should form part of the core vocabulary for that country, and in the definition of what are the acceptable spellings of those words in the case of spelling variation.

For the development of *VON* a local lexicographic team was established in each country, responsible for the final validation of the contents of *VON* and for the definition of the sources to be taken into account for the constitution of the nationally-representative headword lists. Those data comprise the complete set of words already included in *VOC*, and all the words included in the (known) lexicographic tradition for that country. On top of this, the frequency of each of the words in the entire database in the corpus for that country, which is to say the frequency of each word in the sub-corpus of the CPLP corpus for that country, was also considered.

The corpus frequencies were provided as a guiding principle, and each local team was free to tailor the headword list independently of the frequency of each word in the corpus. For lexical selection, this means each *VON* is corpus-driven, but ultimately involves some degree of traditional lexicographic handpicking. And for the case of spelling variation, each local team had to decide whether only one or several of the officially allowed variants was correct for their *VON*. For some of those cases, corpus frequencies were of no use, since many of the words in question were among those having their spelling changed with the orthographic agreement (e.g. in every country apart from Brazil, *facto* would take that form regardless of its pronunciation, <ct> being orthographically opaque). For all those words used in the local variants (often loan words from languages spoken within the country, but not yet lexicographically registered), the corpus was the only legitimizing force, enabling the representation in official sources, for the first time, of a large number of perfectly valid and frequently used words in several countries.

Apart from being a guiding principle, the CPLP Corpus plays a second role in the online interface of *VOC* when displaying a specific *VON*. In order to give an indication of how common a word is, the system displays whether the frequency of the word is high, medium, or low in the corpus, where high means it is amongst the 10% most frequent words, medium means it is amongst the top 40%, and low means anything below that. These data are presented in the interface directly from the sub-corpus of the CPLP Corpus corresponding to the selected *VON*.

## 5 Conclusion

The CPLP corpus is the first pluricentric corpus for Portuguese of a substantial size, large enough to guide the development of pluricentric lexicographic material. It grew out of a corpus designed for the



creation of the official *VOC* spelling dictionary for Portuguese in all countries of the CPLP, which from 2010 onwards provided one of the fundamental resources for the creation of the *VON* for each country. The CPLP corpus finishes what that initial corpus set out to establish, but could not fulfil at the time, which is to be a pluricentric reference corpus that can be used to give corpus-driven usage information on words in the various countries, hence implicitly providing information about whether words are used internationally or specific to a given variety. In time, the CPLP corpus will replace the initial *VOC* corpus as the basis for the frequency information in *VOC*. Moreover, it is also intended to be a basis for future pluricentric lexicographic resources for Portuguese.

The CPLP corpus can be used not only for much-needed lexicographic work, but also for broader linguistic research, such as comparative studies of the different varieties of Portuguese, thus making it a highly relevant resource for the study of Portuguese as a pluricentric language and highly valuable for future lexicographic work, given that there is no single existing dictionary for varieties other than those of Brazil and Portugal. Future work includes further integration of the CPLP Corpus into *VOC*, by extending its role to directly present a selection of example phrases in the corresponding sub-corpus for each word in a given *VON* that is registered in the corpus.

## References

- Almeida, G. B., Ferreira, J.P., Correia, M. & Oliveira, G.V. (2013). Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. In *Estudos Linguísticos*, 42(1), pp. 204-215.
- Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., ... Bacelar, F. (2012). *The Portuguese Language in the Digital Era/A Língua Portuguesa na Era Digital*. White Paper Series. Springer.
- Baxter, A. (1992). Portuguese as a pluricentric language. In M. Clyne (ed.) *Pluricentric languages: Differing norms in different nations*. Berlin, New York: Mouton de Gruyter, pp. 11-43.
- Clyne, M. (1992). Pluricentric languages – introduction. In M. Clyne (ed.) *Pluricentric languages: Differing norms in different nations*. Berlin, New York: Mouton de Gruyter, pp. 1-9.
- Ferreira, J.P., Correia, M. & Almeida, G. B. (orgs.) (2017). *Vocabulário Ortográfico Comum da Língua Portuguesa*. Praia: Instituto Internacional da Língua Portuguesa / Comunidade dos Países de Língua Portuguesa.
- Ferreira, J.P., Janssen, M., Almeida, G. B., Correia, M. & Oliveira, G.M. (2012). The Common Orthographic Vocabulary of the Portuguese Language: A set of Open Lexical Resources for a Pluricentric Language. In *Conference on Language Resources and Evaluation (LREC)*, Istanbul, pp. 1071-1075.
- Gonçalves, P. (2010). *A Génese do Português de Moçambique*. Lisbon: Imprensa Nacional/Casa da Moeda.
- Nascimento, M. F. B., Pereira, L. A. S., Bettencourt, J., Estrela, A., Oliveira, S., & Santos, R. (2008). *Corpus África: as cinco variedades africanas do português*. In S. Frota, A. L. Santos (eds) *Textos Seleccionados. XXIII Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa: APL, pp. 373–384.
- Kuhn, T. Z., Janssen, M., Ferreira, J.P., Kosem, I. & Correia, M. (2017). Dealing with multiple orthographic standards within a single corpus: the case of Portuguese in the CoPEP corpus. In *Actes des 9èmes Journées Internationales de la Linguistique de corpus*, Grenoble, pp. 52-54.
- Janssen, M. (2005). Open Source Lexical Information Network. In P. Bouillon, K. Kanzaki (eds.) *Proceedings of the Third International Workshop on Generative Approaches to the Lexicon, May 19-21 2005*. Geneva: École de Traduction et d'Interprétation – Université de Genève, pp. 79-106.
- Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In *Proceedings of LREC 2016*. Portorož, Slovenia, pp.4037-4043.
- Reto, L., Machado, F.L & Esperança, J.P. (2016). *Novo Atlas da Língua Portuguesa*. Lisboa: Imprensa Nacional-Casa da Moeda.

# Málið.is: A Web Portal for Information on the Icelandic Language

***Halldóra Jónsdóttir, Ari Páll Kristinsson, Steinþór Steingrímsson***

*The Árni Magnússon Institute for Icelandic Studies*

*E-mail: halldo@hi.is, aripk@hi.is, steinst@hi.is*

## Abstract

*Málið.is* is a web portal on the Icelandic language and language use. It currently includes seven different resources, providing reliable information on the language. Six of the seven resources made available on the portal are living projects, constantly being updated. *Málið.is* provides easy and fast access to these, accessible on and designed for desktop and mobile devices. One of these resources contains a number of specialized terminologies. Thus it is a special characteristic of *málið.is* among language portals for other languages that it provides access to abundant domain specific vocabulary. Work is under way to add more resources to the portal, providing more diverse information. The aim of the project is to strengthen the Icelandic language in the digital era by making it easy to access information on the language, helping people becoming more proficient and confident language users.

**Keywords:** e-Lexicography, Icelandic, dictionary portal, orthography, phraseology, etymology, terminology

## 1 Introduction

Users' access to all kinds of information has been subject to great changes over the past two decades, as the demand for easy access to dictionaries and other lexical information has increased. Moreover, online linguistic information that is free of charge can play an important role in language policy and planning, not least for small- and medium-sized language communities, such as the Icelandic one. Iceland is a nation of 340,000 people, and approximately 90% of the population are native speakers of Icelandic.

*Málið.is* is a new web portal on the Icelandic language and language use. It is operated by The Árni Magnússon Institute for Icelandic Studies, a publicly funded academic institute, dedicated to a variety of tasks in the field of Icelandic language and literature. The web portal opened in November 2016 and at present provides access to seven different resources created and/or maintained by the Institute. Together, the resources provide reliable in-depth information on Icelandic. Work is still in progress to add more resources to the portal, which will give more diverse information.

The name (and web address) of the web portal, *málið.is*, translates to 'the language.is', which is descriptive and easy to remember. Access to *málið.is* is free of charge, and the portal is completely clutter free, showing no ads, notifications or other disturbances, only the results from the resources connected to the portal.

Users of the portal can find abundant data and directions on language use and usage, inflections, semantics, stylistic variation, etymology and more. But it also gives access to large amounts of terminological data, as it searches through a database of more than sixty terminological glossaries, each for a specific field of study. The principal target group for *málið.is* are students and professional language users, including writers, journalists and translators. But as the web portal also serves the general Icelandic speaking public, as we strive for plain and non-technical exposition and conciseness.

All but one of the dictionaries and datasets available on *málið.is* are living projects, constantly being updated and revised. The portal runs a separate copy of all the databases, adapted for fast search and access on the *málið.is* website. Every night the server automatically launches a program that checks for recent changes and updates the *málið.is* database accordingly. When searching on *málið.is* users get the local copy of the data, but as most of the resources included in *málið.is* have their own website, users are also provided with a link to the corresponding entries in each of the dictionaries or databases.

## 2 Rationale for Setting up the Portal

The main aim of the *málið.is* project is to strengthen the Icelandic language in the digital era by making it easy to access coded linguistic information, thus helping the public to become more proficient and confident language users. Digital resources for Icelandic are far more limited than for most other national languages in Europe. For that reason, it is even more important that the good quality data that already exists becomes as clearly noticeable as possible for users.

The Árni Magnússon Institute for Icelandic Studies plays a role to publish prescriptive dictionaries as well as descriptive ones, both contemporary and historical. It publishes terminologies and word-lists on language for technical and specialized purposes, offers language consultation and advice to the general public, and compiles text corpora and other language resources. Moreover, researchers at the Institute have edited and published various dictionaries and other diverse language materials over the years. This work constitutes the core of the material we present on *málið.is*.

A great inspiration in preparing *málið.is* was the Institute's participation in the COST Action IS1305: *A European Network of e-Lexicography* (European Network of e-Lexicography 2013), and its ongoing discussions about convenient access to scholarly and advanced dictionaries. In Iceland we have also seen growing demands in recent years for dictionaries and other language resources to be made available free of charge to the general public. Furthermore, there is widespread interest in Iceland among lay people in matters of the Icelandic language, in particular to etymological speculations, the coining of purist neologisms and matters of standard vs. non-standard grammar (Hilmarsson-Dunn & Kristinsson 2013), and the A great number of the queries the Institute gets from the general public shows this. With *málið.is* we try to fulfil these needs as well as possible.

## 3 The Resources

Seven resources are currently accessible through the portal. Initially we wanted the information disseminated through the portal to be highly relevant to everyday use of Icelandic. The decision on what was to be included was therefore based on that, and on the state of the resources' database – whether we could connect it to the portal with relatively little effort. These are the seven resources:

- **The Database of Modern Icelandic Inflection** (Kristín Bjarnadóttir 2012). The DMII contains 278,000 paradigms from Modern Icelandic, with over six million inflectional forms. It was created as a multipurpose resource, for use in language technology, lexicography, and as an online resource for the general public.
- **Spelling Dictionary** (Jóhannes B. Sigtryggsson (Ed.) 2016). The 2<sup>nd</sup> edition of the Icelandic spelling dictionary, revised in accordance to updates to official Icelandic orthography, published by the Ministry for Education and Culture in 2016 (*Íslensk málnefnd* 2016).
- **Dictionary of Modern Icelandic** (*Íslensk nútímamálsorðabók* 2018). The Dictionary of Modern Icelandic is a new dictionary only available online. It is compiled at the Department of

Lexicography at The Árni Magnússon Institute for Icelandic studies. It contains approximately 50 thousand words. The work on the dictionary commenced in 2013 and is ongoing.

- **Language Usage Database** (*Málfarsbankinn* 2018). A compilation of short articles giving advice on language use. It contains more than 7,000 articles on subjects from grammar and syntax to well-crafted and elaborate language.
- **The Icelandic Term Bank** (<http://www.ordabanki.hi.is>). The Icelandic Term Bank (ITB) contains around 60 bilingual or multi-lingual glossaries, and one monolingual glossary, containing terminologies in various fields. The glossaries contain terms in Icelandic, usually with corresponding terms in other languages, most commonly English. Often the terms are accompanied by definitions or explanations. In total the ITB contains more than 180,000 terms and is constantly growing.
- **Icelandic Etymological Dictionary** (Ásgeir Blöndal Magnússon 1989). Icelandic Etymological Dictionary (*Íslensk orðsifjabók*) was compiled by Ásgeir Blöndal Magnússon and originally published in 1989. It is the first and only Icelandic etymological dictionary, and contains approximately 25,000 entries.
- **Icelandic Wordnet** (Jón Hilmar Jónsson 2018). *Íslensk orðanet* (Icelandic Wordnet) is a semantic database with elements that allow it to be used much like an onomasiological dictionary or a thesaurus.

The web portal will grow in the coming years and work is under way to add more resources. Potential datasets to be added include a written language archive with usage examples dating from 1540 to the late 20<sup>th</sup> century, historical dictionaries and bilingual dictionaries. These resources could make the portal more useful to students of the language as well as for everyday users of Icelandic.

## 4 Expanding the General with the Specific

The interests and needs of regular users of language for special purposes (LSP) often tend to be overlooked when information on language and language use is compiled and disseminated through web portals and the like. Among the most innovative and important special features of *málið.is* is that a large number of specialized terminologies, containing domain specific vocabulary, can be accessed through the portal along with information on words and usage in the language in general. Thus, the language portal *málið.is* is particularly useful to authors of textbooks, translators, and scientists, to name but a few groups that write and read LSP texts on a regular basis. *Málið.is* provides information on grammar, orthography, phrases, semantics, etc., of the general vocabulary of Icelandic, and, in addition, it retrieves content from the about 60 different terminologies – containing about 180,000 terms in total – that constitute the Icelandic Term Bank (ITB).

Among the roles of the Árni Magnússon Institute for Icelandic Studies is to facilitate the collection of domain specific terms, and the coordination of their usage and definitions. The Institute operates the ITB (opened on the Internet in 1997) for this purpose. Since Icelandic is the national language of Iceland, it is an important national language policy goal that Icelandic terminologies exist and are readily accessible to language users in a variety of domains of society, science and technology. The Institute is responsible for both the ITB and for the databases that contain information on general Icelandic vocabulary. This favorable situation made it a natural choice for the language portal *málið.is* to incorporate the voluminous terminological data at hand.

Through *málið.is*, one can of course search for specialized terms by choice, as well as for other words, as the need may arise. But the portal also “recruits” new users of terminological resources as they become aware of existing terminologies when they search for information among the other types of

málið.is ferill

ferill farnest í 7 gagnadöfnum

### Beygingarlýsing íslensks nútímamáls

**ferill** Karlkynnaðfornb

### Stafsetningarorðabókin

**ferill** -inn ferill; ferlar á löngum ferli; vera á ferli; ferli; ferli

### Íslensk nútímamálsorðabók

**ferill** nafnord karlkyrn  
 atburðarorð: ferill, ferli  
*ferill hans var kannadur þegar hann stóti um starfið*  
*hinn hóf feril sinn með gleðisveg*  
 Sjúk 3 merkingar í orðabók

### Íslenskt orðanet

**ferill** no kk  
 vegja ferli  
 á ferli  
 vera á ferli  
 vegja ferli að baki  
 fara ferli  
 Sjúk 6 orðanetbrotin á Íslensku orðaneti

### Málfarsbankinn

Ekki er að jafnaði átt við það sama með orðunum **ferill** og **ferli**.  
 1) Orðið **ferill** merkir vegjanga slök, braut, leið, rás, skeið, stór starfsferli, avíllferli o.s.f. Orðið  
 getur líka átt við um línu sem drögn er á milli punkta.  
 2) Orðið **ferli** merkir vegjanga atburðarás, framvinda, röð viðburða.

### Íðorðabankinn

**ferill**  
 [Töfræði]  
 samheiti boglína, tvítt  
 [persón] curve

**ferill**  
 [Eðlisfræði]  
 samheiti lína  
 [persón] line

**ferill**  
 [Eðlisfræði]  
 [persón] curve

**ferill**  
 [Eðlisfræði]  
 samheiti braut  
 [persón] trajectory

**ferill** kk  
 [Fuglfræði]  
 samheiti haldin stefna  
 [skilgreining] Fyrirhugað leið loftfars miðað við yfirborð jarðar eins og hún er sett á flugkort.  
 [skilgreining] Á flugkortum, gerfum með hálvörpun er ferli með breytta- og ákvarðunamat  
 eða milli tvíþriggið haldsmála mæddur við þann tengdarhug sem er næst því að leggja með  
 vegu milli þeirra. Hann er sýndur í gráum frá norðri, ýmist sem réttur ferli, segulferli eða  
 -væðferli. Í stjörnufræði nefnist samsvarendi hugtak „haldin stefna“ .  
 [persón] track

**ferill**  
 [Hugbúnaðisfræðingur]  
 [persón] history

**ferill**  
 [Hugbúnaðisfræðingur]  
 [persón] curve

**ferill**  
 [Læknisfræði]  
 [persón] curve

**ferill**  
 [Læknisfræði]  
 samheiti ferli  
 [persón] curve

**ferill**  
 [Hugvísindisfræði]  
 [persón] curve

**stíð**  
 [Stjörnufræði- og vélfræðingur]  
 samheiti ferli  
 [persón] trail

**ferill**  
 [Stjörnufræði- og vélfræðingur]  
 samheiti stíð  
 [persón] trail

**ferill** fr  
 [Stjörnufræði]  
 samheiti ferli  
 [persón] trajectory

**ferill** kk  
 [Eðlisfræði- og sálfræði]  
 samheiti ferli  
 [skilgreining] Lína, drögn samkvæmt hnotum / hnotakerfi  
 [persón] curve

**ferill**  
 [Fransíska]  
 samheiti ferli  
 [persón] graph

**ferill**  
 [Lísa (Sundfyllingar á Íslandi fyrir alla)]  
 samheiti bogi  
 [persón] arc

**ferill**  
 [Lísa (Sundfyllingar á Íslandi fyrir alla)]  
 [persón] trace

### Íslensk orðsifjabók

**ferill** k. 'ganga, ferðagang, braut, slök', sbt. *feril*, *feril*, *feril* 'menjar (um e-ð)',  
 sbr. mál. *feril* 'skipaleið með ströndum fram'. Af sama toga er **ferli** í *ferlivist*  
 og *ferli*, og nýrðib **ferli** h. um framvindu e-s. Sjá *feri* og -all.

**feri**, **feri** kv. 'ferð', sbt. *feri*, *feri*, *feri* (s.m.), < germ. \**feri*, sk. *feri* og *feri*  
 (1). Af sama toga eru **ferull** 'ferðagangur' og **ferla** (s) s. 'hraka, fara aftur,  
 mistakast', sbt. *ferla* 'farast' og ísl. **ferill**, sbt. *ferill* *ferum* *ferum* *ferum* k.  
 'samferðamaður', c.l.v. < \**feru-g(a)ntau*. Sjá *feri*, *feri* (1) og *ferill*.

Figure 1: The search string *ferill*. Information is retrieved from all seven datasets. The Icelandic Term Bank is most prominent in these results, showing 17 hits.



language resources and dictionaries. A number of users may become aware of a domain specific usage of a number of Icelandic words, for the first time, by learning of it when using *málið.is*.

For example, someone might type in the search string *ferill* ‘trail, course’ in order to find out, for example, something about the orthography or the (complex) inflection of this particular Icelandic noun. In addition to such information, *málið.is* instantly exposes the user to abundant data on *ferill* as a term in a number of specialized terminologies; more precisely, 17 hits, retrieved from 11 different terminologies, in the following domains: statistics, physics, aviation, computer science, medicine, economics, engineering, seamanship, political science, geography, and epidemiology, as shown in Figure 1. By clicking on a red underlined string in the results screen, the user is transferred to that specific database.

## 5 Concluding Remarks

The web portal *málið.is* has been open to the public, free of charge, since November 2016 and more and more language users find their way to it. Bringing seven resources together in one portal is a new way to provide access to reliable in-depth information on Icelandic in the digital era, and by doing this users are directed to resources they do not know exist and are otherwise unlikely to find. Four of the resources were accessible on various websites run by the Institute, prior to the opening of *málið.is*. The others had either not been made accessible or were available only in print.

By giving access to specialized terminologies through the portal we aim to serve users of terminological collections as well as helping others become aware of the domain specific usage of a number of words.

Work is under way to add more resources. Amongst others four historical dictionaries, which have never been published online before, will be made available through *málið.is* as an experiment to explore if there is a demand for such resources on a portal like this one, and therefore whether they belong there.

## References

- Bjarnadóttir, Kristín (2012): The Database of Modern Icelandic Inflection. In: *Language Technology for Normalisation of Less-Resourced Languages SALTMIL 8 - AfLaT 2012*, 13–18. Accessed at: <http://aflat.org/files/saltmil8-aflat2012.pdf> [28/03/2018].
- European Network of e-Lexicography (2013): Cost Action IS1305. Accessed at: <http://www.elexicography.eu/> [01/12/2017].
- Hilmarsson-Dunn, Amanda & Ari Páll Kristinsson (2013): The language situation in Iceland. In: Robert B. Kaplan, Richard B. Baldauf, Jr. & Nkonko M. Kamwangamalu (Eds.): *Language Planning in Europe: Cyprus, Iceland and Luxembourg*. London / New York: Routledge, 100–169.
- Íslensk málnefnd. (2016): Ritreglur. Auglýsing mennta- og menningarmálaráðuneytis nr. 695/2016 með leiðréttingum. Accessed at: <http://islenskan.is/images/ritreglur-IM-2016.pdf> [28/03/2018].
- Íslensk nútímamálsorðabók (2018): Halldóra Jónsdóttir & Þórdís Úlfarsdóttir (Eds.). Accessed at: <http://islenskordabok.arnastofnun.is> [28/03/2018].
- Jónsson, Jón Hilmar (2018): “Íslenskt orðanet: Tekstbasert kartlegging og presentasjon av leksikalske relasjonjer”. Rapport fra Konference om leksikografi i Norden. Island 30. maj-2. juni 2017 [forthcoming].
- Magnússon, Ásgeir Blöndal (1989): *Íslensk orðsifjabók*. Reykjavík: Orðabók Háskólans.
- Málfarsbankinn* (2018): Jóhannes B. Sigtryggsson (Ed.). Accessed at: <http://malfar.arnastofnun.is> [28/03/2018].
- Sigtryggsson, Jóhannes B. (Ed.) (2016): *Stafsetningarorðabókin*. 2. Ed. Accessed at: [www.malid.is](http://www.malid.is) [28/03/2018].



# Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (MultiGenera)

*María José Domínguez Vázquez<sup>1</sup>, Carlos Valcárcel Riveiro<sup>2</sup>, David Lindemann<sup>3</sup>*

<sup>1</sup>Universidade de Santiago de Compostela, <sup>2</sup>Universidade de Vigo, <sup>3</sup>Universität Hildesheim

E-mail: [majo.dominguez@usc.es](mailto:majo.dominguez@usc.es), [carlos.valcarcel@uvigo.es](mailto:carlos.valcarcel@uvigo.es), [david.lindemann@uni-hildesheim.de](mailto:david.lindemann@uni-hildesheim.de)

## Abstract

The aim of the project is to develop a prototype for a generator of argument structure or valency realizations in terms of syntagmatic and paradigmatic combinations of Spanish, German and French nouns. The two main applications of the tool prototype we are aiming to develop, are (1) the generation of noun phrases as argument structure realizations that follow patterns related to semantic features, for the creation of corpus and web query strings; and (2) the knowledge-based generation of simple and complex noun phrases that are acceptable in a coherent sentence context. An essential step in developing these applications is the systematic description of the valency-related syntagmatic and paradigmatic properties of argument combinations. To this end, we have devised a methodology based on bidirectional mutual enrichment (bottom-down and bottom-up). With the aim of generating argument surface realizations, we will mainly use lexical knowledge represented in wordnets for Spanish, German and French for the semantic annotation of lexical prototypes and their subsequent paradigmatic expansion.

**Keywords:** noun valency, argument structure, combinatory patterns, corpus lexicography, ontologies, wordnet, natural language generation

## 1 Introduction: From PORTLEX to MultiGenera

The work done for the PORTLEX dictionary<sup>1</sup> (cf. Domínguez & Valcárcel, in press) has highlighted the limitations of corpus-driven methods in achieving the objective of a lexicographical project, at least from a dependency-valency perspective: to compile all acceptable types of constructions (Mel'čuk 2013). Corpora do not contain examples of all the realizations of the different actants or of their combinations. On the other hand, the examples found in corpora sometimes do not meet the intelligibility or conciseness requirements of a dictionary. Furthermore, the available corpora of a significant size are not semantically annotated, so that it is not possible to apply semantic filters when searching for syntactic patterns. For example, we may search for *muerte de* [NP] *por* [NP] ('death of [NP] by [NP]'), but we may not filter this search by certain semantic values for each of the [NP] slots, like, for example, *death of* [noun: +living being] *by* [noun: +disease], i.e. items the hyponyms of which typically would fit into that slot as argument. On the other hand, there will be examples found in corpora that formally do fit into the queried valency pattern, but that are not argument realizations. For example, the noun+adjective combination *control paterno* ('parental control') does constitute a realization of verb and 'agent', while *control férreo* ('rigid control'), also noun+adjective, does not, since the adjective here is a mere attribute to the noun (cf. Domínguez 2011). A further difficulty is linked to the fact that certain combinations of actants cannot be found in most corpora. For example,

<sup>1</sup> PORTLEX (*Diccionario multilingüe de la frase nominal / Multilingual dictionary of the noun phrase*), accessible at <http://portlex.es>.

in French *La saveur chocolat de vos gâteaux* ‘The chocolate flavor of your cakes’, which is documented in corpora, versus *La saveur citron de vos gâteaux* ‘The lemon flavor of your cakes’, which is not documented, despite being semantically and syntactically similar to the previous one and no less acceptable for a French speaker.

The multilingual tool prototype MultiGenera is designed according to the principles of Valency Grammar and Lexicography (Engel 1995), and following related work on wordnets, i.e. concept-based lexical resources represented as ontologies. This will allow an analysis of different lexical-semantic domains in Spanish, German, and French from an onomasiological-conceptual point of view. The development of such a tool thus involves a contrastive approach to nouns as lexical items that belong to certain lexical-semantic domains. In other words, we propose developing a tool for processing syntagmatic and paradigmatic combinations driven by data extracted from corpora and wordnets.

MultiGenera will apply semantic filters to the results of searches in semantically not annotated corpora, using wordnet ontologies (see the references in Section 2.2) for the different languages under analysis. Existing and newly developed tools will be combined in order to obtain detailed syntactic-semantic information. The tool therefore randomly generates realizations of an argument structure (i.e. in Spanish *el olor a* [noun: +animate/+material] *de* [noun: +material]<sup>2</sup>), by establishing a connection between semantic features ([+material], [+animated], etc.), on one side, and concepts belonging to wordnet ontology categories or that are hyponyms to wordnet concepts that represent the desired semantic feature, on the other: *el olor a humedad de su habitación* (‘the musty smell of his room’), *el olor a hombre de la chaqueta* ‘the jacket’s smell of man’, *el olor a caballo del establo* (‘the smell of horse in the stable’), etc. In order to filter out unacceptable data, generated realizations are then intersected with search results in corpora. However, our methodology also allows us to detect, through expert verification in each language, realizations and combinations that do not appear in corpora or on the web (for different reasons), but that are possible and acceptable. This would enable us to in some way overcome the limitation of corpora as an attestation method, since they only show a limited part of the combination possibilities of a language.

In many cases, it will also generate unacceptable sequences, but this is also a particularly interesting aspect of our research:

All in all we have received several hundred paraphrase clusters on the computer. Some of them contained serious mistakes, and it is precisely those clusters that were of exceptional interest to us. A linguistic processor incorporating a serious linguistic theory becomes, by the very nature of things, a gigantic testing ground for this theory. The computer makes mistakes unimaginable for a human. The analysis of such mistakes can be extremely revealing in the sense that it is a shortcut to correcting lexicographic and grammatical descriptions of which its linguistic software is composed. Moreover, very often experimenting with formal models of language on the computer results in genuine linguistic discoveries (Apresjan et al. 2003, p.11).

## 2 Methodology and Workflow

The design and development of MultiGenera follows five core working steps, the description of argument structure realization patterns (2.1.), expansion of lexical prototypes (2.2.), the generation of argument structure realizations (2.3.), argument combinations within noun phrases (2.4.), and context generation (2.5.).

2 In English, ‘the [adjective: +animate, +material] smell of [noun: + material]’ or ‘the smell of [noun: +animate, +material] from/in [noun: +material]’. For example: *el olor a tabaco de la habitación* ‘the smoke smell of the room’ or ‘the smell of smoke from/in the room’.

## 2.1 Description of Realization Patterns for Argument Structures

The starting point for this description is the set of valency patterns provided in PORTLEX for the ten selected nouns. PORTLEX valency schemata contain patterns for argument structure realization in syntactic slots, in several possible combinations, together with examples for the filling of the slots with lexical items. The following table shows some patterns provided by PORTLEX for the German noun *Geruch* ('smell') together with surface realization examples; note that the syntactic slots of arguments are annotated with semantic categories that will later allow the extraction of lexical data using lexical-semantic relations encoded in wordnet ontologies.

Table 1: Argument structures and semantic features for the German noun *Geruch*.

<b>Det.</b>	<b>{Adjective}<sup>3</sup></b>	<b>Noun Phrase Head</b>	<b>{Genitive Det.}</b>	<b>Noun A1<sup>4</sup> [Material]</b>
Der	angenehme	Geruch	der	Blumen
The	pleasant	smell	of the	flowers
<b>Det.</b>	<b>{Adjective}</b>	<b>Noun Phrase Head</b>	<b>von (+ {Det.})</b>	<b>Noun A1 [Material]</b>
Der	intensive	Geruch	von diesen	Männern
The	intense	smell	of these	men
<b>Det.</b>	<b>{Adjective} A1</b>	<b>Noun Phrase Head</b>	<b>nach(+ {Det.})</b>	<b>Noun A2<sup>5</sup> [Material]</b>
Der	menschliche	Geruch	nach	Schweiß
The	human	smell	of	sweat
<b>Det.</b>	<b>{Adjective}</b>	<b>Adj. A1 [Animate]</b>	<b>Noun Phrase Head</b>	
Der	intensive	männliche	Geruch	
The	intense	male	smell	
<b>Det.</b>	<b>{Adjective}</b>	<b>Noun A1 [Material]</b>	<b>Noun Phrase Head</b>	
Der	stechende	Schweiß	-geruch	
The	pungent	sweat	smell	

## 2.2 Expansion of Lexical Prototypes

The slot-filling lexical items listed in the PORTLEX schemes will be expanded by a list of lexical prototypes, i.e. frequent nouns or adjectives that belong to a semantic category that corresponds to the specified semantic role of an argument. In other words, for each argument-role-slot a general list of prototypical lexical items will be obtained, as shown in Table 2 for the Spanish argument structure Det. + *olor a* + common noun (*aquel olor a tabaco*, 'that smell of tobacco').<sup>6</sup> These lexical items will be collected by queries in *Sketch Engine*,<sup>7</sup> which provides large corpora for the three languages, together with frequency data.

Slot-filling prototypes found in corpora for every argument are associated with a semantic category following an ontology of semantic features, which has been specifically designed for MultiGenera. Four different levels are differentiated, ranging from the most general to the most specific.<sup>8</sup> Table 3

3 Curly brackets mean that an item does not appear necessarily according to the valency pattern.

4 A1' refers to the argument with the meaning 'someone or something, that has something'. In this case: The flowers have a pleasant smell.

5 'A2' refers to the argument with the meaning 'something belongs to a class or type'. In this case: The smell is of sweat.

6 The list is the result of an exemplary query to *eseuTenTen11* corpus using *Sketch Engine*. Results for French were obtained from *frTenTen12* corpus.

7 See <http://the.sketchengine.co.uk/>.

8 The first two levels cover the following main semantic features: [Material]: [substance] and [objects]; [Animated]: [human], [animal], [fungus] and [plants]; [Situation]: [static situations], [processes], [locations]; [Intellectual concepts]. The third and the fourth levels are more specific but they are not always applicable (see Table 3).



Table 2: Example of ranking ten lexical prototypes for the Spanish argument structure *olor a* + common noun.

Prototypes (nouns)	Corpus counts
<i>tabaco</i> ('tobacco')	235
<i>incienso</i> ('incense')	177
<i>pólvora</i> ('gunpowder')	155
<i>humo</i> ('smoke')	153
<i>humedad</i> ('humidity')	142
<i>gasolina</i> ('petrol')	132
<i>azahar</i> ('orange blossom')	92
<i>sudor</i> ('sweat')	90
<i>azufre</i> ('sulfur')	87
<i>naftalina</i> ('naphthalene')	79

shows an example of the semantic annotation carried out for MultiGenera concerning the exemplary lexical prototypes displayed in Table 2:

Table 3: Example of semantic annotation of lexical prototypes for the Spanish argument structure *olor a* + common noun.

Lexical prototypes	1 <sup>st</sup> Order	2 <sup>nd</sup> Order	3 <sup>rd</sup> Order	4 <sup>th</sup> Order
<i>tabaco</i> ('tobacco')	Material	Substance	Solid	Smoke
<i>incienso</i> ('incense')	Material	Substance	Solid	Chemical
<i>pólvora</i> ('gunpowder')	Material	Substance	Solid	Chemical
<i>humo</i> ('smoke')	Material	Substance	Gas	Smoke
<i>humedad</i> ('humidity')	Situation	State	Property	
<i>gasolina</i> ('petrol')	Material	Substance	Liquid	Fuel
<i>azahar</i> ('orange blossom')	Animate	Plant	Flower	
<i>sudor</i> ('sweat')	Material	Substance	Liquid	Excrement
<i>azufre</i> ('sulfur')	Material	Substance	Liquid	Excrement
<i>naftalina</i> ('naphthalene')	Material	Substance	Solid	Chemical

As a result, a conceptual map of the acceptable values for an argument-role slot can be visualized, showing also contrasts from language to language. For the argument A2 of the French noun *odeur* 'smell', for example, there are essentially three main semantic classes of lexical prototypes: [+Material, +Substance] (*sueur* 'sweat', *tabac* 'tobacco', *poudre* 'gunpowder'), [+Animate, +Plant] (*fleur* 'flower', *jasmin* 'jasmine'), and [+Material, +Object] (*pain* 'bread', *crêpe* 'crepe'). By manually associating the MultiGenera semantic category descriptors with wordnet synsets, we can validate the semantic relation (hyponymy, meronymy) between the category descriptor and the lexical prototypes, i.e. their semantic annotation, including disambiguation of word senses.

Beyond its usefulness for describing the semantic features of a nominal argument, this annotation process also makes it easier to expand the prototype lists using item relations encoded in wordnet. For each semantic class of prototypes we search for correspondences with categories or subcategories in wordnet ontologies. Thus, for example, for the group of lexical prototypes [+Material, +Object, +Food] of the argument A2 of the Spanish noun *olor* or its French equivalent *odeur*, it has been possible to establish a connection with the subcategory [+Food] of the TOP ontology. For the task of exploring the different semantic relationships with which wordnets operate, a specific API for database queries has been developed for each language of the project. Connections are not only established between groups of lexical prototypes and categories of the TOP (Rodríguez et al. 1998) and

SUMO ontologies (Niles & Pease 2003), but also with WordNet Domains (Gonzalez, Rigau & Castillo 2012)) and epinonyms (Guinovart & Solla 2018), and using hyponymy or meronymy relations encoded in wordnets for the three languages.<sup>9</sup>

When a link to wordnet items is set, all the existing items in the pertaining ontological category are extracted to form a set of candidates as argument slot fillers. In the case of the argument A2 for *olor* and *odeur*, all items belonging to the subcategory [+Food] of the TOP ontology will be considered. All candidate lists will be validated in two steps: an automated and a manual one. In automated data validation, an intersection is made between the list of lexical items in a subcategory or hyponym cluster in Wordnet and the lexical items collected from corpora for a specific syntactic slot. All items present in the corpus are automatically validated and the remainder are manually validated by experts in each of the languages of the project. Thus, following the previous example, items such as *liqueur* ‘liqueur’ (present in the corpus) and *hachis* ‘hashish’ (not present) would remain in the set, while *vitamine B2* or *vendange* ‘grape harvest’ would be excluded. The result is an expansion of the initial group of lexical prototypes that can go beyond the limitations of corpora.

### 2.3 Generation and Assessment of Single-Argument Surface Realizations

The automatic generation of the argument structures of the ten selected nouns to develop the prototype implies, once again, the joint use of several tools. In this case, in addition to wordnets for the extraction of the lexical knowledge, *Freeling*<sup>10</sup> will be used for the introduction of morphological data (gender, number and, for German, also the case). Noun phrases in Spanish, French and German are subject to rules of agreement between the determiner and head. Moreover, these languages often present contractions of prepositions and the articles that follow them. For example, for the argument A1 of *olor* we have: *el olor del pan fresco* (‘the smell of the fresh bread’).

For the automatic generation of the argument structures, we use own python scripts and our own API for accessing wordnet and semantic ontologies. The lists of candidate lexical items to fill in each argument slot will allow users of our tool to choose those they prefer to generate simple noun phrases.

### 2.4 Argument Combinations within Noun Phrases

We also aim to generate argument combinations within a noun phrase, as, for example, an argument structure with two semantic roles realized in prepositional phrases (e.g., in Spanish *el olor a sudor de tu ropa* ‘the smell of sweat from your clothes’). The aim is to assess the combinatorial compatibility of the paradigmatic sets defined before (see Section 2.3). In this way, the candidate lists for each argument will be combined with those of the other noun arguments in different positions within the noun phrase. This raises the issue of semantic constraints governing combinations of arguments. Separately acceptable semantic categories for filling in different argument slots can present problems when combined in the same nominal phrase. Thus, for the noun *olor*, ‘smell’, it is usually not possible to combine arguments A1 and A2 belonging respectively to the categories [+Animal] and [+Food]: (\**el olor a tomate de los ratones* ‘the tomato smell from mice’, \**el olor cárnico del cocodrilo* ‘the meaty smell of the crocodile’, \**el olor del caballo a tortilla* ‘the horse’s smell of omelet’).

Although the combinations already registered in PORTLEX will be used as a starting point, we suggest that the acceptability of argument structure realizations not attested in corpora can be validated by this method. As many acceptable argument combinations are not included in corpora, the

<sup>9</sup> At present, we use MCR 3.0 (Gonzalez, Laparra & Rigau), available at <http://adimen.si.ehu.es/web/MCR>, and the Extended Open Multilingual Wordnet (Bond & Foster 2013), <http://compling.hss.ntu.edu.sg/omw/summx.html>.

<sup>10</sup> See <http://nlp.lsi.upc.edu/freeling/>.

assessment of the generated argument combinations will have to be carried out manually by members of the project team. At the end of this phase, an exhaustive list of the acceptable combinations of arguments will be obtained for each noun, as well as valuable information on grammatical constraints.

## 2.5 Context generation

The main objective at this fifth stage is to create contexts of whole sentences for the generated nominal phrases. The results of a preliminary study on the noun *muerte* ‘death’ (Valcárcel & Domínguez 2016), where the acceptability of the generated nominal phrases was assessed by anonymous participants, suggests that the assessment of semantic acceptability of automatically generated nominal phrases could be improved by providing a whole-sentence context. This conclusion led us to MultiComb, a parallel project to MultiGenera but longer in duration. MultiComb, which is funded by the Spanish Ministry of Economy, Industry and Competitiveness, is focused on generating more acceptable and familiar output for a human speaker. It is necessary to distinguish here the generation of context at the phrase level, on one hand, and at sentence level, on the other.

### 2.5.1 Context Generation at the Phrase Level

For context generation at the phrase level, we add adjective attributes to the ten nouns selected for developing the prototype within MultiGenera (i.e. in Spanish *un fuerte olor a tabaco* ‘a strong smell of tobacco’), *aquel agradable olor a madera de su habitación* ‘that pleasant smell of wood in his room’). For a formal modelling and machine-readable annotation of these structures, a selection of basic lexical functions (LF) related to qualifiers is carried out, following the proposal of Mel’čuk (2013, 2015). In similarity to the working steps described above for the extraction of slot-filling lexical items, a selection of lexical prototypes according to frequency in corpora will allow the definition of paradigmatic sets for each LF. The lexical items to represent the paradigmatic restrictions will be collected from corpora and collocation dictionaries for the three languages involved. Obviously, these paradigmatic sets associated with LF will depend not only on each noun, but also on the specific lexical restrictions of each of the three languages. For example, in the case of the Spanish noun *olor* ‘smell’, we would obtain the following prototype lists for the selected LF:

- Magn (*olor*) = *fuerte* (‘strong’), *intenso* (‘intense’), *penetrante* (‘pungent, penetrating’)
- AntiBon (*olor*) = *malo* (‘bad’), *desagradable* (‘unpleasant’), *nauseabundo* (‘nauseating’), *rancio*, (‘rancid’), *insoporable* (‘unbearable’), *asqueroso* (‘nasty’), *fétido* (‘foul’)
- Bon (*olor*) = *agradable* (‘pleasant’), *fresco* (‘fresh’), *dulce* (‘sweet’)
- Ver (*olor*) = *característico* (‘characteristic’), *genuino* (‘genuine’), *verdadero* (‘real’)

This allows us to randomly program the appearance of adjectives linked to a noun by an LF, and obtain a more varied and human-like output. Again, the issue of semantic constraints arises here. For example, some semantic categories of the argument A1 of the Spanish *olor* (‘smell’) such as [+Flower] imply Bon adjectives while others such as [+Excrement] demand AntiBon qualifiers (see Table 4). Thus, it is at least striking for a speaker to hear or read a nominal phrase such as *el agradable* (Bon) *olor de las cloacas* [+Place, +Building, +Excrement] (‘the pleasant smell of the sewers’).

### 2.5.2 Context Generation at the Sentence Level

In this stage, the previously generated noun phrases (Det + noun + arguments) will fill in the valency slots of a verb. These sentence contexts will be limited to four basic syntactic structures: [Subject (NP) + Verb: *el olor a tabaco de la casa se disipó*, ‘the tobacco smell in the house faded away’], [Subject (NP) + Copula + Attribute: *el olor a tabaco de la casa resultaba insoporable*, ‘The tobacco

smell in the house was unbearable’], [Subject + Verb + Object (NP): *el vecindario sentía el olor a tabaco de la casa*, ‘the neighborhood noticed the tobacco smell of the house’] and [Subject + Verb + Prepositional Complement (Prep + NP): *Me enamoré del olor a campo de su ropa*, ‘I fell in love with the country smell of their clothes’]. This will allow us to generate sentence contexts with the most frequent valency patterns. Again, new sets of lexical prototypes will be created for the rest of slots of the sentence contexts on the basis of frequency queries in corpora and dictionaries.

As in the preceding working steps, it will be necessary to perform a manual assessment of the acceptability of a representative amount of output generated for each language. As a final result, users should be able to decide in a web interface the types of context they want to be displayed.

### 3 Conclusions

The combined methodology for data extraction based on the interoperability of different resources, as presented in this paper, will lead to the development of the MultiGenera prototype. Beyond this specific outcome, the project not only entails the description of lexical-semantic fields, but also the study of the noun as a valency carrier. To this end, a multi-layered analysis of syntagmatic and paradigmatic combination patterns and their distribution is being conducted for three different languages. Among potential applications of the results of MultiGenera, we may highlight, on the one hand, that the project will help to explore new ways for the semantic annotation of corpora using semantic categories and existing lexical knowledge ontologies such as wordnets. This might be interesting to improve several Natural Language Processing tasks, such as Natural Language Generation or rule-based Machine Translation. On the other hand, MultiGenera will also allow us to generate examples of valency patterns and argument structure realizations in three languages, which well may find application in language teaching and Lexicography.

### References

- Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL 2013, Sofia, pp. 1352–1362.
- Domínguez Vázquez, M.J. & Valcárcel Riveiro, C. (in press). PORTLEX as a multilingual and cross-lingual online dictionary. In M.J. Domínguez Vázquez, M. Mirazo Balsa, C. Valcárcel Riveiro (eds.) Studies on multilingual lexicography.
- Domínguez Vázquez, M.J. (2011). Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens. München: Iudicium.
- Engel, U. (1995). Tiefenkasus in der Valenzgrammatik. In L. Eichinger, H.-W. Eroms (eds.) Dependenz und Valenz, Hamburg: Buske, pp. 53–65.
- Gómez Guinovart, X. & Solla Portela, M.A. (2018). Building the Galician wordnet: methods and applications. In *Language Resources and Evaluation*, 52(1), pp. 317–339.
- Gonzalez, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12). Istanbul: ELRA
- Gonzalez, A., Rigau, G. & Castillo, M. (2012). A Graph-Based Method to Improve WordNet Domains. In *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science. International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Berlin, Heidelberg, pp. 17–28.
- Mel’čuk, I. (2015). *Semantics. From meaning to text*, vol. 3, Amsterdam/Philadelphia: John Benjamins.
- Mel’čuk, I. (2013). *Semantics. From meaning to text*, vol. 2, Amsterdam/Philadelphia: John Benjamins.
- Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), CSREA Press, pp. 412–416.

- Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F. & Roventini, A. (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In *Computers and the Humanities*, 32, 117–152.
- Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*. Marrakech: ELRA.
- Valcárcel Riveiro, C. & Domínguez Vázquez, M.J. (2016). Teste ‘muerte’: falantes a avaliar a aceitabilidade de frases nominais geradas artificialmente. Blog Post, Carlos Valcárcel Riveiro. Retrieved from <https://carlosvalcarcel.net/2016/11/30/teste-muerte-falantes-a-avaliar-a-aceitabilidade-de-frases-nominais-geradas-artificialmente/> [28.03.2018]

## Acknowledgements

The results of this work are related to the research project “Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos”, financed by the BBVA Foundation Grants for Scientific Research Teams 2017, and to the research project “Multilingual generator of noun argument structures with application in foreign language production”, financed by the Spanish Ministry of Economy, Industry and Competitiveness (Scientific and Technical Excellence Research Program, FFI2017-82454-P).



# A Lexicon of Albanian for Natural Language Processing

**Besim Kabashi**

*Friedrich-Alexander-Universität Erlangen-Nürnberg, Ludwig-Maximilians-Universität München*

*E-mail: [besim.kabashi@fau.de](mailto:besim.kabashi@fau.de)*

## Abstract

For a lot of applications in the field of natural language processing a lexicon is needed. For the Albanian language a lexicon that can be used for these purposes is presented below. The lexicon contains around 75,000 entries, including proper names such as the names of inhabitants, geographical names, etc. Each entry includes grammatical information such as part of speech and other specific information, e. g. inflection classes for nouns, adjectives and verbs. The lexicon is a part of a morphological tool and generator, but can also be used as an independent resource for other tasks and applications or can be adapted for them. Both information from some traditional dictionaries, e. g. spelling dictionaries, and a balanced linguistic corpus using corpus-driven methods and tools are used as sources for the creation and extension of the presented lexicon. The lexicon is still a work in progress, but aims to cover basic information for the most frequent tasks of natural language processing.

**Keywords:** Albanian, NLP lexicography, lexicon updating, corpus linguistics

## 1 Introduction

Lexicons are very important for a lot of tasks in the field of natural language processing / human language technology, where either only part of the information is extracted or the unabridged dictionary is used. For the Albanian language there are now many types of dictionaries, cf. Lloshi (1988), for an overview of the time before 1988. In the three decades since Lloshi's report, new dictionaries or new types of dictionaries for Albanian have been compiled, e.g. synonym dictionaries, cf. Thomai et al. (2004), and Dhrimo et al. (2002), antonym dictionaries, cf. Samara (1998), bilingual dictionaries, e.g. Newmark (1994), and many specialized dictionaries in the fields of social, natural, technical, and computer sciences.

With the beginning of the digital age and the intensification of natural language processing, there has been an increasing need for more lexical data. These can be used in many areas, either as final product, or to support the creation of other resources and tools/applications in the field of natural language processing, e.g. spell checkers, morphological analyzers and generators, or part-of-speech taggers.

For Albanian, only Murzaku (1994), a kind of orthographical/spelling dictionary, is available (in electronic form), which is a lexicon with ca. 32,000 entries, supplied with information about parts of speech and linguistic gender, which can be adapted for natural language processing. In particular, new vocabulary of the last two decades, after the social and political changes that occurred in 1990–1991, is not covered. For a lot of tasks more information is needed. Another dictionary, Snoj (1994), a reverse dictionary of the Albanian language, lists more detailed information than Murzaku (1994), i.e. four forms for nouns (Sg. Indef. Nom., Sg. Def. Nom., Pl. Indef. Nom., and Pl. Def. Nom.), and three forms for verbs (1P. Sg. Ind. Pres. Act. N.Adm., 1P. Sg. Ind. Aor. Act. N.Adm., and Participle). It corresponds with the information given in the traditional dictionaries of the Albanian language like Kostallari et al. (1980) and Kostallari et al. (1984).

Until the year 2010 the maximum number of lexical entries in a dictionary of the Albanian language was 48,000, cf. Thomai et al. (2006). The spelling dictionary by Dhrimo and Memushaj (2010) increased this number up to around 75,000 lexical entries, which is more than double the number in the spelling dictionary of Kostallari et al. (1976). Dhrimo and Memushaj (first edition, 2010 with around 75,000 lexical entries, second edition 2015 with around 81,000 lexical entries) also has more information, e.g. about syllabification (hyphenation, word division), for the first time for Albanian, and about rarely used word forms, which are given in addition to standard forms. Other dictionaries, e.g. Samara (1998), Dhrimo et al. (2002), and Thomai et al. (2004), also extend the lexical information that is available about Albanian. Both properties, the higher number of lexical entries as well as the new type of information, offer the possibility to use, combine and organize this information in different forms and ways for the tasks of natural language processing.

In addition to the creation of dictionaries in traditional ways, the enrichment of lexical data and types of data is very important to cover as much lexis and language properties as possible. For this purpose we have started using a 100 million word corpus, named AlCo (Albanian Corpus), which is compiled from a variety of sources, cf. Kabashi (2017). This corpus is used to update and revise the lexical data based on linguistic features/attributes, and on data like frequencies, collocations, or n-grams, extracted from the corpus. It is annotated with a fine-grained tagset designed by Kabashi and Proisl (2016). Together with morphological tools based on Kabashi (2015), a full form lexicon can be generated or word-forms can be lemmatized.

## 2 Some Notes on the Albanian Language

The Albanian language is used by ca. 5.5 million people in South-Eastern Europe, and ca. 1.5 million people in other parts of the world. Albanian is an Indo-European language that constitutes a subgroup of its own. It is on the same level as the Hellenic, Romance, Slavic or Germanic subgroups. The language is characterized by a diverse vocabulary with many loan words due to language contact with Greek, Latin/Italian, Slavic languages and Turkish, and due to the influence of French and especially English as world languages.

Albanian as a writing system is based on the Latin alphabet and writing. The Albanian alphabet is an extended one with combinations of basic letters of the Latin alphabet, i. e. digraphs (*dh*, *gj*, *ll*, *nj*, *rr*, *sh*, *th*, *xh*, and *zh*) and two letters with diacritic signs (*ë*, and *ç*). Seven of the thirty six letters of the Albanian alphabet are vowels (*a*, *e*, *ë*, *i*, *o*, *u*, and *y*).

Albanian has a rich morphological system. Nouns, adjectives and numerals have 20 forms each, combined from five cases (Nominative, Genitive, Accusative, Dative and Ablative), two numbers (singular and plural), as well as definiteness (indefinite and definite). Proper names are also declinable.

The use of multi-word units is typical of the Albanian nominal system, i. e. some words have articles or particles as their first part, written as two separate graphical tokens e. g. *mirë* adv., engl. good, vs. *i mirë*, masc. / *e mirë*, fem. adj., engl. good. According to Newmark et al. (1982) the categories of verbs are as follows: person (1st, 2nd, 3rd), number (singular and plural), voice (active and non-active, i. e. passive, middle, reflexive or reciprocal), mood (indicative, subjunctive, optative, admiring, and imperative), tense (present, past and future), aspect (common, perfect, progressive, inchoative, definite, and imperfect), finiteness (finite and non-finite, i. e. infinitive, participle, gerundive, and absolutive). Verbs (counted with infixed pronominal clitics) have up to 90 forms.

### 3 A Standard Lexicon

A dictionary, e. g. a spelling dictionary, as one type with minimal information, lists the lexical entries, separated in hyphenation places, and gives additional notes in relevant cases, e. g. a variable writing form of the entry. The lexical entries are ordered alphabetically. Each lexical entry contains at least information about writing, grammatical category (part-of-speech), and other properties like grammatical gender, or valency (in/transitivity) of the verb. The lexical entries of verbs and nouns in the *Spelling Dictionary of the Albanian Language* (1976), and also in later dictionaries e.g. Dhrimo & Memushaj (2010), are taken as the standard, and look like examples 1 and 2:

- (1) bím/ë, ~a f., sh. ~ë, ~ët (engl. plant)
- (2) sjëll fol. kal. ~ólla ~jëllë (engl. to bring)

The lexical entry (1) has the lemma (bímë), alternation of the definite form in singular (~a, i.e. bíma), the part-of-speech information (f. i.e. feminine and means the gender and so finally noun). Next the alternations of plural forms are given (i.e. sh.), in the indefinite (~ë, i.e. bímë) and definite (~ët, i.e. bímët). The lexical entry (2) has the lemma (sjëll), the part-of-speech information (fol. i.e. verb, kal. i.e. transitive), followed by the form alternation of the verb in the aorist (~ólla, i.e. sólla), and finally the participle of the verb (~jëllë, i.e. sjëllë).

The information in the dictionaries mentioned above can be adapted into a lexicon for natural language processing purposes. The information can also be combined in order to compile a new type of lexical data. For more details about the different types of lexical entries in the dictionaries of the Albanian language, see Kabashi (2015: 99–123).

### 4 Compiling an Albanian Lexicon for the Purposes of Natural Language Processing

We first give some notes on the work on and improvements to compiling lexicons for the purposes of natural language processing of the Albanian language.

#### 4.1 Improvements and Work in the Past

Kabashi (2003) compiled an electronic lexicon based on word lists extracted from different texts. The lexicon benefits from Kostallari et al. (1976) as well as from M. Snoj (Ljubljana), i.e. a wordlist, dated 1993, with grammatical information like in the *Spelling Dictionary of the Albanian Language* by Kostallari et al. (1976). The lexicon was primarily designed as component of a morphological tool (Kabashi 2003, 2004). The information in the lexicon was similar to a spelling dictionary with additional data about the inflection of each lexical entry of nouns, adjectives, and verbs. The number of the lexical entries comprised around 55,000.

Tromer and Kallulli (2004) presented a morphosyntactic tagger for the Albanian language. This uses “three source lexica for the operative lexicon: 1) the full-form lexicon 2) the stem lexicon and 3) the regular lexicon” (2004: 1237). The operative lexicon has around 53,000 lexical entries.

Piton et al. (2007) created an electronic dictionary and finite state automata/transducers for automatic processing of the Albanian language in the framework of the NooJ platform. It is not clear whether the lexicon can be used separately from this platform, or whether there are two parallel lexicons which correspond to each other.

Kadriu (2013) uses a lexicon with around 32,000 entries, together with their correspondent part-of-speech information. She uses the lexicon within the NLTK framework, i.e. a natural language toolkit written in the Python programming language, together with a set of regular expressions rules that correspond to them.

Kabashi (2015), based on previous work (2003, 2004), created a lexicon which is used as a base for a morphological analyzer and generator for word forms of Albanian. On the one hand it is integrated in the morphological tool, and on the other it can be used as an independent resource. For more details about the lexicon see Kabashi (2015: 99–123).

## 4.2 The New Idea

In all the above-mentioned works about the lexicons (in electronic form), the lexicon was somehow integrated in a framework or directly in the program code of the tool. The idea in Kabashi (2003) and Kabashi (2015) was to develop/compile a lexicon as a parallel and independent resource that can be used with other tools and applications. This means the data are machine readable and can be used for different tasks in natural language processing. The idea and work presented here is to extend the information of lexical entries in the lexicon presented in Kabashi (2015), beginning with orthographic/spelling information of difficult forms, syllabification information, updating of the morphological information (classification of words into part-of-speech inflection subclasses that make the application of exact rules corresponding to the respective regular expressions possible). A completely new kind of data is the phonetic information about the lexical entries. These data have already been created and are currently in the process of being proofread. The goal is to convert the data into the Sampa format.

In general, the new lexicon presented here aims to follow the *CELEX Lexical Database*, cf. Baayen et al. (1995), but with state-of-the-art methods and goals, as linked data, as well as data supplied with up-to-date information on statistics and other data derived from corpora. As an independent resource the lexical data can be revised, extended and updated more easily. Also, eventually more authors can collaborate on the resource.

In the following we present the compilation process of the lexicon.

## 4.3 Parts-of-Speech and Their Subclassification

As a first step we gave every noun and adjective, including numerals, a numerical declension class, as well as every verb their conjugation class. In this way the saved data are tested and can serve as reliable information. Eventually new additional lexical entries can be recognized, lemmatized and collected preliminarily using regular expressions, extraction rules and other methods. At this stage lexical entries appear as shown in example 3.

(3) ... adhuroj 7, afroj 7, aftësoj 7, agjëroj 7, ajkoj 7, ajoj 7, ajroj 7, ...

This information is needed for the modeling of morphological tools and grammars. An important part of the lexical entries are nouns, which are declinable in Albanian, e.g. the name Tirana can occur in the forms Tiranë, Tirana, Tiranës, Tiranën, Tirane. Most other names also have definite and indefinite plural forms, e.g. standard names, but also family names. They all need to be classified and supplied with these numbers.

## 4.4 Morphological Information as a Full-form Lexicon

As the next step we generate a full-form lexicon with the corresponding morphological information for each word-form. This data can be used for lemmatization of word-forms, generation of a

word-form using lemma and the morphological information, or for tagging any word-form with the morphologic information. Examples 4 and 5 show this data for a noun respectively a verb.

(4) Sample of the full-forms of nouns:

```
...
bimë/bimë/S-020_NS-;S-020_AcS-;S-020_NP-;S-020_AcP-
bima/bimë/S-020_NS+
bimën/bimë/S-020_AcS+
bimës/bimë/S-020_GS+;S-020_DS+
bimët/bimë/S-020_NP+;S-020_AcP+
bimëve/bimë/S-020_GP-;S-020_DP-;S-020_AbP-;S-020_GP+;S-020_DP+;S-020_AbP+
```

(5) Sample of the full-forms of verbs:

```
...
sjellim/sjell/V-036_1P.Pl.Ind.Prs.Act.Adm-;V-036_1P.Pl.Sbj.Prs.Act.Adm-
sjellin/sjell/V-036_3P.Pl.Ind.Prs.Act.Adm-
sjellka/sjell/V-036_3P.Sg.Ind.Prs.Act.Adm+
sjellkam/sjell/V-036_1P.Sg.Ind.Prs.Act.Adm+
sjellkan/sjell/V-036_3P.Pl.Ind.Prs.Act.Adm+
sjellke/sjell/V-036_2P.Sg.Ind.Prs.Act.Adm+
sjellkemi/sjell/V-036_1P.Pl.Ind.Prs.Act.Adm+
sjellkeni/sjell/V-036_2P.Pl.Ind.Prs.Act.Adm+
sjellkësh/sjell/V-036_3P.Sg.Ind.Ipf.Act.Adm+
sjellkësha/sjell/V-036_1P.Sg.Ind.Ipf.Act.Adm+
...
```

This data can be generated based on the inflection classes, i.e. conjugation and declension classes, and the corresponding paradigms. Moreover, new lexical entries can be easily integrated if they are classified as preliminary ones.

## 4.5 Lexicon Size

The presented lexicon includes the vocabulary which is covered by traditional dictionaries, and also additional lexical entries which are not covered by these. The lexicon has around 75,000 lexical entries, and includes 45,500 nouns, 18,500 adjectives, 5,800 verbs, 3,200 adverbs and other parts of speech and abbreviations.

## 4.6 Structure

The lexicon is organized in alphabetical order as one file, which has a clear and strict data structure (as tables), and as such they can be exported, converted and transformed in other structures or in any database. Each lexical entry, firstly organized as lines, separated in fields, has the properties of the part of speech which it belongs to, i.e. the structure of a noun is different to that of adjectives, to that of verbs, to that of adverbs and that of parts of speech, cf. the examples given below.

(6) Sample lexical entry of one noun and verb entry:

```
06241\bimë\bi-m\ë\bIm\ë\bimə\cv][cv]cvcv\4\2\3\4\bím~ë~a~ë~ët\ſ\020\
57195\sjell\sjell\sjell\sjë.ł.\ccv.cc.]\ccvcc\5\2\1\4\s~jèll\s~ó~lla\s~jé~llë\t\Ŵ\036\
```

The data in example 6 are as follows: The first field is the ID of the lemma, followed by the lemma itself, the syllabification of the lemma with the marking of the alternation segment. Next the information from the third field is converted in another writing form in the fourth field. Then the IPA



representation of the lemma follows. The syllabification segments are shown in the next field. Next is the queue of the consonants and vowels, followed by the number of letters of the lemma, the position of the accent, position of the alternation of the possible word-form(s), and the number of letters, where the digraphs count as one. The next four fields contain the word-forms Sg. Indef. Nom., Sg. Def. Nom., Pl. Indef. Nom., and Pl. Def. Nom. The last three fields show the gender, part of speech and the declension class of the noun. The data for a verb lexical entry given in example 6 can be interpreted in a similar way. The .ɫ is an IPA representation of the digraph “ll”, in the following field marked with .cc. because the two letters belong together. The number 4 means that “sjell” has four letters of the Albanian alphabet.

#### 4.7 Technical Aspects

The data are encoded in ISO/IEC-8859-1 (latin-1), ISO/IEC-8859-16 (latin-16) and Universal Coded Character Set (UCS), UNICODE, and saved in different formats, as well as UTF-8 parallel. For more detailed information on coding of the Albanian alphabet see Kabashi (2009).

The linguistic data themselves are correlated, but not in the desired form because there is still a need for manual intervention to link some data, e. g. update the number (IDs) of the lemmata and each word-form. Apart from this, other issues are managed well.

#### 4.8 Interoperability with other Resources

The main part of the data is taken from the lexicon compiled by Kabashi (2015). Other data are taken from the AlCo-Corpus, cf. Kabashi (2017). Some data, e.g. about syllabification, are compared with the corresponding data in Dhrimo and Memushaj (2015). Some data about syllabification and about some word-forms, that are not used so often, classified as difficult, as well information about accent/stress in some compound words, have been discussed with R. Memushaj (Tirana). The lexicon also benefits from some other data obtained directly from R. Memushaj in electronic form from time to time. New word-forms found extracted from the AlCo-Corpus can be lemmatized, and from the lemmata the full form paradigms can be generated, i.e. the new full-form lexicon with neologisms.

#### 4.9 Comparisons with other Albanian Resources and Lexicons

As mentioned and briefly introduced in Section 4.1, there are only a few resources for the Albanian language that are created and compiled for natural language processing purposes. The availability of the lexicon offered online by Murzaku (2003) is the first step to start with a lexicon with more than the basic vocabulary. Other resources and tools are not freely available at present.

#### 4.10 Status of the Project

The current state of the project is a work in progress, and new entries are added from time to time. This makes it necessary to recount the entries and to give a new number to the entries. In this context, linking of the data still presents some difficulties and needs to be revised. Linking data in the lexicon is currently being defined and can be changed.

The phonetic data for the word-forms are currently in the compiling process. The problems here are on the one hand the definition and marking of the syllabification and the accent, and on other hand the IPA-transcription of some of the lexical entries. At the moment this issue requires the most time working on the lexicon. Morphological data needs to be changed only in rare cases, when errors are detected.

As usual during electronic lexicographic work, some corrections are possible at any time. However, the work shown in detail in example 6 is already done.

## 5 Conclusion

The Albanian lexicon presented in this work for the purposes of natural language processing is a work in progress. The aim is to have an up-to-date, state-of-the-art, and contemporary lexicon, that can be used directly or with small adaptations, or can be easily converted into other formats or structures. As this is a one-man project, the work is proceeding slowly, based on current needs for some additional new data.

## References

- Baayen, R., Piepenbrock, R. & Gulikers, L. (1995): *The CELEX Lexical Database*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. Accessed at: <http://celex.mpi.nl> [28/7/2014].
- Dhrimo, A., Tupja, E. & Ymeri, E. (2002): *Fjalor sinonimik i gjuhës shqipe*. Tiranë: Toena.
- Dhrimo, A. & Memushaj, R. (2010): *Fjalor drejtshkrimor i gjuhës shqipe*. Tiranë: Infbotues.
- Dhrimo, A. & Memushaj, R. (2015): *Fjalor drejtshkrimor i gjuhës shqipe*. Botimi i dytë. Tiranë: Infbotues.
- Kabashi, B. (2003): *Automatische Wortformererkennung für das Albanische*. Master's thesis in Linguistische Informatik/Computational Linguistics. University of Erlangen-Nürnberg.
- Kabashi, B. (2004): Analiza automatike e fjalëformave të gjuhës shqipe. In: *Seminari XXIII Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*. Universiteti i Prishtinës, Prishtinë. Libri 23/1. 129-135.
- Kabashi, B. (2005): Disa propozime për modelimin e informacionit në leksikografinë kompjuterike. In: *Seminari XXIV Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*. Universiteti i Prishtinës, Prishtinë. Libri 24/1. 179-184.
- Kabashi, B. (2009): Das albanische Alphabet aus sprachtechnologischer Sicht. In: Demiraj, B. (Hrsg.): *Der Kongress von Manastir. Herausforderung zwischen Tradition und Neuerung in der albanischen Schriftkultur*. Hamburg: Verlag Dr. Kovač, 2009. 175-208.
- Kabashi, B. (2015): *Automatische Verarbeitung der Morphologie des Albanischen*. Erlangen: FAU University Press.
- Kabashi, B. & Proisl, T. (2016): A Proposal for a Part-of-Speech Tagset for the Albanian Language. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia. Ed. by Nicoletta Calzolari etc. European Language Resources Association (ELRA) Paris. 4305-4310.
- Kabashi, B. (2017, in publication process). AlCo – një korpus tekstesh i gjuhës shqipe me njëqind milionë fjalë. In: *Seminari XXXVI Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*. Universiteti i Prishtinës, Prishtinë.
- Kabashi, B. & Proisl, T. (2018): Albanian Part-of-Speech Tagging: Gold Standard and Evaluation. In: *Proceedings of the 11<sup>th</sup> International Conference on Language Resources and Evaluation (LREC 2018)*. 7-12 May 2018, Miyazaki, Japan. European Language Resources Association (ELRA) Paris. 2593-2599.
- Kadriu, A. (2013): NLTK Tagger for Albanian using Iterative Approach. *Proceedings of the 35<sup>th</sup> International Conference on Information Technology Interfaces (ITI 2013)*, June 24-27, 2013, Cavtat, Croatia.
- Kostallari, A., Domi, M., Lafe, E. & Cikuli, N. (1976): *Fjalori drejtshkrimor i gjuhës shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë.
- Kostallari, A. (Kryeredaktor), Thomaj, J., Lloshi, Xh., & Samara, M. (1980): *Fjalor i gjuhës së sotme shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë.
- Kostallari, A. (Kryeredaktor), Thomaj, J., Samara, M., Kole, J., Daka, P., Haxhillazi, P., Shehu, H., Sima, K., Feka, Th., Keta, A. & Hidi, A. (1984): *Fjalor i gjuhës së sotme shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë.
- Lloshi, Xh. (1988): Compiling and Editing Bilingual Dictionaries in Albania. In: *EURALEX 1988*.
- Murzaku, A. (1994): Albanian. In: *European Corpus Initiative Multilingual Corpus I (ECI/MCI)* CD-ROM. Utrecht: ELSNET.

- Murzaku, A. (2003): *Inverse Dictionary of Albanian*. Lissus Language, Literature, Computing. Albanian Linguistics. Accessed at: <http://www.lissus.com/albanian> [18/02/2018].
- Newmark, L. (1994): *Albanian–English Dictionary*. London etc.: Oxford University Press.
- Newmark, L., Hubbard, P., & Prifti, P. (1982): *Standard Albanian – A Reference Grammar for Students*. Stanford University Press, Stanford, CA.
- Piton, O., Lagji, K., and Përnaska, R. (2007): Electronic dictionaries and transducers for automatic processing of the Albanian language. In: *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007)*. 407–413.
- Samara, M. (1998): *Fjalor i antonimeve në gjuhën shqipe*. Shkup: Shkupi.
- Snoj, M. (1994): *Rückläufiges Wörterbuch der albanischen Sprache*. Hamburg: Buske.
- Thomai, J., Samara, M., Shehu, H. & Feka, Th. (2004): *Fjalori sinonimik i gjuhës shqipe*. Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.
- Thomai, J., Samara, M., Haxhillazi, P., Shehu, H., Feka, Th., Memisha, V. & Goga A. (2006): *Fjalor i gjuhës shqipe*. Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.
- Trommer, J. & Kallulli, D. (2004): A Morphological Analyzer for Standard Albanian. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. 26–28 May 2004, Lisbon, Portugal. 1271–1274. European Language Resources Association (ELRA) Paris.

## Acknowledgements

Many thanks to three anonymous reviewers for their valuable comments on a draft of the paper.

# Building a Gold Standard for a Russian Collocations Database

**Maria Khokhlova**

*St. Petersburg State University*

*E-mail: m.khokhlova@spbu.ru*

## Abstract

In the last decade, linguists have become increasingly interested in corpus material, which allows for a fresh approach to the phenomena that have already been extensively described in academic works. The dual nature of the co-occurrence phenomenon itself lies, on one hand, in its linguistic component and, on the other, in the probabilistic (combinatorial) characteristics. The former has been described in numerous papers and explicitly defined in dictionaries, while the latter can be identified by a statistical approach. The present paper focuses on the process of building a gold standard that will include data from Russian dictionaries and corpora. The standard is being prepared for a Russian Collocations Database that already includes information on words' collocability and was extracted from text corpora by statistical measures and linguistic filters. The gold standard will be also used for the evaluation of the extracted collocations and for marking them as "true" collocations with references to the dictionaries.

**Keywords:** database, collocations, corpora, dictionaries, Russian language

## 1 Introduction

The study of a language assumes a large amount of data about word usage, and about their joint occurrence. The competence of a native speaker involves not only knowing numerous meanings of different words of the language, but also understanding which words, if connected, can form a single semantic unit. One can speak about a general population, which would cover absolutely all the examples available in a language. However, it is not possible to collect such a population for objective reasons. Nevertheless, information on collocability can be obtained from already created resources. Information on lexical and syntactic co-occurrence of various words can usually be found in dictionaries (explanatory, specialized, etc.), and, although less often, in grammar books. Large text corpora that have been actively appearing recently also can be seen as a source of such data. But they cannot serve *a priori* as a source of the relevant material without an appropriate "superstructure", as they contain typos, occasionalisms, errors and repetitions of the same fragments (such problems often arise when the data is automatically collected). As such, information on collocability that is automatically obtained on the basis of text corpora should be accompanied by some evaluation, allowing them it to be verified. It should be noted that there is no unified source of information on collocability that researchers can consult. Despite some criticisms, it should also be noted that dictionaries and other lexicographical resources are the most valuable sources of information on the collocability that must be used when preparing a gold standard. Therefore, the project is aimed at solving the following two tasks: 1) creating a single resource of "reference" (verified) data; 2) the representation of "reference" data on the basis of the text corpora, i.e., real language sets. Here we mean a kind of a "gold standard" of collocability, or a reference list containing as much information as possible about the word combinations and collocations for the Russian language. By a gold standard we understand a collection of collocations extracted from explanatory and specialized Russian dictionaries with some statistical evaluation measured on corpus data. The present paper describes an ongoing project and is structured as follows. The Introduction presents the basic idea of the research. Section 2 describes the related

projects and provides an overview of the topic. Section 3 discusses the main ideas underlying the gold standard and gives an example of its part. The last section concludes the paper and proposes plans for future work.

## 2 Related Work

Over the last decade, linguistics has been less focused on prescriptive tools, giving scholars more opportunities to draw their own conclusions based on the analysis of examples. Besides, the work of a lexicographer is believed to possess an inevitable element of subjectivity that influences the words and word combinations selected for the dictionary, as well as the way and order in which they are grouped into dictionary entries. It is not uncommon for a dictionary to become outdated in the time period between the start of its compilation and its publication date. It should also be mentioned that there is no single concept for data presentation in various explanatory dictionaries. For instance, the Russian noun “*nadezhda*” ‘hope’ has only the following standard collocations listed in the dictionaries (Kuznetsov 1998; Dictionary of the Russian Language 1981-1984): “*vozlagat nadezhdu*” ‘to pin hopes’, “*pitat nadezhdu*” ‘to nourish hopes’, “*podayot nadezhdu*” ‘to give hope’ and “*l’stit sebya nadezhdoj*” ‘to flatter oneself with a hope’. These examples, however, do not include other collocations, which are also characteristic of this lexeme (for instance, “*opravdyvat nadezhdy*” ‘to justify hopes’ or “*vselyat nadezhdu*” ‘to inspire a hope’). Therefore, the presence of such a non-uniform representation of collocational preferences of lexical units indicates that there is a need for a single resource that would absorb information from different sources.

There are a number of online systems developed for the Russian language, which provide information on collocations. Among these, we can name such resources as the Lexicograph database, the FrameBank database of collocations (Lyashevskaya 2010) which includes descriptions of valency frames for verbs and constructions, and the “Collocations, Colligations, Constructions” database (Kopotev et al. 2015), providing information about collocations on the basis of the Russian National Corpus (RNC) and the ruWac corpus. One can name dictionaries created on the basis of RNC as a source of lexicographic data. The RNC provides a range of instruments (n-gram search with statistical analysis, lists of established words and set expressions, lexical graphs), and has also served as a basis for compiling dictionaries. There is the *Dictionary of Verbal Collocations for Abstract Nouns of the Russian Language* (Biryuk, Gusev, & Kalinina 2008). It lists more than 1,000 collocations based on the following models: 1) noun + verb; 2) verb + noun; 3) verb + adjective + noun. We should also note the *Dictionary of Russian Idioms: Combinations of Words with the Meaning of High Degree* (Kustova 2008). Their differences from the proposed resource include the following: these dictionaries provide information on co-occurrence for a limited set of key words (for instance, only for verbs); they also offer an interface that is non-intuitive for a general user. Another unique lexicographic project is the *Active Dictionary of the Russian Language* compiled under the guidance of Yu. D. Apresyan, which offers vast amounts of information on co-occurrences separately represented in dictionary entries. The material is well-structured and includes data on syntactic actants, collocations and constructions. Another resource developed for the Russian language is the RNC Sketches project aimed at creating patterns of word sequences based on the material from the National Corpus of the Russian Language, which offers syntactic models for word sequences with examples but does not provide any quotations. Sketch Engine is another tool that provides information on syntactic co-occurrence on the basis of text corpora for different languages, including Russian (Kilgariff et al. 2014; Khokhlova & Zakharov 2010).

Despite the fact that there are dictionaries of collocations, there exists, nevertheless, a need for resources that would describe the data more consistently. Currently there is no single system for the



Russian language which would allow scholars to obtain not only the corpus data, but also vocabulary information. At the moment there is no single system for the Russian language that would allow researchers to obtain not only corpus data, but also “reference” information on the vocabulary and behavior of lexical units. There are a number of unique and valuable lexicographic projects that describe the collocability, although in different ways. The purpose of our project is thus to consolidate the information on the collocability, which is presented in different ways in different sources. There are reference web-sites (for example, [slovari.ru](http://slovari.ru) or [gramota.ru](http://gramota.ru)), which provide an opportunity to learn the meaning of a word and to look through dictionary entries, which also contain information about collocability. But at the same time, users may have difficulty in reading the articles, as the information on collocability can be represented both in the “phraseological” part, and also directly in the quotations themselves. When it comes to collocations, a full dictionary entry with explanation is not always necessary, but rather examples of real data, appropriately designed and accompanied by values of the correctness or frequency of use in speech. Therefore, at the moment there is no such system, which would combine “reference” data from a lot of recognized sources, as well as real case examples. Such a system could be in demand both by ordinary users (for example, studying the Russian language) and by specialists. The given goal can be achieved right now, when large data sets are available (large text corpora), along with the software tools and computational power needed to process them.

### 3 Russian Collocations Database

#### 3.1 Evaluation of Collocability

The task of creating corpora that comprise large amounts of data has a long history. Researchers have long been attracted by the opportunity to test their hypotheses on quantitatively new material, but only with the advent of new technologies has this been practical. Numerous works discuss different approaches to the automatic collection of material from the web (see, for example, Kilgarrieff and Grefenstette (2003), Belikov et al. (2012), among others) and the creation of large text corpora. The threshold of 1 million tokens or even 100 million tokens has already been passed, and a number of papers discuss the merits of such large corpora (Belikov, Selegey, & Sharoff 2012; Benko & Zakharov 2016).

Along with the existence of active and passive vocabulary, the concept of active and passive dictionaries has also been introduced (Active Dictionary of the Russian Language 2014: 5-7). The first covers the needs of speaking and producing texts (*ibidem*). Thus, it gives a “deeper” idea of lexical units, describing not only their meanings, but also syntagmatic and paradigmatic properties. While the latter type of dictionary is aimed at covering as much material as possible, including low-frequency lexis. Large corpus of texts can be used to obtain data on collocability and its full representation within the paradigm of the active dictionary approach. The method used in the development of the Tolkovo-combinatorial dictionary (Melchuk & Zholkovsky 1984) showed the possibility of a structural approach to the description of collocability on the example of lexical functions.

When processing a large amount of data, it is difficult to use only a “manual” approach. Researchers thus work with automatic methods, which may include a quantitative approach, a rule-based approach, and a combination thereof. There are various statistical metrics for evaluating collocability. In other words, we are talking about the statistical non-randomness of word combinations, which can be evaluated quantitatively. Thus, some stability inherent to lexical units can be calculated, which allows them to be put on a scale: from free combinations to phraseological structures. Both statistical methods (including machine learning) and rules-based approaches can be used. In total, there are more than 80 measures to assess the strength of the relatedness of word combinations

(Pecina 2005). Not all of these were tested on language data, but the most frequently mentioned ones in the literature (MI, t-score, Dice, log-likelihood, chi-square) have proved successful when working on material from various different languages, including Russian. To evaluate the results of automatic collocation detection, both data from lexicographic sources are used (see, for example, Khokhlova (2008)) and the results of experiments with native speakers (Pivovarov et al. 2017). At the same time, there should be a lot of reference data to cover as many automatic results as possible, in order to evaluate them.

### 3.2 Representation of Information on Collocability

To build the gold standard, at the present stage we selected four explanatory Russian dictionaries (*Dictionary of Contemporary Literary Russian Language* 1948-1965; *Dictionary of the Russian Language* 1981-1984; *Big Academic Dictionary of Russian* 2004-2018; Kuznetsov 2014) and two specialized ones (Denisov & Morkovkin 1983; Borisova 1995). The given dictionaries differ in their representation of collocations and their example coverage. Explanatory dictionaries implement various ways to represent the information on combinatorial restrictions. One can find set phrases not only in special sections of the entries but also in the examples, sayings and quotations. The entry structure can also vary and depends on a dictionary. For example, the diamond symbol  $\diamond$  is used to designate set expressions and phraseological units in the *Dictionary of the Russian Language* 1981-1984, while these are indicated in the *Big Academic Dictionary of Russian* 2004-2018 with a tilde symbol.

We have collected collocations from the phraseological section in *Dictionary of the Russian Language* 1981-1984; the whole list comes about 13,000 examples (while the vocabulary list has more than 80,000 words). In the dictionary by Borisova (1995) collocations are structured according to their semantics and represented with a font. The analysis showed that both explanatory dictionaries (the *Dictionary of Contemporary Literary Russian Language* 1948-1965 and *Big Academic Dictionary of Russian* 2004-2018) overlap with each other in their representation of collocations.

One single format implies that part-of-speech tags will be assigned to all the extracted collocations, as well as information as to in which dictionaries they were described.

In order to obtain data on co-occurrences in the Russian language, we process the Araneum Russicum Maximum corpus (with about 15 billion words), which was created automatically and is based on web texts of different genres, and is one of the largest collections of Russian texts (Benko & Zakharov 2016). We use a statistical approach for automatic extraction of word combinations from corpora that implied several association measures (t-score, MI, log-likelihood). We focused our attention on the bigrams within the span [-1; 1] from the node. Thus the following models were extracted: noun + verb, verb + noun, adjective + noun, noun + noun etc.

We analyzed phraseological sections of the entries marked with special symbols and extracted data from them. Then we merged two lists, hence there are three categories of collocations: 1) the overlapping collocations that have both references to dictionaries and statistical values (see Table 1); 2) collocations described in the dictionaries but not extracted from the corpus (as they can be longer than bigrams); 3) collocations that were not found in the dictionaries. In our study we focus on the first and second groups of word combinations.

Table 1 gives an example of collocations from the gold standard for the headword “*nadezhda*” (‘hope’) described in the dictionaries. The second column indicates the dictionaries that list a collocation. One can see that only one collocation is present in the entries of all the dictionaries (“*pitat nadezhdu*” ‘to nourish hopes’), while other phrases were described in fewer lexicographic resources. As noted in Section 2, the coverage of specialized dictionaries can be even wider than that of

explanatory dictionaries. We introduced a simple metric called “dictionary index” that is given in the third column. It shows the number of dictionaries that include the collocation. It can vary within 0 and 6 (0 means that the collocation was not listed in any dictionary but nevertheless was extracted from the corpus). Large values of the index imply that the collocation is reproduced in speech quite often, and thus should be learned by heart (if we speak about students of Russian). The last three columns show the values of the association measures (t-score, MI, LL).

Table 1: Representation of the results for the headword “*nadezhda*” (‘hope’).

Collocation	Dictionaries <sup>1</sup>	Dictionary Index	Syntactic Structure	t-score	MI	LL
“ <i>pitat’ nadezhdu</i> ” ‘to nourish hopes’	1, 2, 3, 4, 5, 6	6	V+N	43,445	7,466	15991,086
“ <i>podavat’ nadezhdy</i> ” ‘to give hopes’	1, 2, 3, 4, 6	5	V+N	73,600	7,208	44053,620
“ <i>podavat’ nadezhdu</i> ” ‘to give a hope’	1, 3, 5, 6	4	V+N	73,600	7,208	44053, 620
“ <i>pitat’ nadezhdy</i> ” ‘to nourish hopes’	1, 2, 3, 4	4	V+N	43,445	7,466	15991,089
“ <i>nadezhda na</i> ” ‘hope on’	1, 3, 5	3	N+Prep	405,451	3,796	682980,452
“ <i>v nadezhde</i> ” ‘in hope’	1, 2, 3	3	Prep+N	229,645	1,842	118188,997
“ <i>vozlagat’ nadezhdy</i> ” ‘to pin hopes’	2, 4, 6	3	V+N	67,042	10,257	55472,800
“ <i>vselyat’ nadezhdu</i> ” ‘to inspire a hope’	5, 6	2	V+N	75,031	11,344	78560,080
“ <i>poslednyaya nadezhda</i> ” ‘last hope’	3, 5	2	Adj+N	109,620	4,753	60613,064
“ <i>vozlagat’ nadezhdu</i> ” ‘to pin a hope’	3, 5	2	V+N	67,042	10,257	55472,800
“ <i>vyrazhat’ nadezhdu</i> ” ‘to express a hope’	5, 6	2	V+N	77,419	7,599	51871,075
“ <i>s nadezhday</i> ” ‘with hope’	1, 3	2	Prep+N	148,848	2,044	51765,512
“ <i>ostavlyat’ nadezhdu</i> ” ‘to give up a hope’	5, 6	2	V+N	64,155	5,806	25996,563
“ <i>lelyat’ nadezhdu</i> ” ‘to cherish a hope’	5, 6	2	V+N	36,618	9,744	15561,769

It can be seen that the dictionaries list not only collocations but also constructions and colligations. At the present stage of the study we deal with lemmatized word combinations. This is a restriction if it comes to Russian, as certain phrases are used in a certain morphological form (e.g. “*tret’yego dnya*” ‘the day before yesterday’) and such preferences should be studied separately from other forms. Aspectual verb forms we considered as one item (“*podavat’*” ‘to give’ vs “*podat’*” ‘to give’). The examples (see Table 1) suggest that the same headword can be used in both singular and plural forms in a collocation, but these phrases are not equally presented in the dictionaries (cf “*vozlagat’ nadezhdy*” ‘to pin hopes’ and “*vozlagat’ nadezhdu*” ‘to pin a hope’). At the moment the collocations have the same statistical measures, but if we distinguish between word forms they will differ accordingly.

The above-mentioned second group of collocations has the following examples: “*obmanyvat’ sebya*

<sup>1</sup> Here we use the followings symbols: 1 (*Dictionary of Contemporary Literary Russian Language 1948–1965*); 2 (*Dictionary of the Russian Language 1981–1984*); 3 (*Big Academic Dictionary of Russian 2004–2018*); 4 (Kuznetsov 2014); 5 (Denisov & Morkovkin 1983); 6 (Borisova 1995).

*nadezhday*” ‘to disappoint oneself with a hope’, “*teshit’ sebya nadezhday*” ‘to please oneself with a hope’, etc. They do not have any statistical evaluation as the trigrams were not extracted from the corpus. But nevertheless the given collocations will be added to the gold standard.

The Russian Collocations Database is already partially available upon request on the web (<http://collocations.spbu.ru>), and includes information about lexical collocations with statistical evaluations. For our research we have developed a special database to store pairs of collocated words and their correlation values according to various collocation metrics. The database is implemented by means of the MySQL engine and consists of three main tables:

- words table;
- collocations table;
- metrics table.

The gold standard will help to distinguish between different types of collocations, e.g. high-frequency and specialized phrases.

## 4 Conclusion and Further Work

The paper describes work in progress. At the present stage the database includes automatically extracted collocations from a large web corpus. The collocations are marked according to their presence in the Russian dictionaries (gold standard).

Further work will be focused on extraction of collocations from quotations and other examples used in the entries, as they can contain significant data. Moreover, the analysis confirmed a need to find correlation between the ranking of collocations from the gold standard and their statistical coefficients.

The results of this research project can be used in courses on lexicology, morphology, and syntax of the Russian language; they will be helpful for compiling dictionaries and grammar books, as well as for teaching Russian. The proposed resource will also be useful for studying Russian as a foreign language. The obtained results can be used for machine learning in programs connected with automated language processing, for instance, in systems for automatic clustering of word combinations and disambiguation.

## References

- A Dictionary of the Russian Language* [Slovar’ russkogo yazyka v 4 tomakh]. (1981–1984). Yevgen’yeve, A. P. (ed.-in-chief). Vol. 1–4, 2nd edition, revised and supplemented. Moscow: Russkij yazyk.
- Active Dictionary of the Russian Language* [Aktivnyy slovar’ russkogo yazyka]. (2014–2017). Apresyan, Ju. D. (ed.) Vol. 1-3. M.: Yazyki slavyanskoy kul’tury.
- Belikov, V.I., Selegey, V.P., Sharoff, S.A. (2012) Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k projektu General’nogo internet-korpora russkogo yazyka (GIKRYa)]. In *Computational linguistics and intellectual technologies*. Vol. 11 (18). Moscow: Izd-vo RGGU, pp. 37-49.
- Benko, V., Zakharov, V. (2016). “Very large Russian corpora: New opportunities and new challenges.” *Computational Linguistics and Intellectual Technologies* 15:22, pp. 79–93. Moscow: Izd-vo RGGU.
- Big Academic Dictionary of Russian* [Bolshoy akademicheskiy slovar v 30 tomakh]. (2004–2016). Moscow-Saint-Petersburg: Nauka.
- Biriuk, O. L., Gusev, V. Iu., Kalinina, E. Iu. (2008). *Dictionary of Russian Abstract Nouns’ Verbal Collocability. A Dictionary based on the Russian National Corpus* [Slovar’ Glagol’noi Sochetaemosti Nepredmetnykh Imen Russkogo Iazyka. Slovar’ na osnove Natsional’nogo Korpusa Russkogo Iazyka]. Accessed at: <http://dict.ruslang.ru> [15/05/2018].

- Borisova, E. G. (1995). *A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords* [Slovo v tekste. Slovar' kollokatsiy (ustoychivyykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov]. Moscow.
- Denisov, P. N., Morkovkin, V. V. (comp. and ed.-in-chief), (1983). *An Academic Collocation Dictionary of Russian* [Uchebnyi slovar' sochetaemosti slov russkogo iazyka]. Moscow: Russkij jazyk.
- Dictionary of Contemporary Literary Russian Language* [Slovar sovremennogo russkogo literaturnogo yazyka v 17 tomakh], (1948–1965). Chernyshev, V.I. (ed.). Moscow-Leningrad: Izd-vo Akademii nauk SSSR.
- FrameBank*. Accessed at: <http://framebank.ru/> [15/05/2018].
- Khokhlova, M. (2008). Evaluation of Methods for Collocation Extraction [Eksperimental'naja proverka metodov vydeleniya kollokatsij]. In *Slavica Helsingiensia 34. Instrumentarij rusistiki: Korpusnye podhody*. Eds. A. Mustajoki, M.V. Kopotev, L.A. Birjulin, J.J. Protasova. Helsinki. pp.343–357.
- Khokhlova, M., Zakharov, V. (2010). Studying Word Sketches for Russian. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (Valetta, Malta, 19–21 May 2010). Eds. Nicoletta Calzolari (Conference Chair), Khalid Choukri, et al., pp. 3491–3494.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, 1, pp. 7–36.
- Kilgariff, A., Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. In *Computational Linguistics*, 29 (3), pp. 333–347.
- Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L., Yangarber, R. (2015). CoCoCo: Online Extraction of Russian Multiword Expressions. In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing* (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, pp. 43–45.
- Kustova, G.I. (2008). Slovar' russkoj idiomatiki. Sochetaniya slov so znacheniyem vysokoy stepeni. [Dictionary of Russian Idioms: Combinations of Words with the Meaning of High Degree]. Accessed at: <http://dict.ruslang.ru> [15/05/2018].
- Kuznetsov, S. (ed.). (2014). *Large Explanatory Dictionary of Russian* [Bolshoy tolkovyi slovar russkogo yazyka]. Norint, St. Petersburg.
- Lexicograph*. Accessed at: <http://lexicograph.ruslang.ru/> [15/05/2018].
- Lyashevskaya, O. (2010). Bank of Russian constructions and valencies. In *LREC 2010*. Malta, Valletta, May 19–21, 2010.
- Mel'čuk, I., Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian* [Tolkovo-kombinatornyj slovar russkogo jazyka]. Vienna.
- Pecina, P. (2009). Lexical Association Measures. Collocation Extraction. Prague: Institute of Formal and Applied Linguistics.
- Pivovarova, L., Kormacheva, D., Kopotev, M. (2017). Evaluation of collocation extraction methods for the Russian language. In *Quantitative Approaches to the Russian Language* (ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki). London, New York: Routledge. pp. 137–157.
- Russian National Corpus*. Accessed at: <http://ruscorpora.ru> [15/05/2018].

## Acknowledgements

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-2513.2018.6).





# Rethinking the Role of Digital Author's Dictionaries in Humanities Research

**Margit Kiss<sup>1</sup>, Tamás Mészáros<sup>2</sup>**

<sup>1</sup>*Institute for Literary Studies, Hungarian Academy of Sciences*, <sup>2</sup>*Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics*

E-mail: [kiss.margit@btk.mta.hu](mailto:kiss.margit@btk.mta.hu), [meszaros@mit.bme.hu](mailto:meszaros@mit.bme.hu)

## Abstract

Although it is true that computers make one's work significantly easier during some of the work phases of creating an author's dictionary, our goal is to completely revise conventional computer-based dictionary work processes as well as to extend the possible range of applications for author's dictionaries. In our study we examine the role of the digital author's dictionaries in humanities research. We developed the extended digital author's dictionary based on the oeuvre of Kelemen Mikes (1690-1761). We present the advantages of the extended digital author's dictionary and demonstrate its benefits in literary and linguistic research. The main advantages of our system include easily accessible up-to-date encyclopedic information and the improved efficiency of historical text analysis methods. The benefits of the extended functions of the digital author's dictionary help scholars answer more specialized and complex research questions, and reconsider expectations towards the new generation of author's dictionaries.

**Keywords:** author's digital dictionary, historical text analysis, Kelemen Mikes

## 1 Introduction

The number of author's dictionaries available today is quite significant, and is continually increasing; still, the literature on lexicography pays little attention to this topic (Karpova 2011; Mattausch 1990). Paper-based author's dictionaries play a limited role in humanities research, their function is essentially limited to contextual interpretation and presenting the linguistic functions of the given word in the text, which makes it difficult to make it a source of comprehensive, complex analyses concerning a larger lexicon. In contrast, the digital environment is ideal for the author's dictionary, since structured data handling and data extraction have a number of unexploited advantages compared to paper-based volumes, and this strengthens and extends the function which the author's dictionary was originally designed to fulfill. We have two aims: on the one hand, we would like to introduce the new generation of digital author's dictionaries, for which we have created the extended digital dictionary through making use of state-of-the-art opportunities in information technology; on the other hand, we would like to sketch the specific application areas to demonstrate the methodological changes and new possibilities this type of dictionary creates in humanities research. In our study, through presenting a digital author's dictionary under preparation based on the oeuvre of an 18<sup>th</sup>-century Hungarian writer, we describe how we have rethought the role of the digital author's dictionary.

## 2 The definition and development of the author's dictionary

The main aim of the author's dictionary is to show elements of language use, style, vocabulary and expressions through presenting a significant part of an author's oeuvre, or even its complete vocabulary.

Author's dictionary: a type of reference work which provides information on the vocabulary of a specific author. The material is usually based on a text corpus of one, several or all of the works of the author, and often presented in alphabetical order, with examples or contexts (but not definitions) of the words cited (Hartmann & James 2001: 10).

Many author's dictionaries have been created throughout the world, such as for the works of Johann Wolfgang Goethe, Friedrich Schiller, Thomas Mann, Bertolt Brecht, Jean Racine, Victor Hugo, Dante, Henrik Ibsen, Alexander Pushkin, Mikhail Lermontov, William Faulkner, Mark Twain, Ernest Hemingway, and William Shakespeare. Not only the work of writers of fiction, but also other significant authors' work may constitute the basis of such dictionaries, e.g. the *Kant Dictionary*, the *Hegel Dictionary*, the *Historical Dictionary of Leibniz's Philosophy*, etc. (see Karpova 2011).

The precursors of the author's dictionary as defined today in terms of genre are concordances and glossaries (Karpova 2011: 1-7). While the aim of a concordance is orientation and search within a text (e.g. in the Bible), glossaries help the reader interpret specialist texts (e.g. the writing of ancient Greek and Roman authors). The next milestone in the development of the author's dictionary as a genre happened when the emphasis shifted from sporadic word analyses to complete vocabulary analyses, to processing the whole authorial lexicon (e.g. Clarke 1846). This change in approach marked a generic determination for the author's dictionary that is still valid today.

In 20<sup>th</sup>-century dictionary literature previous methods used in the processing of authorial oeuvres were replaced by indexing, preparing concordances, and interpretation with stylistic and grammatical qualifications: emphasis shifted from sporadic interpretation to methodical systematization. The most important task of the author's dictionary became the lexicographic processing of the entire oeuvre, by supplying context-dependent interpretation. The function aiding textual understanding became a scientific genre of its own. The complete author's dictionary is also a significant qualitative change, a basic generic characteristic. These functional dictionaries present the words in their real context, based on principles of general grammar, semantics, and stylistics. Being exhaustive implies processing every occurrence of every word. Due to space limitations and because it requires substantial manpower, in the case of paper-based dictionaries this aim is often very difficult to achieve, so there is often a need for solutions that limit size or create a complex, complicated reference system. Influential examples of this interpretative-qualifying complete dictionary type are the Pushkin (Vinogradov 1956-1961), Goethe (GWB 1978-), Ibsen (Ibsen-Ordbok 1958), and Petőfi (J. Soltész et al. 1973-1987) dictionaries. These were typically created slowly, with a detailed entry structure, a rich data structure, providing the occurrence data, with a detailed, sophisticated meaning description, stylistic qualifications, and phraseological references, and all made with virtually no computer support.

With the improvement of technology and the spread of information technology (IT), another type of author's dictionary also appeared (today this has a well-established tradition), which, instead of semantic-stylistic finishing, represented a more formal, grammar- and IT-centered approach. Among the first computer applications in connection with authorial texts was the preparation of concordance lists (Busa 1980). At the dawn of computer-based lexicography (in the 1950s and 1960s), analyzing the works of Thomas Aquinas, Kant, Shakespeare, Goethe, Racine, Baudelaire, Dante, and others on the basis of corpora became the center of attention, in an index and concordance style (Gouws et al. 2013: 974). In the 1970s in Hungary the work of Ferenc Papp also strengthened this line of work through the computer-aided processing of Ady's oeuvre, and by emphasizing the importance of concordance as a raw material of author's dictionaries (Mártonfi 2014). Using the concordance list generated by the machine as a basis, we can create works resembling dictionaries with limited editorial work, which are a transition between concordance lists and author's dictionaries. These works, in contrast with previous ones, can be prepared with much less effort; at the same time, they also provide less in terms of their lexicological-semantic finesse. A concordance list that is the basis of dictionaries

created with computer assistance is often a product on its own and an interesting source of research, but as intermediate textual material between corpus and dictionary it also makes lexicographers' work significantly easier (e.g. OSS 2003-2018).

The IT solutions that are widespread today in the creation of author's dictionaries have made it substantially easier to create dictionaries and thus sped up the process (Čermák & Cvrček 2010). The spread of computers in lexicology work has caused changes in several areas (Gouws 2014). The easy and fast digitalizing of texts, using text recognition processes, creating and using text corpora and databases, producing automatic concordances, and using specialized software for editing entries, are all due to the application of IT, and are now of substantial help in the process of creating a dictionary, with such technologies replacing manual tools. Storing texts and handling data has become easier, and this change has not only made dictionary makers' task easier, but also changed the users' expectations towards dictionaries. These IT innovations have taken over some areas associated with paper-based dictionaries, and acquired new functions (Lew 2013).

IT tools, however, have even more to offer for author's dictionaries than what we are used to today, pointing beyond storing data and handling it in different ways. In addition to their role in the process of dictionary making, such tools can play a part in research and the different ways of processing data. They can integrate and represent the dispersed knowledge elements (linguistic, stylistic, historical, etc.) located in different places, and can make them processable and further applicable. While linguistic corpora have been the main tool in linguistic research for a long time, their capabilities for dictionaries remain underestimated (Apresjan & Mikulin 2016). By using the opportunities of a state-of-the-art set of tools provided by IT, we need to rethink how else we can use digital author's dictionaries beyond the usual applications, which also means a rethinking of their possible roles. In the future this change will result in creating author's dictionaries with greater added value.

### 3 The extended author's dictionary

Although it is true that computers make one's work significantly easier during some of the work phases of creating an author's dictionary, our goal is to completely revise conventional computer-based dictionary work processes as well as to extend the possible range of applications for such dictionaries. We set out to exploit modern information technology tools and prepare a type of digital author's dictionary which adjusts to the new opportunities and demands, and supports humanities research in a more efficient way than usual. Our main aims cover three major areas.

- In terms of a quantitative change, we are preparing a complete dictionary: we present every single word and sample sentence, without any space limitations.
- We provide much more information content for the individual entries than usual, which can be stored and searched in a structured manner.
- We link the individual entries with external data sources so that we can link the relevant knowledge elements digitally stored elsewhere to the particular entry.

We have implemented this new type of author's dictionary by processing the oeuvre of Kelemen Mikes, who played an influential role in 18<sup>th</sup>-century Hungarian prose literature. With its 1.5 million words of text, his work stands out among Hungarian authors' oeuvres, and his complete oeuvre is not part of the corpus of the comprehensive dictionary (HHC). The Mikes dictionary is based on the textual material of the complete critical edition, and this is the largest lexicon based on which a Hungarian author's dictionary has been created. The Mikes dictionary is the first complete Hungarian digital author's dictionary (Kiss 2012). It presents all the author's words with all the sample sentences. In a printed format the dictionary would comprise approx. 20,000 pages.

The use of paper-based dictionaries essentially concentrates on interpreting a given authorial text location, and search capabilities are basically limited to individual entries. The structure of the entries is rigid, and word interpretations often contain too much or too little explanation. Publishing in a printed volume makes it necessary to apply a range of limiting functions, which may include limiting the sample material, establishing a complicated reference system, or other space-saving strategies. In a traditional paper-based author's dictionary the typical entry structure of a geographical name is the following:

**Konstantinápoly** tulfn 2 | -ß 1 | -nak 1  
(földrn) 'nagyváros a Boszporusz partján, a mai Isztambul Törökországban': (mint a keletrómai birodalom fővárosa, Bizánc:) Konstantinápolynak ment Botond keletre, S szörnyű taglójával kapuját betörte (LV/3 : 307) | (mint a török birodalom fővárosa:) elzúgtak a kemény csaták, Mellyek Konstantinápoly tornyain A büszke félholdat megingaták. (851/3 : 230)

Figure 1: *Konstantinápoly* in a paper-based author's dictionary.

From this entry we can learn the following pieces of information concerning the given textual location: the headword of the place name, authorial form variations, number of occurrences, paradigmatic forms, a one-sentence interpretation concerning place location, and one or two selected sample sentences for illustration.

However, much more semantic content can fit into the computer-based representation of an author's dictionary than in a paper-based format. In order to provide the dictionary with as much semantic information content as possible, we extended the dictionary in two ways: on the one hand, we allowed the dictionary maker to include lexicographical knowledge about the entries; on the other hand, we enriched the dictionary by linking it to already existing external knowledge sources. We extended the entries of the Mikes dictionary in such a structured way that it contains extra knowledge suitable for computer representation, search, and processing. We completed this extension in the areas listed below. We linked all the sample sentences to each authorial word form, in some cases several thousands of them. From the individual sample sentences we can get to the wider textual context, since the sample sentences are linked with the corpus in the dictionary. We supplemented the word forms with the form variation of Mikes, as well as the contemporary headword. Beside the modern dictionary headwords, we also defined so-called reference headwords where necessary, so we can extend the searchability of the words by adding different headword variations and by presenting their connections to other headwords. Types of these include: reference to headwords within the dictionary based on etymology, e.g. *hívség* - *hűség*; reference to current Hungarian (or foreign-language) equivalents of words functioning as foreign words and loanwords, e.g. *decembris* - *december*; presenting the headword variations also existing in contemporary Hungarian together, e.g. *caritas* - *karitás*; presenting the proper name variations also existing in contemporary Hungarian together, e.g. *Kroiszosz* - *Krózus*; presenting the Hungarian equivalents of Latin proper names, e.g. *Casimirus* - *Kazimír*; associating archaic words with their current form, e.g. *milliom* - *millió*; presenting connections with foreign words, e.g. *clanicus* - *klinikus*; presenting the Turkish originals of the Turkish words used by Mikes, e.g. *cház oda* - *has oda*. Proper names, words rooted in foreign languages, words invented by Mikes (the words which are not included among the headwords of other dictionaries or in the text corpus of the comprehensive Hungarian dictionary), and those headwords which differ from the form used by Mikes, have received further annotations for content. We have attached the part-of-speech category to every occurrence of a word. The dictionary also assigns semantic information to named entities. Geographical names are supplemented by further metadata, for example, geographical coordinates, which, besides interpretation, provide a clear point of reference for determining exact location.



We supplemented the extra knowledge entered by the dictionary maker with a second pillar by linking already existing knowledge elements available in external data sources: adding critical notes, and connecting to external databases such as DBpedia using the LOD technique. We have created links to these data sources, and thus we have made other specialist knowledge that is available in external sources accessible to the users of the dictionary. During the linking we have lifted a narrow set of knowledge from DBpedia and attached it to the relevant dictionary entries with the help of an RDF graph. This form of extension thus does not entail the recording of a specific piece of information content in an entry, as in the first case, but links an external data source to an entry. For example, by extending an entry with a DBpedia identifier we can get from the dictionary entry to a source containing encyclopedic knowledge, where the user can find a significant amount of additional information. This linking of dictionary entries and external databases may be created automatically, but in several cases it requires manual revisions (e.g. to perform semantic disambiguation).

The other element of extending from an external data source relates to linking those information contents to the entry which come from the notes made to the critical edition. The notes of the critical edition provide help during the interpretation of the text. They help the reader with information such as explanations of the less well-known proper and geographical names that appear in the work, as well as those of obsolete and dialectical words, and not commonly known phraseological expressions, among others. It extends to the historical aspects of the given work, points out the sources of the author's views and philosophical influences, and refers to the work's genesis.

### rodosta - 19 9bris 1724

JNúndkor pirongat. ked leveleiben hogy meg nem iron kediek, mint tiltjuk itt az idet, rigasságban, csak suhajjunk, olyan jó kedünk van, hogy majd meg halunk bunkban, mit k egyebet, ha jó volnék jobban tölthetném, mert arra elég jó példát ad a mi urunk, de rossz va attól tartok, hogy az is ne maradjak, de talán az idő okosabbá tesszen, vagy akarom vag, kétslen való okosságnak pedig semmi érdeme nincsen, akkor volna valami kis érdemünk, chebünk a meg tiltat gyűmelemből, de nem eszünk, és nem akar, amideen arra nem nagy egy vagyon, de már most ha csak egy néhány napig is, okosab leszek, mert tegnap idő érkezett a érsejre, itt fog egy néhány napot tölteni, és adlyg rea tartjuk magunkat, valamint a kompa aszszony, edes nénem ez után hintet, mit kellett küldeni, arégi püspökök pedig az nehezteltek volna, mivel az előtt feképpen anap keleti országokban, apüspökök, közü gyalog jártak, nem szollok az oregeköl, akik számárna, vagy észre utenek, a görög an egyházban emúndkor így való szokásban, mert a püspökök mint hogy csak közinséges valának, azért nem vágytanak a fele alkalmatosságokra, constantinapolyban annyi pátriárkák közöt, talán csak egy volt, akiről mondják hogy hét száz paripát tartot, act

#### Constantinapolyban

— TL.1

Type: B,TÖ

Előfordul még Constantinapoly (168, 183. lev.), Constancinapoly (25, 41 Constantinopolis; Ottoman Turkish: قسطنطينية, Kost'an(i)nye) was the capital city of the Roman/Byzantine Empire 123, 127, 138, 178, 179, 204. lev.), Constāncinapoly (115, 123, 200. lev (330–1204 and 1261–1453), and also of the brief Latin (1204–1261), and the later Ottoman (1453–1923) empires.

### Konstantinápoly

Megjegyzés: ok

#### Constancinapoly

Constancinapoly  
Iora ülén, a csauz, a Constancinapoly mellett lévő retnek avégin, egy (TL.37)  
Constancinapoly 5 (ML.305)

#### Constancinapolyban

lovát, és a pápa azon Constancinapolyban megyen, és onnét viszá küldi (TL.95)  
szerencsétlen lévén hadakozása, viszá tere Constancinapolyban, a hadának negyec hogy a szegény fejdelem testit, Constancinapolyban vigyék, azért tegnap este, egy amurates látván hogy mlsoda szükséges Constancinapolyban való menetele, magá semmi szándékát nem látta volna Constancinapolyban való igyekezetéről, azért nem portának nagy készületin, követet küldének Constancinapolyban, de a császár olyan erre való nézve követeket küldte Constancinapolyban, a többi közöt vala Cardinals

#### constantinapoly

constantinapolyban  
a jesuitákhoz viszik bé innét, constantinapolyban., tudom hogy őt lesz két (TL.79)



Browse using

Formats

#### About: Constantinople

An Entity of Type: city, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Constantinople (Greek: Κωνσταντινούπολις Konstantinoúpolis or Κωνσταντινούπολη Konstantinoúpoli; Latin: Elİflordul még Constantinapoly (168, 183. lev.), Constancinapoly (25, 41 Constantinopolis; Ottoman Turkish: قسطنطينية, Kost'an(i)nye) was the capital city of the Roman/Byzantine Empire 123, 127, 138, 178, 179, 204. lev.), Constāncinapoly (115, 123, 200. lev (330–1204 and 1261–1453), and also of the brief Latin (1204–1261), and the later Ottoman (1453–1923) empires.

Figure 2: Dictionary headword (top right) with full text citation (top left), the attached critical annotation (bottom left), and related DBpedia entry (bottom right).

Figure 2 shows a sample headword *Konstantinápoly* 'Constantinople'. It contains all its writing variations, word forms, citations from the corpus in their extended textual context. All sample sentences are assigned a part-of-speech designation. The dictionary also marks that it is a geographical name, which is part of a country and region, as well as its exact coordinates. Linking to DBpedia we can access a huge amount of information that would not be possible to include with a traditional, paper-based entry structure. The notes on *Konstantinápoly* (Figure 2, bottom left) are attached to the entry as an external data source, and provide detailed information about Kelemen Mikes' personal attachment to Constantinople.

The result is a digital author's dictionary which supports humanities research by meeting newly emerged opportunities and needs. We exchanged the tools designed to reduce size and content to

exploit opportunities for extension, while the chopped-up content found in different data sources was linked within the entry in order to uncover connections between distant textual locations. We supplemented the traditional entry structure with information from other data sources or databases. With the help of this extended method we have created a structured knowledge base. By being linked to other databases, the dictionary now has new applications that go beyond describing the author's language. Reaching beyond a contextual interpretation that explains a specific textual location, and building a knowledge base, encyclopedic knowledge also becomes easily accessible and analyzable with computer tools.

## 4 Applications of the extended author's dictionary

### 4.1 Dictionary as the object of analysis

The extended digital author's dictionary not only constitutes a new type of dictionary, but beyond that the dictionary itself also becomes an object of further analyses. With its help we can choose research directions that have not been covered before, the analysis of which was not possible due to the lack of tools or opportunities. Our system can help researchers perform new kinds of analyses on the author's oeuvre to uncover novel linguistic and literary results. These may include a large-scale analysis of the author's language, a statistical survey of different linguistic features, computing statistics about words with certain kinds of annotations, comparing various author's dictionaries, and analyzing vocabulary similarities and influences or changes over time. Since this extended dictionary organizes the corpora in a very different way from the text source, its analysis may also uncover previously hidden connections or similarities between distant pieces of text.

The Mikes dictionary was part of a comprehensive analysis of language history in which a grammaticalization process that had taken place over approximately 500 years was researched. During this period the complemented participial form *mondván* started to become a conjunction, and through analyzing the morphological and syntactic structures collected from the Mikes dictionary one of the important stages in this change can be found (Dömötör 2013). Creating a digital dictionary also enables us to uncover previously unknown text creation processes in the author's oeuvre. We have uncovered textual connections between different, distant points of the oeuvre, which would be impossible to collect manually. In the Mikes corpus we have uncovered approximately 500 textual similarities and parallels, which, although they are not a complete textual match, present smaller variations between the different excerpts. We have automated the computer-aided listing of text migration from the historical text corpus, and we can access these without a targeted word search in the 18<sup>th</sup>-century text. From this we found that some excerpts may even appear seven times in different works of the author, and we have identified which works of the oeuvre are connected in this respect. With the help of the Mikes dictionary we have been able to reevaluate previous findings in the literature concerning his lexicon. In the Mikes oeuvre, translations, as compared to his own texts, did not play an important role in the literary canon. Analyzing approximately 10,000 entries shows that the lexicon of the translations and that of Mikes' own texts differ substantially, they are a match for a maximum of 30% of cases, and so they cannot be left out of period analyses. From the analysis of the lexicon it has also become apparent that in his translations Mikes took a more modern stylistic approach not only compared to his era, but also compared to his own works, since it is in the lexicon of his translations where we can discover progressive linguistic and stylistic changes. By the 19<sup>th</sup> century the meaning condensation capability of compounds is stronger, and the perfective function of prefixes becomes determinant, as does its word creating role. These progressive changes can already be detected from the lexicon of Mikes' 18<sup>th</sup>-century translations (Kiss 2016). While analyzing Mikes' hapax legomena, we uncovered the common etymological roots of two words (*gyaur*, *kaffer*), and their entire history

of meaning (Kiss 2017). Further results can be expected from using the digital dictionary in the analysis of excerpts of debated authorship, in uncovering the characteristics and reasons for changes in individual style and sets of expressions, as well as exploring parallels among other authors and eras. In the works of Mikes we encountered several unresolved philological-textological problems while creating the dictionary, which due to the size of the oeuvre and use of manual tools had so far been unresolved. We were previously not able to examine these issues with such thoroughness.

## 4.2 Using the dictionary during corpus analysis

As further added value, the digital dictionary also enhances the efficiency of different IT-supported processes of analysis. The digital author's dictionary is a reliable tool in standardizing the morphological variety of historical texts. A dictionary with a structured setup makes it easier to search for both historical and contemporary word forms, which makes it possible to avoid the difficulties of the morphological analysis of historical texts, and through its use we can carry out the normalization of historical texts with an uneven, unregulated writing style. Whether we search for a historical or a contemporary entry, or any reference entry within those, we will get to all the other forms found in the entry. Beyond this, the search is also aided by semantic information in the corpus. With the help of the dictionary and the knowledge base linked to it, if, for example, we determine as a search parameter that we would like to list the place names connected with Turkey in the dictionary, as a result we will receive all Turkish place names listed in the dictionary, including *Konstantinápoly*. Beyond this, we can also display the geographical places collected according to different parameters on a map – for example, if we want to know where Mikes wrote each of his letters.

As an additional benefit, the dictionary can improve the quality of computational stylometry analysis. In order to compare Mikes' own writing with his translations, and to analyze the author's idiolect based on the similarities and differences between his works, we have conducted stylometry analysis on the corpus. We have used the digital author's dictionary to improve the efficiency of the statistical analytical methods, in which during the lexically-based analysis the dictionary had a lemmatization role by normalizing the different historical form variations. During the computational stylometry analyses we supplemented the statistical analysis with a preparatory (text normalizing) phase using the dictionary, in order to enhance the effectiveness of the analysis. Running a multidimensional analytical process on the entire authorial lexicon, we could determine the similarities and differences between the historical texts more exactly (Kiss & Mészáros 2016). We can thus see the differences between the analyses of the original and the normalized texts.

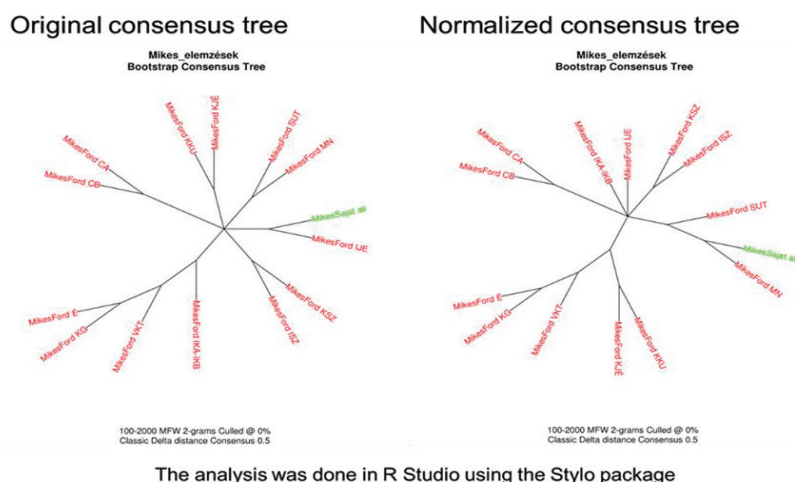


Figure 3: The results of the stylometry analysis: original [left], normalized [right].

## 5 Summary

If we do not consider the storage capability of large texts as the only opportunity among the advantages provided by information technology, but we also take advantage of the possibilities provided by the computer, we can change more than just the dictionary creation processes. With the extended digital dictionary we have created a type of dictionary which, extended to a knowledge base, makes it possible to access much more semantic content than before. The dictionary itself then becomes the basis of further analyses, as a new data source. With a synoptic, systematic analysis we can conduct discovery analyses for which we did not have the appropriate tools before. The methodology creates a completely new perspective and makes it possible to conduct badly needed tasks in basic research, to get a better awareness of linguistic, literary, historical, and cultural trends and processes.

The advantages that go with the extended functions of the digital author's dictionary help solve some more sophisticated, specific, and complex research problems than before. This leads us to reevaluate our expectations towards the new generation of author's dictionaries. Having rethought the role of digital author's dictionaries in the humanities, we have created an extended version of this type of dictionary, so that easily accessible, encyclopedic information content suitable for machine analysis can become even more readily accessible, and we have shown how we can enhance the efficiency of the analytical methods of historical texts with the help of our dictionary.

Most of the software tools developed during this project are open source, and they are available in the following GitHub repository: <https://github.com/mtwebit/dhmine>. The Mikes dictionary is also available on-line at <http://mikesszotar.iti.mta.hu/>, with a revised version at <https://dh.mit.bme.hu/mikes/>.

## References

- Apresjan, V., Mikulin, N. (2016). Dictionary as an Instrument of Linguistic Research. In T. Margalitadze & G. Meladze (eds.), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 224-231.
- Benkő, L. (1979). *Az írói szótár*. Budapest: Akadémiai Kiadó.
- Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. In *Computers and the Humanities*. 14(2), pp. 83-90.
- Čermák, F., Cvrček, V. (2010). Author Dictionaries Revisited: Dictionary of Bohumil Hrabal. In A. Dykstra, T. Schoonheim (eds.), *Proceedings of the XIV EURALEX International Congress 6-10 July 2010*. Leeuwarden/Ljouwert: Fryske Akademy – Afûk, pp. 595-598.
- Clarke, C. (1846). *The Complete Concordance to Shakespeare*. New York: Wiley and Putnam.
- Dömötör, A. (2013). Idéző szerkezetből diskurzuszjelölő elem: a mondván szerepei és története. In M. Csepregi et al. (eds.) *Grammatika és kontextus. Új szempontok az uráli nyelvek kutatásában III*, Budapest, ELTE, 20-30.
- Gouws et al. (2013). *Dictionaries. A International Encyclopedia*. Berlin, Boston: De Gruyter Mouton.
- Gouws (2014). Article Structures: Moving from Printed to e-Dictionaries. In *Lexikos 24*, AFRILEX-reeks/series 24, pp. 155-177.
- GWB (1978-). *Goethe-Wörterbuch*. Berlin-Brandenburgischen Akademie der Wissenschaften, der Akademie der Wissenschaften zu Göttingen und der Heidelberger Akademie der Wissenschaften (Hrsg.), Stuttgart: W. Kohlhammer.
- Hartmann, R.R.K., James, G. (2001). *Dictionary of Lexicography*. London and New York: Routledge.
- HHC. *Hungarian Historical Corpus*. Accessed at <http://www.nytud.hu/hhc/> [31/03/2018]
- Ibsen–Ordbok (1958). *Ibsen–Ordbok*. Ordforradet i Henrik Ibsene samlede Verker. Oslo.
- Karpova, O. (2011). *English author dictionaries the (XV<sup>th</sup>–XXI<sup>st</sup> cc.)*. Cambridge: Cambridge Scholars Publishing.
- Kiss, M. (2012). The Digital Mikes-Dictionary. In G. Tüskés et al. (eds.) *Literaturtransfer und Interkulturalität im Exil [...]*. Bern: Peter Lang Verlag, pp. 288-297.



- Kiss, M. (2016). „más értelmet adni ezeknek a szoknak”: Mikes Kelemen szóhasználatához. In R. Lengyel (eds.) *Nunquam autores, semper interpretes: A magyarországi fordításirodalom a 18. században*, Bp, MTA BTK, pp. 58-68.
- Kiss, M. (2017). Hitetlenek: mi köze a gyaur-nak a kaffer-hez? In *Magyar Nyelv*, 113(1), pp. 80-87.
- Kiss, M., Mészáros, T. (2016). Creating an extended author's dictionary to support digital literary research, In *DH Benelux 2016*, Luxembourg. Accessed at [http://www.dhbenelux.org/wp-content/uploads/2016/05/89\\_Kiss-Meszaros\\_FinalAbstract\\_DHBenelux\\_2016\\_long.pdf](http://www.dhbenelux.org/wp-content/uploads/2016/05/89_Kiss-Meszaros_FinalAbstract_DHBenelux_2016_long.pdf) [31/03/2018]
- Lew, R. (2013). From paper to electronic dictionaries: Evolving dictionary skills. In D. A. Kwary et al. (eds.) *Lexicography and dictionaries in the information age*, Selected Papers from the 8th ASIALEX International Conference, [Bali, 20-22 August, 2013], Airlangga University Press, pp. 79-84.
- Mártonfi, A. (2014). Számítógép és írói szótár – különös tekintettel a készülő József Attila szótárra. In *Magyar Nyelv*, 110(1), pp. 30-46.
- OSS (2003-2018). *Open Source Shakespeare*. Accessed at <http://www.opensourceshakespeare.org/concordance/> [28/03/2018]
- J. Soltész, K. et al (1973-1987). *Petőfi-szótár*. Budapest: Akadémiai Kiadó.
- Vinogradov, V. V. [виноградов, виКтор владимирович] (1956–1961). *Словарь языка Пушкина 1-4.*, Государственное Издательство Иностранных и Национальных Словарей, Москва.





## European Lexicographic Infrastructure (ELEXIS)

*Simon Krek<sup>1</sup>, Iztok Kosem<sup>1</sup>, John P. McCrae<sup>2</sup>, Roberto Navigli<sup>5</sup>,  
Bolette S. Pedersen<sup>6</sup>, Carole Tiberius<sup>4</sup>, Tanja Wissik<sup>3</sup>*

<sup>1</sup>*Jožef Stefan Institute*, <sup>2</sup>*Insight Centre for Data Analytics, National University of Ireland Galway*,  
<sup>3</sup>*Austrian Academy of Sciences*, <sup>4</sup>*Dutch Language Institute*, <sup>5</sup>*Sapienza University of Rome*, <sup>6</sup>*University of Copenhagen*

*E-mail: simon.krek@ijs.si, john@mccr.ae, iztok.kosem@ijs.si, tanja.wissik@oeaw.ac.at,  
carole.tiberius@ivdnt.org, navigli@di.uniroma1.it, bspedersen@hum.ku.dk*

### Abstract

In the paper we describe a new EU infrastructure project dedicated to lexicography. The project is part of the Horizon 2020 program, with a duration of four years (2018-2022). The result of the project will be an infrastructure which will (1) enable efficient access to high quality lexicographic data, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. One of the main issues addressed by the project is the fact that current lexicographic resources have different levels of (incompatible) structuring, and are not equally suitable for application in Natural Language Processing and other fields. The project will therefore develop strategies, tools and standards for extracting, structuring and linking lexicographic resources to enable their inclusion in Linked Open Data and the Semantic Web, as well as their use in the context of digital humanities.

**Keywords:** lexicography, research infrastructure, natural language processing, computational linguistics, semantic web, artificial intelligence, linked open data, digital humanities

## 1 Introduction

Reliable and accurate information on word meaning and usage is important in the information-driven society of the 21<sup>st</sup> century. In most European countries, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale has long been limited. Consequently, the lexicographic landscape in Europe is rather heterogeneous. Firstly, it is characterized by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts, prohibiting reuse of this valuable data in other fields. Secondly, there is a significant variation in the level of expertise and resources available to lexicographers across Europe. Both issues contribute to the fact that the data from these resources is lost for extensive, interoperable and generally accessible computer use.

On the other hand, the language technology community, for their part, have created an overwhelming number of different types of lexical resources over the last thirty years, which are used for natural language processing tasks. These include corpora, lexicons, glossaries (used in machine translation), machine-readable dictionaries, lexical databases, and many others. One of the crucial issues addressed by ELEXIS is the fact that in the past the impressive results of the language technology community have rarely found their way into the practical work of creating lexicographic resources. This can be largely attributed to the lack of a common platform for building, sharing and exploiting knowledge and expertise between computational linguistics and lexicography, which is one of the goals of the proposed infrastructure.

In 2013, the European lexicographic community was brought together in the European Network of e-Lexicography (ENeL) COST action.<sup>1</sup> This initiative was set up to improve the access for the general public to scholarly dictionaries and make them more widely known to a larger audience. In the context of this network, a clear need has emerged for a broader and more systematic exchange of expertise, for the establishment of common standards and solutions for the development and integration of lexicographical resources, and for broadening the scope of application of these high quality resources to a larger community, including the Semantic Web, artificial intelligence, NLP and digital humanities. At the end of the COST action, the initiative had been successfully transformed into a H2020 infrastructure project – European Lexicographic Infrastructure (ELEXIS).<sup>2</sup>

The objectives emphasized in ELEXIS are the following: the infrastructure will (1) foster cooperation and knowledge exchange between different research communities in lexicography in order to bridge the gap between lesser-resourced languages and those with advanced e-lexicographic experience; (2) establish common standards and solutions for the development of lexicographic resources; (3) develop strategies, tools and standards for extracting, structuring and linking of lexicographic resources; (4) enable access to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders; (5) and promote an open access culture in lexicography, in line with the European Commission recommendation on access to and preservation of scientific information.

## 2 The Consortium

The consortium is composed of content-holding institutions and researchers with complementary backgrounds in terms of lexicography, digital humanities, standardization, language technology, the Semantic Web and artificial intelligence, and it cooperates strongly with the existing CLARIN and DARIAH<sup>3</sup> infrastructures. ELEXIS project partners are:

1. The Jožef Stefan Institute, Slovenia (leading partner)
2. Lexical Computing, Czech Republic
3. Dutch Language Institute, Netherlands
4. Sapienza University of Rome, Italy
5. National University of Ireland, Galway, Ireland
6. Austrian Academy of Sciences, Austrian Centre for Digital Humanities, Austria
7. Belgrade Center for Digital Humanities, Serbia
8. Hungarian Academy of Sciences, Research Institute for Linguistics, Hungary
9. Institute for Bulgarian Language, Prof. Lyubomir Andreychin, Bulgaria
10. Universidade Nova de Lisboa, Faculty of Social Sciences and Humanities, Portugal
11. K Dictionaries, Israel
12. Institute for Computational Linguistics A. Zampolli, Italy
13. The Society for Danish Language and Literature, Denmark
14. University of Copenhagen, Centre for Language Technology, Denmark
15. Trier University, Centre for Computational Linguistics and Digital Humanities, Germany
16. Institute of the Estonian Language, Estonia
17. Spanish Royal Academy, Spain

<sup>1</sup> [www.elexicography.eu](http://www.elexicography.eu)

<sup>2</sup> <http://www.elex.is/>

<sup>3</sup> Web sites: <https://www.clarin.eu/>, <https://www.dariah.eu/>.

Broadly speaking, work in the consortium focuses on three different types of activities: (1) joint research activities, (2) networking activities and (3) development of ELEXIS infrastructure through what is defined as “virtual access” and “trans-national access” activities. In the following sections we describe the three types of activities.

### 3 Research activities in ELEXIS

Research in ELEXIS is generally focused on two areas: lexicography and natural language processing. Lexicographic resources, both born-digital and retrodigitized, have different levels of structure and are not equally suitable for application in advanced NLP technologies. ELEXIS goal is thus to develop strategies, tools and standards for extracting, structuring and linking the high quality semantic data from lexicographic resources and make them available to the Linked (Open) Data family. In addition to linking lexicographic content, we also work on interlinking lexical content with other structured or unstructured data – corpora, multimodal resources, etc. – on any level of lexicographic description: semantic, syntactic, collocational, phraseological, etymological, translation equivalents, examples of usage, etc. The ultimate goal is the creation of a universal registry/network of semantic relations used as a semantic intermediary language for global knowledge exchange, focused on difficult polysemous vocabulary (single-word and multi-word), modern and historical; the realization of a universal lexicographic metastructure, i.e. a matrix dictionary spanning across languages and time.

In order to motivate interoperability, we enable partners and other stakeholders to encode their data with common concepts from models such as the BabelNet (Navigli & Ponzetto 2012) and other Semantic Web models, such as DBpedia. To ensure that there is integration at even the most basic level, ELEXIS partners will define a minimal common data model capturing the basic concepts of a lexicographic resource such as entries (single-word, multi-word), senses, syntactic frames, etymologies, etc. and linguistic relationships such as synonymy/antonymy, translation, domain/region/register classification, relatedness, and so on that will be compatible with existing models used in the community, including TEI (Text Encoding Initiative), LMF (Lexical Markup Framework) and OntoLex-Lemon (McCrae et al. 2017), a model for modelling lexicon and machine-readable dictionaries, and linked to the Semantic Web and the Linked Data cloud.

Ultimately, research results in ELEXIS will be integrated in a platform that allows lexicographers to see the results coming from information extraction and corpus information, as well as crowdsourcing. This will enable lexicographers to make more detailed and consistent analyses of words in context. However, it is important to note that from the point of view of human-oriented lexicography, faced with abundance of data, the emphasis is on knowing what not to say and on how to say it efficiently. Therefore, methods and tools for visualization and presentation of lexicographic data are also extremely important, and will receive due attention. These research activities create a so-called virtuous cycle of cross-disciplinary exchange of knowledge and data, as shown in Figure 1.

The virtuous cycle of eLexicography includes experimental validation of the integrated LLOD data in Natural Language Processing tasks (lexicography for NLP) and the use of NLP for lexicography. As regards the former, the following tasks will be shown to benefit from the huge amount of multilingual information integrated in this project:

- **Multilingual Word Sense Disambiguation**, addressing the paucity of sense-annotated sentences. The ELEXIS lexicographic resources will be utilized to bootstrap large training datasets for WSD in dozens of languages.

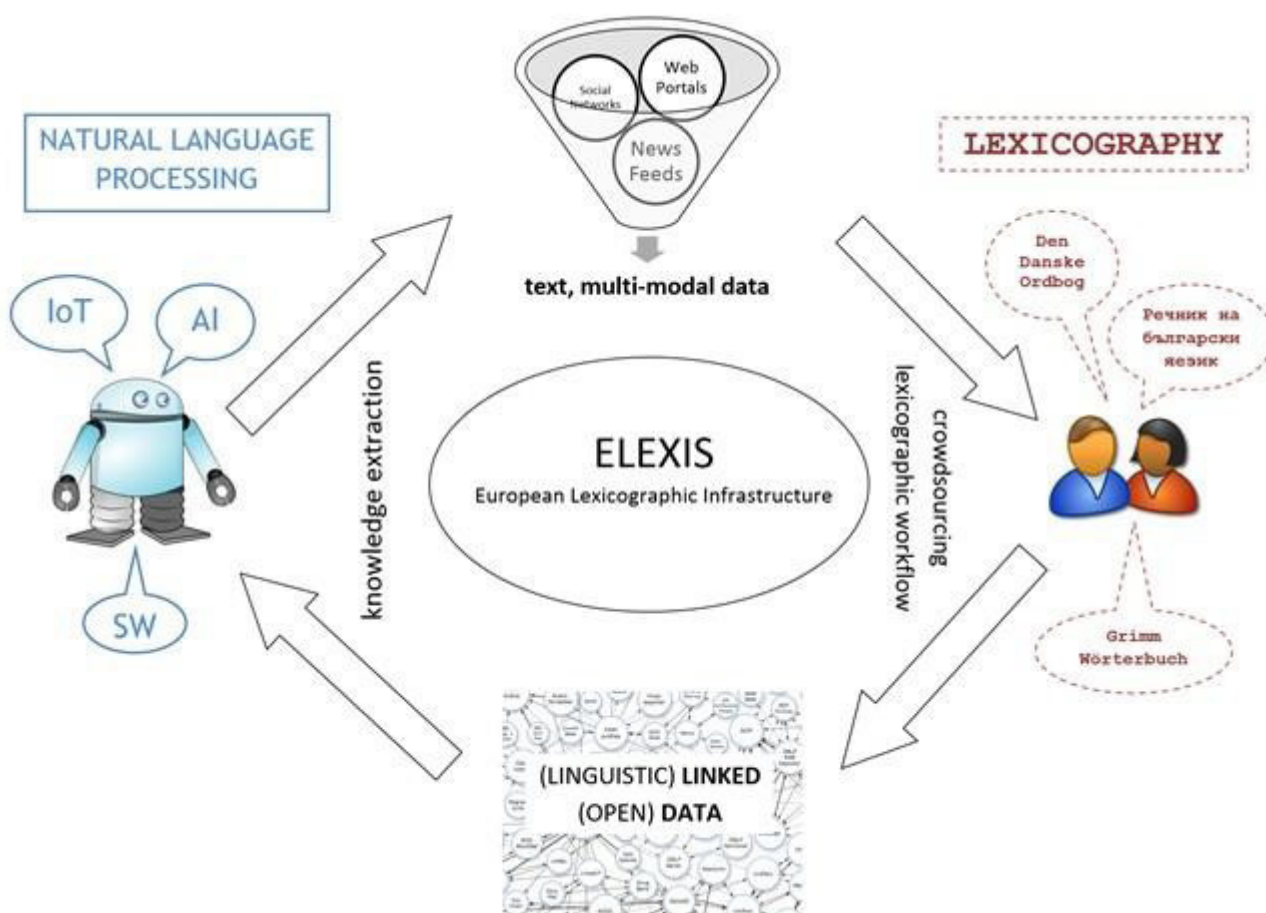


Figure 1: The virtuous cycle of eLexicography

- **Multilingual Semantic Parsing:** semantic parsing aims to map sentences to formal representations of their meaning. In ELEXIS we will develop innovative algorithms that exploit the huge multilingual network of interlinked lexical knowledge to perform multilingual semantic parsing.
- **Word sense clustering,** where the development of semi-automatic procedures to bring together subtle sense distinctions in clusters of meanings will be shown to improve the performance of tasks such as Word Sense Disambiguation;
- **Domain labelling of text,** where the aggregated information obtained from the lexicographic network of resources will be shown to improve automatic tagging of text with domain labels in arbitrary languages, thanks to developing innovative neural techniques.
- **Study of the diachronic distribution of senses:** the use of the most frequent sense in NLP is a solid baseline used in WSD and other tasks, but it is available only in the English language. We will develop novel techniques for aggregating the predominance information of senses a) from the multitude of resources and b) considering evolution over time, which will have an important impact on disambiguation and corpus analysis.

Advances in AI and NLP will, in turn, enable the development of improved tools for the production of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques resulting in a new type of lexicographic resource: a dictionary-on-the-fly. These new methods and tools will significantly shift the lexicographer's starting point and reduce the time-consuming parts of lexicographic work. In principle, having enough web or corpus data in a particular language will be a sufficient condition for a dictionary of that language



to be created, bridging the gap between lesser-resourced languages and those with advanced e-lexicographic experience. In addition to developing methods and tools for the automatic acquisition of lexicographic data, methods and tools for introducing crowdsourcing and gamification in the lexicographic process, and methods and tools to enrich lexicographic resources with multi-modal data will also be developed.

## 4 Networking Activities in ELEXIS

In order to reach the goals described in the introduction in Section 1, not only are research activities carried out, but networking activities are also needed, especially because ELEXIS will build and foster a community around the infrastructure and will support the exchange of knowledge. Therefore, special attention is given to networking activities that are reflected at different levels, as explained below.

### 4.1 Organization

The objective of ELEXIS is to foster cooperation and knowledge exchange among different research communities in lexicography in order to bridge the gap between less-resourced languages and those with advanced e-lexicographic experience, and one of the impacts of ELEXIS is defined as the emergence of a new type of lexicography that no longer views languages as isolated entities, but fully embraces the pan-European nature of those spoken in Europe. This ambition also extends to the global level. ELEXIS plans reflect this with an inclusive multi-layered organization that aims at engaging different user groups with various levels of intensity during the project, as shown in Figure 2.

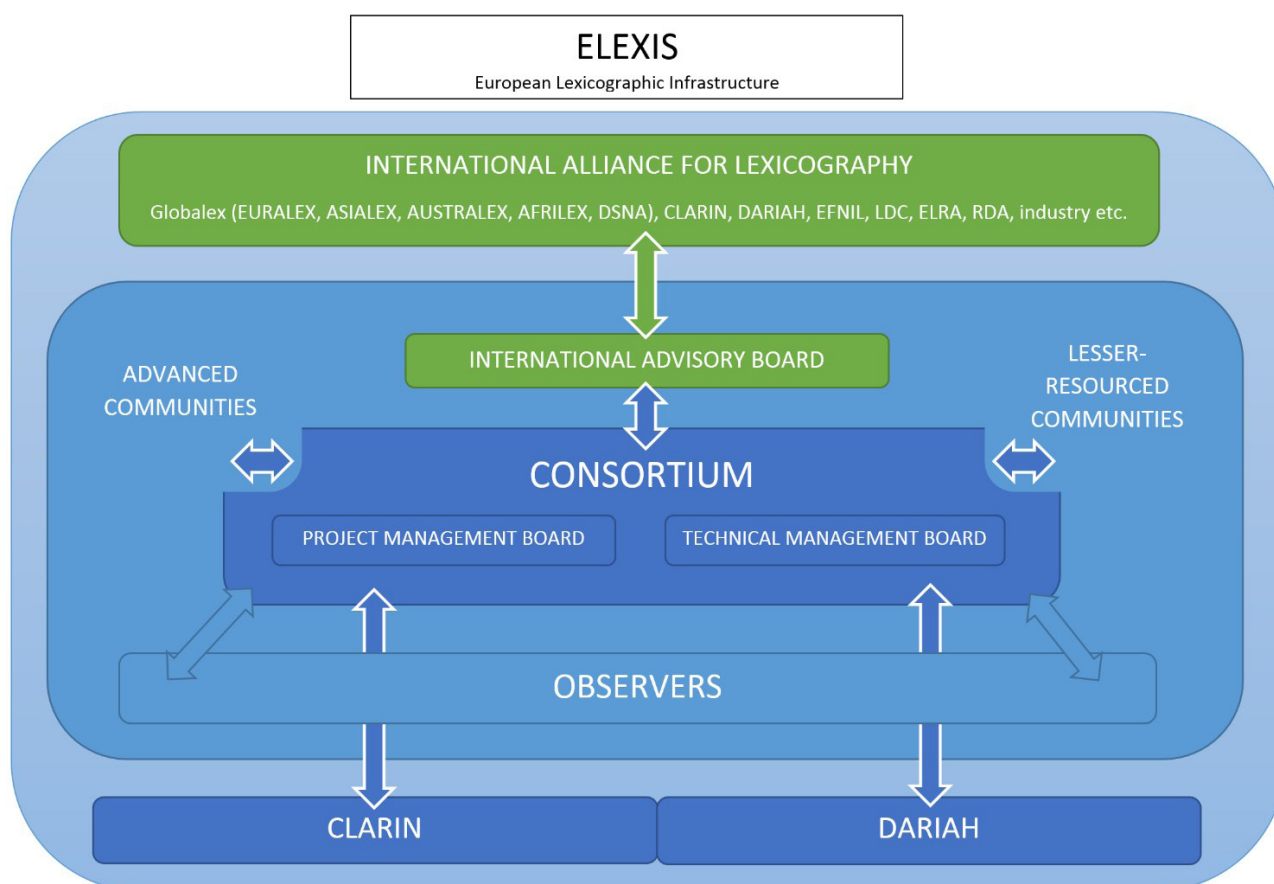


Figure 2: ELEXIS organization

#### ***4.1.1 The Consortium and Observers***

Elements of the structure consist of ELEXIS consortium partners as the core group responsible for the development of the infrastructure. Another organizational layer is observing institutions that will be directly included in outreach and dissemination activities through various channels. The central group of institutions that fall under the observer category are those producing quality lexicographic data and resources, filtered by the criteria of inclusion in the European Dictionary

Portal developed by the European Network of e-Lexicography COST action.<sup>4</sup> Typically but not exclusively, these institutions include (European) national language institutes, large dictionary publishers and other prominent producers of lexicographic data.

#### ***4.1.2 Advanced and Less-resourced Communities***

The broadest and also least defined user group consists of less-resourced communities and advanced communities working in lexicography. In practical terms, these include all researchers interested in lexicography who will gain access to newly developed data, tools and services through ELEXIS virtual access activities. The consortium itself reflects this division, as it includes partners working with less-resourced languages (Serbian, Slovenian, Irish, etc.) and from more advanced communities (English, German, Dutch, Italian, etc.). The infrastructure will provide networking activities (training, online training material, conferences, workshops, etc.) to enable researchers working with less-resourced languages to benefit from the results of the project, building on the expertise and data available in the more advanced communities.

#### ***4.1.3 International Alliance for Lexicography***

ELEXIS consortium partners will be advised by an International Advisory Board (IAB). Members of IAB include appointed representatives of five continental societies and associations for lexicography and top experts in both relevant fields, lexicography and NLP, from academia and industry. Through the International Advisory Board, and using its outreach and community building activities, ELEXIS will strive to form an International Alliance for Lexicography, which will include stakeholders from various fields. The alliance will be formed as a loose organization dedicated to the advancement of lexicography in the digital age, with the ELEXIS International Advisory Board serving as the initial governing body. It is expected that the alliance will be joined by the five continental lexicographic associations who form the Globalex initiative,<sup>5</sup> standardization bodies organized in EFNIL (European Federation of National Institutions for Language), data providers such as LDC (Linguistic Data Consortium), ELRA (European Language Resources Association), RDA (Research Data Alliance), and representatives of both language technology (Language Technology Industry Association – LT Innovate) as well as lexicography and language learning industries (e.g. Oxford University Press, MacMillan Publishers). The aim of the alliance is to consolidate the field of lexicography and enable its transition to the digital environment, bringing together all stakeholders interested in language description and semantic data.

#### ***4.1.4 CLARIN and DARIAH***

ELEXIS will also serve as a hub between CLARIN and DARIAH: dictionaries are essential language resources whose quality, reliability and coverage can be vastly improved by means of harmonizing formats and optimizing points of infrastructural access. At the same time, however, dictionaries are

<sup>4</sup> Web page: <http://www.dictionaryportal.eu/en/catalog/>.

<sup>5</sup> Web page: <http://globalex.link/>.

also objects of humanities research in their own right. Humanities scholars study dictionaries in terms of their cultural and ideological values, or their role in language standardization and nation-building, to name just a few different perspectives.

ELEXIS as a new infrastructure builds upon the existing tools and services of CLARIN and/or DARIAH with the goal of achieving something that neither infrastructure can at the moment provide on its own: a concerted pan-European effort aimed at combining and advancing the state-of-the-art in three distinct fields — lexicography, NLP and digital humanities. CLARIN already has a leading role in providing language resource repositories, linguistic annotation pipelines and federated search facilities, whereas DARIAH is a leader in facilitating long-term access to and use of arts and humanities research data. By creating a common platform for building, sharing and exploiting high-quality, multilingual lexical data, ELEXIS will aim to serve as a catalyst for closer cooperation between the two existing infrastructures. ELEXIS can succeed in this role because it has assembled a critical mass of eminent stakeholders from various disciplines who have both the technical and scholarly potential to: 1) help lexicographers build better dictionaries using the most advanced NLP techniques; 2) provide NLP researchers with high-quality lexicographic data to test and improve their algorithms on; and 3) aid humanities scholars in accessing social, historical and cultural data contained in legacy dictionaries in order to develop new procedures and tools for analyzing, visualizing and interpreting large sets of lexical data.

## 4.2 Dissemination and Community Building

Due to the nature of the ELEXIS infrastructure, various communities and types of users (professional, semi-professional and general public) will benefit on account of the scalable outcomes and services. We identified several major target groups, those who are undertaking lexicographic projects and those who are applying the high quality lexicographic data e.g. in the context of the Semantic Web, artificial intelligence, natural language processing and the digital humanities (Declerck et al. 2018). Next, we describe in more detail the different usage scenarios related to lexicographic projects.

### 4.2.1 Professional Large-scale Lexicography

This group includes commercial and non-commercial entities undertaking large-scale lexicographic projects carried out by professional specialized teams from these areas:

- national language institutes (including consortium members)
- academia, universities, research institutions outside the ELEXIS consortium
- language standardization bodies and their umbrella organization EFNIL (European Federation of National Institutions for Language).
- industry (publishing houses, also software developers and language industry – in connection with large lexicographic projects).

### 4.2.2 Professional Small-scale Lexicography

This group includes entities undertaking small-scale lexicographic projects carried out on a highly professional level either for research purposes or to address the needs of a small, well-defined community. The following groups may fall into this category:

- individual researchers (from the field of lexicography, also language studies, translation studies or the sister field of digital humanities, as well as natural language processing in connection with lexicography)
- trainers and students (interested in the educational aspects of the ELEXIS projects, such as learning material, training events)

- professionals and practitioners (language professionals, translators, proofreaders and others who use or produce linguistic resources in their daily professional life)
- freelance terminologists.

#### ***4.2.3 Spontaneous and Small-scale Lexicography***

This group includes an enormous number of small projects often carried out without expertise in lexicography to address very specific needs of highly-specialized or very small professional or general public communities. A typical example would be a highly specialized domain-specific glossary. The following groups may fall into these categories:

- professional organizations, associations and authorities, non-profit organizations
- general public

Each group will be targeted with tailor-made messages that address the needs of the community to maximize the impact of any such activity.

### **4.3 Trans-national Access**

During the lifetime of the project ELEXIS will organize trans-national calls enabling researchers (a) to work with data with restricted access at host institutions; and (b) to gain knowledge and expertise in close contact with lexicographers and experts in NLP and artificial intelligence. One of the reasons for the limited accessibility of lexicographic data outside institutions which are the creators and copyright holders of such data is the effort needed for their compilation, which necessitates tighter control over the access and availability of raw data. Trans-national activities represent one of the mechanisms of ELEXIS to enable access to such content for researchers from other institutions or countries. However, the results of research conducted in trans-national activities will be available under open access licenses according to the rules of the call enabling the international community to familiarize itself with previously inaccessible resources. In total, ELEXIS will give access to eleven European infrastructures/lexicographical milieus where researchers/lexicographers within the EU member states or associated countries are invited to apply for free-of-charge access via grant visits. During the visits, the hosting institutions will provide support in terms of both lexicographical and technical expertise.

## **5 ELEXIS Infrastructure**

ELEXIS “virtual access” infrastructure – providing online access to data, tools and services – will consist of three sub-infrastructures: LEX1, LEX2, LEX3.

### **5.1 LEX 1**

The first part of the infrastructure is dedicated to automatic segmentation and structuring of content for dictionaries that are currently produced in digital environments, but are typically encoded in their own custom data format. Conversion and alignment tools provide users of the infrastructure with the possibility to harmonize and convert their lexicographic resources into a uniform data format that allows their integration in Linked Open Data. Standards will be developed and tested during the project on the data provided by the lexicographic partners and implemented in the newly-developed service.

To provide conceptual interoperability, services enabling the linking of lexicographic resources will be developed and made available in the linking tools segment of the platform. This will provide the possibility to link lexical entries, senses and fundamental concepts in different lexical resources,

using a semi-automatic approach. BabelNet, as an existing multilingual resource to provide cross-lingual linking, is exploited for this purpose. Extensive linking of existing lexicographic resources by pivoting through BabelNet will enable the creation of what we call ELEXIS matrix dictionary – a universal repository of linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical, etc.

## 5.2 LEX 2

Based on the contribution of lexicographic data, a new infrastructure is being developed that includes word sense disambiguation and entity linking tools dedicated to semantic processing of corpus data. These tools will have an important impact on disambiguation and corpus analysis, and will open up the possibility to create lexicographic data from corpora in a fully automated process. This is included in the dictionary-on-the-fly segment of the platform. The service will be able to produce a proto-dictionary with sense distribution, extracted definitions, collocations, multi-word expressions, (good dictionary) examples, translation equivalents and data in other modalities.

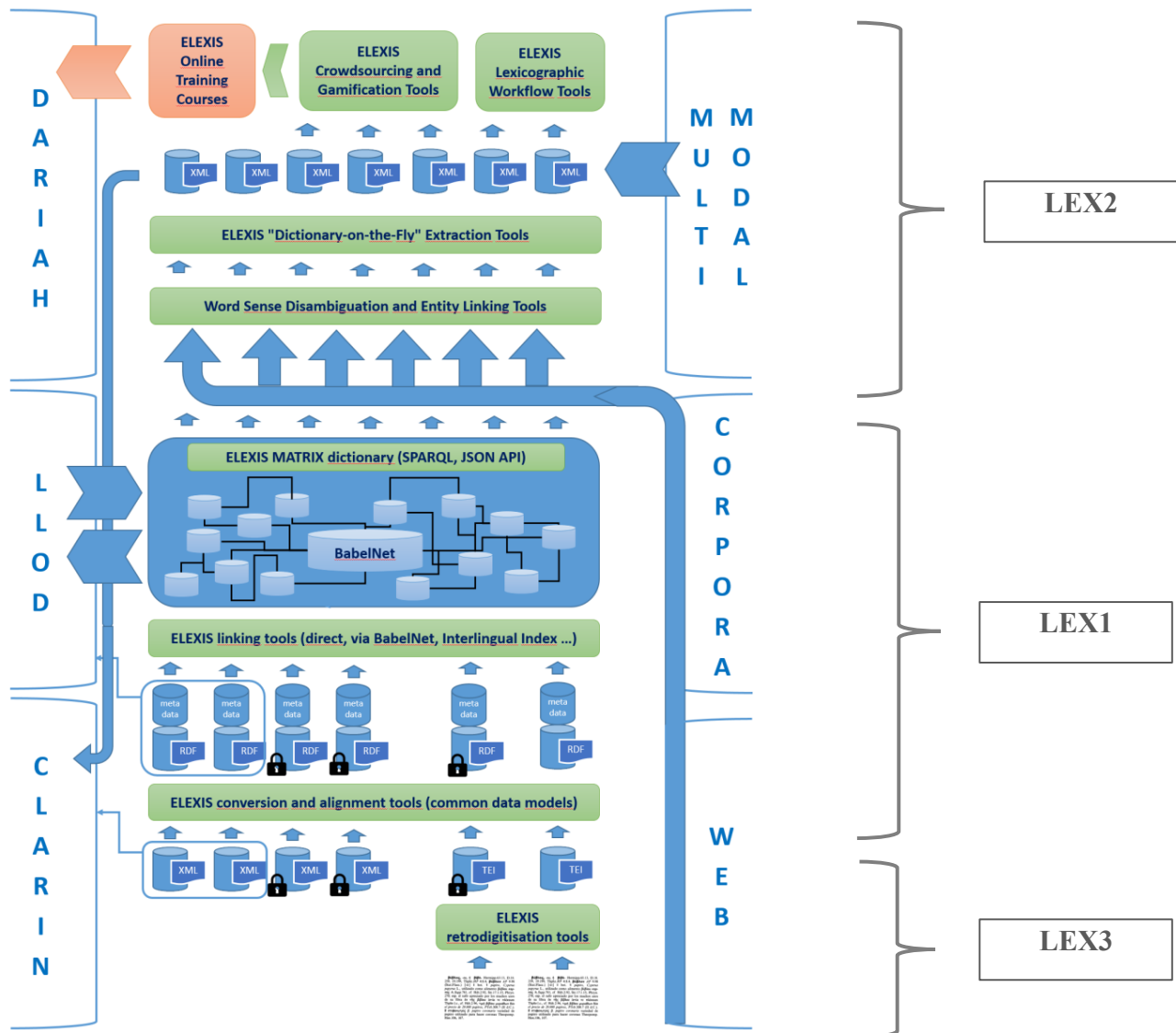


Figure 3: Tools and services: green, (lexicographic) data: blue, online training and education: brown.



To enable online lexicographic work on both existing and new (extracted) lexicographic data, two complementary sets of tools are provided: lexicographic workflow tools and crowdsourcing and gamification tools. The first include an open source online dictionary writing system, with the aim to provide the central dictionary writing platform which also includes new possibilities of online collaboration. The second provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification).

### 5.3 LEX 3

The third set of services is dedicated to retrodigitized dictionaries in the part of the platform that includes (1) tools for automatic segmentation and structuring of content in retro-digitized dictionaries, and (2) an online publication tool for retrodigitized dictionaries which also offers interfaces for the analysis and profiling of the underlying lexical data.

## 6 Conclusion

As a new infrastructure, ELEXIS brings together research communities and consortium partners working in different fields, in order to support the community working in the emerging field of e-lexicography. In particular, ELEXIS builds on the existing expertise and knowledge of partners in the fields of lexicography, computational linguistics and artificial intelligence in an interdisciplinary effort to make existing lexicographic resources available on a significantly higher level compared to their availability as stand-alone resources, which is the current state of affairs.

To support the lexicographic process and contribute to lexicography-oriented language description, ELEXIS will:

- develop methods and tools for the automatic processing and extraction of data from corpora and other (multimodal) resources for lexicographic purposes;
- develop methods and tools for the inclusion of extracted data into interlinked (open) lexicographic data;
- develop methods, guidelines and tools enabling the use of crowdsourcing and citizen science in the lexicographic process;
- elaborate on the guidelines and solutions for handling copyright and authorship protection to enable inclusion of extracted data into the lexicographic workflow.

To support the natural language processing community, several steps are needed to make existing lexicographic resources globally available. Therefore, ELEXIS will:

- develop methods, guidelines and tools for harmonization of dictionary formats, building on the existing standards within the lexicographic and NLP community;
- develop methods and tools for automatic segmentation and identification of dictionary structure, enabling interlinking of dictionary content;
- develop methods and tools for interlinking, maintenance, reuse, sharing and distribution of existing lexicographic resources;
- define evaluation and validation protocols and procedures (lexicographic data seal of compliance);
- elaborate on the guidelines and solutions for handling copyright and authorship protection to enable open access to lexicographic data in the LOD framework.

## References

- Declerck, T. et al (2018)....
- McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pp. 587-597.
- Pilehvar, M. T., Navigli, R. (2014). A Robust Approach to Aligning Heterogeneous Lexical Resources. *Proceedings of ACL 2014*, pp. 468-478.
- Navigli R., Ponzetto, S. (2012): BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, pp. 217-250.

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 731015.



# The EcoLexicon English Corpus as an Open Corpus in Sketch Engine

**Pilar León-Araúz<sup>1</sup>, Antonio San Martín<sup>2</sup>, Arianne Reimerink<sup>1</sup>**

<sup>1</sup>Department of Translation and Interpreting, University of Granada, <sup>2</sup>Department of Modern Languages and Translation, University of Quebec in Trois-Rivières

E-mail: pleon@ugr.es, antonio.san.martin.pizarro@uqtr.ca, arianne@ugr.es

## Abstract

The EcoLexicon English Corpus (EEC) is a 23.1-million-word corpus of contemporary environmental texts. It was compiled by the LexiCon research group for the development of EcoLexicon (Faber, León-Araúz & Reimerink 2016; San Martín et al. 2017), a terminological knowledge base on the environment. It is available as an open corpus in the well-known corpus query system Sketch Engine (Kilgarriff et al. 2014), which means that any user, even without a subscription, can freely access and query the corpus. In this paper, the EEC is introduced by describing how it was built and compiled and how it can be queried and exploited, based both on the functionalities provided by Sketch Engine and on the parameters in which the texts in the EEC are classified.

**Keywords:** specialized open corpus, terminology, corpus exploitation

## 1 Introduction

Corpora have become a key element of almost all language studies, as any assertion about language requires verification through real linguistic data to be deemed credible (Teubert 2005: 1). Having access to general and specialized corpora is thus essential for anyone involved in research or any professional activity related to language. However, many of these professionals do not have time to compile large corpora. The EcoLexicon English Corpus (EEC) is a 23.1-million-word specialized corpus of contemporary environmental texts. It was compiled by the LexiCon research group for the development of EcoLexicon (Faber et al. 2016; San Martín et al. 2017), a terminological knowledge base on the environment.<sup>1</sup> In EcoLexicon, the EEC and its Spanish counterpart (together over 50 million words) can be queried with pragmatic restrictions such as author, date of publication, target reader, contextual domain, and keywords. However, its search engine does not provide all the functionalities of the well-known corpus tool Sketch Engine (Kilgarriff et al. 2014). This is why the EEC was made available as an open corpus in Sketch Engine, which means that any user, even without a subscription, can freely access and query the corpus.<sup>2</sup> One very interesting module provided by the query system is information extraction through word sketches, which are automatic corpus-derived summaries of a word's grammatical and collocational behavior (Kilgarriff et al. 2010). Apart from the built-in word sketches, Sketch Engine allows users to customize sketches for their specific needs. In the case of the EEC, this has enhanced the extraction of semantic information.

In this paper, the EEC is introduced by describing how the corpus was built and compiled (Section 2), and how it can be queried and exploited (Section 3), based on the functionalities provided by Sketch Engine, the parameters in which the texts in the EEC are classified and the word sketches exclusively created for the EEC. Finally, Section 4 offers some concluding remarks.

<sup>1</sup> EcoLexicon is freely available at <<http://ecolexicon.ugr.es>>.

<sup>2</sup> Certain advanced functionalities are only available for subscribed users.

## 2 Creating the EcoLexicon English Corpus

The EEC is a 23.1-million-word corpus of contemporary environmental texts. It was first created as an internal tool for knowledge extraction while building EcoLexicon. However, it was made publicly available because it evolved to be a tool in itself that terminologists, translators or even experts could exploit for different purposes (i.e. modeling, comprehension and production tasks) within the specialized domain of the environment. As Sinclair (1991: 24) pointed out, we should not expect a general reference corpus like the British National Corpus to adequately document specialized genres and domains. It follows that we need more specialized corpora, compiled with enough texts and text types to represent a knowledge domain, as they are more likely to document the conventions of the genre and the concepts and terms of the domain.

Each text in the EEC is tagged according to a set of XML-based metadata, some of which are based on the Dublin Core Schema, while others have been included to meet the needs of the research group. Corpus metadata permit users to constrain corpus queries based on pragmatic factors, such as environmental domains and target reader. Thus, for instance, the use of the same term in different contexts can be compared. Tags are based on the following main parameters:

- Domain: the EEC encompasses all the domains and subdomains of environmental studies (e.g., Biology, Meteorology, Ecology, Environmental Engineering, Environmental Law, etc.).
- User: the corpus includes texts for three types of user, depending on level of expertise (i.e., expert, semi-expert, general public).
- Geographical variant: it comprises American, British, and Euro English.
- Genre: it covers a wide variety of text genres (e.g., journal articles, books, websites, lexicographical material, etc.).
- Editor: it distinguishes texts edited by scholars/researchers, businesses, government bodies, etc.
- Year: it includes texts from 1973 to 2016.
- Country: the texts are tagged according to the country of publication.

The EEC was processed and compiled in an internal application of the research group. Then it was recompiled within Sketch Engine with the Penn Treebank tagset (TreeTagger version 3.3) and with the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz & San Martín 2018; León-Araúz, San Martín & Faber 2016), a CQL-based (Corpus Query Language) (Jakubíček et al. 2010) customized sketch grammar separate from the default sketch grammar. The ESSG was developed for the extraction of semantic word sketches based on some of the most common semantic relations in terminology: generic-specific, part-whole, location, cause, and function.

When a corpus is compiled with a collection of different pattern-based grammar rules such as the above, new word sketches can be queried within the Sketch Engine (see Section 3.2). The ESSG thus has three aims: (1) extracting semantic relations for building EcoLexicon; (2) offering semantic word sketches in the EEC; and (3) providing other users with the possibility of reusing them in their own corpora.<sup>3</sup>

## 3 Exploiting the EcoLexicon English Corpus

The combination of pragmatic, syntactic and semantic information that can be extracted from the corpus makes the EEC an adequate resource for all kinds of end users with an interest in environmental science, such as domain experts, professional writers, translators, terminologists, ESP researchers,

<sup>3</sup> The latest version of the ESSG can be downloaded from <<http://ecolexicon.ugr.es/essg/>>.



etc., as stated above. Thanks to Sketch Engine's automation capabilities, users are able to analyze and extract a sizable quantity of linguistic data that would have been unmanageable in the past (Kosem et al. 2014: 362). In the following sections, different queries will be provided by combining the main functionalities of Sketch Engine with the parameters according to which the EEC is tagged.<sup>4</sup>

### 3.1 Search and Text Types

The feature *Search* is the main way to access concordances in Sketch Engine. Different types of queries are possible (simple, lemma, phrase, word, character and CQL), and they can be combined with the contextual filter, which allows the user to limit the lemmas that should appear around the word or words of the query. Additionally, in the case of the EEC, any query performed through the *Search* feature can be filtered according to text type based on the tagging of the EEC (domain, genre, editor, etc.) (Figure 1).

The filtering by text type can be chosen manually for each query. However, the user can also create subcorpora based on text types. For instance, a user may want to create a simple subcorpus for the domains of Hydrology or Renewable Energy, or complex subcorpora, such as one containing only articles and books in British English from the domain of Biology for experts in the field. Additionally, the EEC comes with several subcorpora created by default (i.e. American English, British English, Year 1973–1999, Year 2000–2009 and Year 2010–2016).

Figure 1: Sketch Engine's Search and EEC Text types.

All these possibilities of query customization allow the user to retrieve, for instance, all the concordances where *recycle* is a verb in texts addressed to the general public (lemma search filtered by user) or where *climate change* occurs in Environmental Law texts (phrase search filtered by domain). Additionally, the *Context* option can be combined with any search, permitting the user to find, for example, all the concordances in Oceanography academic articles where the lemma *wind* appears in a window of  $\pm 15$  tokens of the lemma *wave*.

However, given that the EEC was recompiled with TreeTagger, it is possible to perform more fine-grained queries in CQL, allowing for the formalization of grammar patterns in the form of regular expressions combined with POS-tags. CQL queries used together with text-type filtering are a powerful tool to research the workings of environmental English. An example of a CQL query is `([tag="N.*"] [lemma="amount" & tag="N.*"]) | ([lemma="amount" & tag="N.*"] [word="of"] [tag="N.*"])`, which finds concordances of the lemma *amount* either preceded by any noun or followed by *of* and any noun. Figure 2 shows a sample of the resulting concordances limited to the Meteorology subdomain.

<sup>4</sup> Due to space restrictions, no instructions are provided. However, interested readers can consult the user-friendly Sketch Engine manual at: < <http://sketchengine.co.uk/user-guide/> >

fall centers in June. However, the simulated nuclear power plants do not lead to significant rainfall amount and distribution in June in the uncoupled model and some extreme dust events bring the same amounts of contamination, potential accidents in the nuclear industry (Chetti et al., 2008). Such fires produce large amount of dust in a few hours as the cumulated amount over the plots and the loss of cross-correlation between amounts of gases and particles from biomass burning and soil (Rajagopalan et al., 2008). The plots of the right panel show that high precipitation amounts and other weather variables are derived from the following specific weather conditions: from the north and north-east is slightly undervapor than cool air, so when air cools it is not it can hold. That is why the flat regions below is facing water shortages due to drought and the Earth loses into space. The greenhouse effect in the atmosphere has gone up and down during the month of October as the month of November flowing through. Carbon dioxide (CO<sub>2</sub>) species

Figure 2: Sample of the results for the CQL query *amount* preceded by a noun or followed by *of* and any noun in the Meteorology subdomain.

With CQL queries, a user can also compare the frequency of different variants of multiword expressions. For example, in the term *geologic time scale*, *geologic* can be replaced by *geological* and *time scale* can be written as a single word. With the CQL query `[lemma="geologic.*"] ([lemma="timescale"])([lemma="time"] [lemma="scale"])` we can retrieve all the concordances where all the variants appear, and with the *Frequency – Node forms* feature we can see which form is more frequent (Figure 3).

word	Frequency	Items: 9    Total frequency: 139
P   N geologic time scale	41	
P   N geological time scale	37	
P   N geological timescale	18	
P   N geological timescales	17	
P   N geological time scales	10	
P   N geologic timescale	5	
P   N geologic time scales	5	
P   N geologic timescales	3	
P   N Geologic time scale	3	

Figure 3: Frequency of variants of *geologic time scale* in the EEC.

Another feature of Sketch Engine that permits users to fully exploit the EEC is *Frequency – Text type*. With this feature, users can observe how language expression changes across different levels of expertise in the environmental domain. For instance, when searching for the verb *liquefy*, concordances can be filtered according to the user type parameter. Not surprisingly, the verb appears more often in expert-related texts than in texts addressed to the general public (Figure 4).

User	Frequency	Rel [%]	Items: 3    Total frequency: 217
P   N Expert	153	125.10	
P   N Semi-expert	52	83.60	
P   N General public	12	39.10	

Figure 4: Frequency of *liquefy* in the EEC according to user type.

With this feature the frequency of terms in different domains can also be observed, thus verifying if a term is more specific to one domain or another. For instance, by searching the lemma *photovoltaic* and looking up its frequency according to domain, the results show that it is a term mainly linked to the domain of Renewable Energy, although it also occurs, but with much lower frequency, in Climatology and Air Quality Management (Figure 5).

Domain	Frequency	Rel. [%]	Items: 4    Total frequency: 174
P   N 3.5.1 Renewable Energy	106	2,683.70	
P   N 2.7.2 Climatology	34	104.10	
P   N 3.2.5.3 Air Quality Management	17	249.40	
P   N 0 General	17	85.00	

Figure 5: Frequency of *photovoltaic* in the EEC according to environmental subdomain.

### 3.2 Word Sketch and Sketch Diff

The EEC employs both the default sketch grammar for English underlying the word sketches in the tool in combination with the ESSG. Users can benefit from Sketch Engine's default word sketches when searching for the collocations that are used more often in specialized discourse in combination with a certain term. For instance, Figure 6 shows the modifiers of *methane*, the nouns modified by *methane* and the verbs that collocate with *methane* both as object and subject.

modifiers of "methane"	nouns modified by "methane"	verbs with "methane" as object	verbs with "methane" as subject
17.94	34.14	15.42	13.06
dioxide 35 9.80	emission 75 8.24	produce 41 6.58	be 151 3.45
atmospheric 27 6.47	gas 54 7.70	be 36 3.08	have 24 3.52
coalbed 17 10.49	production 43 7.40	release 29 8.35	increase 7 5.57
co2 15 8.71	hydrate 38 10.60	include 20 4.85	react 5 8.23
gas 9 6.33	oxide 30 8.64	emit 14 7.93	cause 4 4.00

Figure 6: Word sketches of *methane* extracted from the EEC.

Thanks to the ESSG, users can access ready-made semantic word sketches such as those shown in Figure 7, where search terms may appear related to their hyponyms (i.e. *microorganism*), the whole they are part of (i.e. *oxygen*), their underlying causes (i.e. *tsunami*), etc.

"microorganism" is the generic of...	"oxygen" is part of...	"tsunami" is caused by...
13.42	6.36	16.05
bacterium 29 10.78	atmosphere 23 9.80	earthquake 93 11.41
fungus 15 10.49	molecule 21 10.18	landslide 43 10.71
pathogen 5 9.35	compound 18 9.51	eruption 26 9.79
alga 5 8.42	water 16 8.44	water 22 8.14
virus 4 8.76	earth 15 9.26	slide 13 9.14

Figure 7: Semantic word sketches of *methane* extracted from the EEC.

The word sketch queries can be complemented with the text type filters provided by the tags of the EEC (or subcorpora based on them). In this sense, users can also observe how concepts can change their relational behavior across different environmental subdomains. For example, Figure 8 shows how *nitrogen* is mainly categorized as a type of *pollutant* in the domain of Air Quality Management and as a type of *nutrient* in that of Biology.



"nitrogen" is a type of...		8.06
pollutant	9	8.61
gas	5	6.27

"nitrogen" is a type of...		3.27
nutrient	4	8.55
gas	3	5.56

Figure 8: *Nitrogen* generic-specific semantic word sketches in Air Quality Management (left) and Biology (right) subcorpora.

Additionally, if users access the concordances extracted with the ESSG, they can extract knowledge-rich contexts (i.e. contexts containing domain knowledge potentially useful for conceptual analysis (Meyer 2001)) like the ones in Table 1.

Table 1: Sample of knowledge-rich contexts extracted from the EEC with the aid of the ESSG.

<i>generic-specific</i>	<i>A <u>hydrograph</u> is a <u>graph</u> that reflects the discharge of a river over a period of time.</i>
<i>part-whole</i>	<i>The <u>astronomical tide</u> refers to the regular oscillations of the sea or ocean surface[...].</i>
	<i><u>Sand grains</u> usually consist of <u>quartz</u> but may also be fragments of feldspar, mica, and, [...].</i>
	<i><u>Seawater</u> contains <u>sodium chloride</u> and other salts in concentrations three times greater [...].</i>
<i>location</i>	<i><u>Lagoons</u> commonly form on <u>coastlines</u> that are subsiding, or where sea level is rising.</i>
	<i>Most <u>ozone</u> is found in the <u>stratosphere</u> at elevations between 10 and 50 kilometers [...].</i>
<i>cause</i>	<i>[...] the human costs of malaria outweigh the <u>environmental damage</u> caused by the <u>use of DDT</u>.</i>
	<i><u>Logging</u> may also contribute to <u>deforestation</u> by making it easier for agriculture to [...].</i>
<i>function</i>	<i><u>Membrane-assisted BAC</u> is used for the <u>removal of priority pollutants</u> from secondary [...].</i>
	<i><u>Liquid-in-glass thermometers</u> are often used for <u>measuring surface air temperature</u> because [...].</i>

Another word-sketch based feature that can be especially exploited with the EEC is *Sketch diff*. It allows the user to compare either the word sketches of two lemmas, the word sketches on the same lemma in two subcorpora, or two different word forms of the same lemma. Figure 9 shows an example of each type. At the left, the modifiers of *risk* (in green) and *hazard* (in red) in the whole EEC are contrasted. As it can be observed, these two semantically related terms tend to co-occur with different modifiers, although they also share some of them (in white). At the center, there is a sketch diff that shows how *water* takes different verbs as an object in Hydrology (in green) and Water Treatment and

modifiers of "risk/hazard"	2,393	2,099	0.36	0.70
extinction	30	0	8.4	--
cancer	27	0	8.4	--
disaster	30	0	8.3	--
disease	21	0	7.7	--
great	87	15	8.1	5.6
flood	180	39	9.8	7.7
erosion	106	24	9.0	6.9
flooding	23	5	8.1	6.0
drought	25	7	7.9	6.2
potential	99	49	8.6	7.6
health	142	81	10.1	9.4
serious	20	35	7.5	8.4
tsunami	18	53	7.1	8.8
safety	6	31	5.8	8.3
climate-related	4	37	5.7	9.1
geologic	4	42	5.1	8.6
earthquake	1	16	3.5	7.6
natural	7	290	4.1	9.5
aviation	0	13	--	7.5
geological	0	37	--	8.3

verbs with "water" as object	1,003	1,304	0.15	0.14
infiltrate	8	0	7.9	--
entrain	9	0	7.9	--
ground	8	0	7.8	--
flow	15	1	8.4	4.2
pour	7	2	7.7	5.5
evaporate	13	8	8.4	7.4
divert	11	10	8.2	7.7
contaminate	8	8	7.6	7.3
withdraw	7	8	7.6	7.5
pump	15	17	8.1	8.1
disinfect	6	9	7.5	7.8
treat	25	55	8.4	9.4
supply	9	24	7.1	8.3
drink	35	134	9.5	11.1
receive	10	37	6.5	8.3
purify	1	8	4.9	7.5
produce	13	117	4.9	8.0
conserve	1	14	4.3	7.9
intend	1	44	4.4	9.6
regulate	0	18	--	8.1

verbs with "gas" as subject	983	877	0.10	0.20
clean	16	0	8.9	--
stage	8	0	8.0	--
fire	7	0	7.8	--
exit	6	0	7.5	--
enter	11	0	7.3	--
reheat	4	0	7.1	--
flare	4	0	7.0	--
mix	15	1	7.8	4.0
play	14	2	7.5	4.8
exert	12	4	8.1	6.6
leave	4	6	6.1	6.8
build	3	5	5.9	6.7
pass	4	7	5.8	6.7
contribute	7	13	6.4	7.4
escape	2	5	5.9	7.3
absorb	11	30	7.6	9.1
expand	2	8	5.4	7.5
radiate	0	4	--	6.9
diffuse	0	5	--	7.4
trap	0	10	--	8.2

Figure 9: Sample of sketch diffs extracted from EEC.

Supply (in red), as well as a considerable number of shared results. Finally, the sketch diff at the right outlines the verbs that tend to have *gas* as subject in singular (in green) and in plural (in red) in the whole EEC.

### 3.3 Word List

The *Word list* feature can be used to extract frequency lists with many different settings including n-gram extraction, filtering based on regular expressions or keyword extraction with the aid of a user-chosen reference corpus. This feature can be used in combination with an EEC subcorpus, which allows the user to generate very specific frequency lists. Some examples of frequency lists that could be useful to generate from the EEC are: nouns specific to Energy Engineering academic texts using the British National Corpus as a reference; most common 4-grams in Zoology texts; adjectives containing *-friendly* in the whole EEC; or the most common verbs in Geology texts (Figure 10).

lemma (lowercase)		Frequency	Items: 1,195    Total frequency: 14,845	
P   N	form	631		
P   N	see	251		
P   N	cause	231		
P   N	occur	230		
P   N	move	216		
P   N	become	208		
P   N	determine	188		
P   N	represent	175		
P   N	produce	167		
P   N	rise	166		
P   N	make	154		
P   N	provide	142		
P   N	include	124		
P   N	take	108		
P   N	show	108		
P   N	change	108		
P   N	grow	105		
P   N	develop	105		
P   N	increase	104		
P   N	reach	98		
P   N	flow	98		

Figure 10: Frequency list of verbs in Geology texts.

## 4 Conclusion

In this paper, we have shown how the EEC was built and compiled and how it can be queried and exploited in Sketch Engine. The EEC's metadata, the default sketch grammar and the ESSG make the EEC a useful resource for any user interested in environmental science. As future work, we will refine, improve and update the ESSG and develop new rules for Spanish. Furthermore, in the short term, we plan to upload an improved version of the EEC (with more words and some minor codification issues solved) and a first version of the Spanish counterpart. In the long run, we will enhance the EEC with a new annotated version, where different semantic tags will be added to improve its querying potential. These semantic tags will include semantic categories and argument structure.



Sketch Engine's API also allows for the exploitation of the EEC from external applications. An example of this is EcoLexiCAT, a terminology-enhanced computer assisted translation (CAT) tool that provides easy access to domain-specific terminological knowledge in context (León-Araúz & Reimerink, 2018; León-Araúz, Reimerink & Faber, 2017). EcoLexiCAT integrates different features of the professional translation workflow in a stand-alone interface where a source text is interactively enriched with terminological information (i.e., definitions, translations, images, compound terms, corpus access, etc.) from EcoLexicon, BabelNet, IATE, and Sketch Engine. In the Sketch Engine module of EcoLexiCAT's interface, terms from both the source and target segments can be selected and direct access is given to concordances, CQL queries and word sketches of the selected terms. For a more detailed analysis, the output of the queries can be opened in a new tab that sends users to the website of the Sketch Engine Open Corpora.

## References

- Faber, P., León-Araúz, P. & Reimerink, A. (2016). EcoLexicon : New Features and Challenges. In *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with LREC 2016*, pp. 73–80.
- Jakubíček, M., Kilgariff, A., McCarthy, D. & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. *Proceedings of the PACLIC 24*, pp. 741–747.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36. <http://doi.org/10.1007/s40607-014-0009-9>
- Kilgariff, A., Kovář, V., Krek, S., Srdanovic, I. & Tiberius, C. (2010). A Quantitative Evaluation of Word Sketches. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress*. pp. 372–379. Leeuwarden/Ljouwert, The Netherlands: Fryske Akademy.
- Kosem, I., Gantar, P., Logar, N. & Krek, S. (2014). Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies. In A. Abel, C. Vettori & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*, pp. 355–364. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.
- León-Araúz, P. & Reimerink, A. (2018). Evaluating EcoLexiCAT: a Terminology-Enhanced CAT Tool. In *Proceedings of the 11th International Language Resources and Evaluation Conference (LREC2018)*. Miyazaki: ELRA.
- León-Araúz, P., Reimerink, A. & Faber, P. (2017). EcoLexiCAT: a Terminology-enhanced Translation Tool for Texts on the Environment. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pp. 321–341. Leiden: Lexical Computing.
- León-Araúz, P. & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In I. Kerneman & S. Krek (Eds.), *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, pp. 94–99. Miyazaki: Globalex.
- León-Araúz, P., San Martín, A. & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology*, pp. 73–82. Osaka.
- Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault, C. Jacquemin & M.-C. L'Homme (Eds.), *Recent advances in computational terminology*, pp. 279–302. Amsterdam/Philadelphia: John Benjamins.
- San Martín, A., Cabezas-García, M., Buendía Castro, M., Sánchez Cárdenas, B., León-Araúz, P. & Faber, P. (2017). Recent Advances in EcoLexicon. *Dictionaries: Journal of the Dictionary Society of North America*, 38(1), pp. 96–115.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), 1–13. <http://doi.org/10.1075/ijcl.10.1.01teu>

## Acknowledgements

This research was carried out as part of projects FF2014-52740-P, Cognitive and Neurological Bases for Terminology-enhanced Translation (CONTENT), and FFI2017-89127-P, Translation-oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.



# A Call for a Corpus-Based Sign Language Dictionary: An Overview of Croatian Sign Language Lexicography in the Early 21st Century

**Klara Majetić, Petra Bago**

*Faculty of Humanities and Social Sciences, Zagreb*

*E-mail: klaramajetic13@gmail.com, pbago@ffzg.hr*

## Abstract

Many sign languages today are still not standardized nor accessible to a wider audience. Sign languages with high quality dictionaries are scarce. In this paper we give a brief description of sign languages and how they differ from spoken ones, in order to better understand the issues lexicographers might face when compiling dictionaries for these languages. We focus on Croatian Sign Language (HZJ), giving an overview of the current situation in HZJ lexicography following some criteria we find relevant for both online and printed sign language dictionaries. The criteria have been classified into twenty-five categories and applied to create a model for an online HZJ dictionary, briefly presented in this paper. By presenting an unsatisfactory status of HZJ lexicography, we are issuing an urgent call for the compilation of a HZJ corpus as a basis for a high quality dictionary that could benefit both the potential hearing and deaf users.

**Keywords:** Croatian Sign Language (HZJ), sign language lexicography, dictionary evaluation, e-dictionary model

## 1 Introduction to Sign Languages and the Croatian Sign Language

Sign languages worldwide have not always been recognized as natural, genuine languages with an extensive vocabulary and complex grammar. It was only in the middle of the 20<sup>th</sup> century that the sign languages have been perceived to be as flexible and as expressive as spoken languages. The pioneer in this paradigm shift was William Stokoe and his linguistic research of the American Sign Language (ASL). His work has influenced other sign language researchers worldwide to investigate sign languages not as mere systems of gesture or simplified versions of their spoken counterparts, but as complex and independent natural languages in their own right. In order to better understand issues in lexicographical works on sign languages, a brief description of the Croatian Sign Language (HZJ) follows.

As with all sign languages when faced with their spoken counterparts, HZJ has a different modality from spoken Croatian – instead of an oral-auditory modality, sign languages have a spatial-visual modality (Kavčić 2012). A sign is a basic unit in sign languages, like a word in spoken ones. A sign consists of five elements: handshape, location, movement, orientation, and non-manual markers (mouthings, facial expressions, etc.). One of the first problems lexicographers face when approaching sign languages is the fact that they do not have a standard written form – there are a few notation systems (Stokoe notation<sup>1</sup>, *HamNoSys*<sup>2</sup>, and *SignWriting*<sup>3</sup> being some) and the option of glossing<sup>4</sup> (Filić

1 Stokoe 1960

2 Hanke 2004

3 URL: <http://www.signwriting.org/>

4 Glossing is a way of writing down signs using words – in HZJ glosses are verbs in infinitive form, nouns in singular and nominative case, adjectives in nominative case, singular and masculine grammatical gender.

2016). Notation systems use symbols and abstract pictures to describe a sign and all its elements, and are not written in a single line, but also use the vertical plane to add information (Stokoe 1960; Costa & Dimuro 2003). As a result of such writing requiring a lot of space on a page, traditional printed dictionaries had to have a rigorous lemma selection. Ordering of entries is an additional complication if lemmas are written in one of the mentioned notation systems, due to the lack of standardization. On the other hand, glossing conveys too little visual-spatial information – it gives very little information about sign elements because it assumes the user already knows what the sign looks like, and is therefore not intended for beginners. Sign languages provoke another problem by having several ways of modifying signs in different contexts. Those modifications provoke a question as to whether they should each be separate entries or all part of one entry. In recent years many researchers worldwide have been struggling with lexicographical issues regarding sign languages, such as Zwitserlood, Hedegaard Kristoffersen and Troelsgård (2013), Zwitserlood (2010), Capovilla (2003), Hanke and Storz (2008), König, Reiner and Langer (2004), among others. Most of the issues could be solved by utilizing new technologies rich with multimedia and thus moving on from the traditional printed dictionaries (Singleton 2000).

## 2 Current HZJ Lexicography

The current Croatian lexicography lacks a comprehensive sign language dictionary. In this section we describe existing dictionaries of HZJ using an instrument for evaluating (online) dictionaries of sign languages we have developed (Majetić & Bago 2017). The dictionaries are described according to the following criteria: intended users, subject field covered, type of dictionary according to the norm, number of languages in the dictionary, scope, function, direction, comprehensiveness, source and target language, data collection, data selection, multimedia, written form of the sign, other information regarding the sign, additional information in an entry, restrictions on content, content creators, content updates, content download feature, user interface design, searching and search results, browsing, and extra content.

At the moment HZJ has just one small traditional printed dictionary, *Hrvatski znakovni jezik*<sup>5</sup>, published in 2015. All other dictionaries are not standalone works but a part of textbooks, such as *Znak po znak*<sup>6</sup> 1-3 (ZPZ) that was published in 2006 and 2007, and *Gluhi i znakovno medicinsko nazivlje: kako komunicirati s gluhim pacijentom*<sup>7</sup>, published in 2010. The above-mentioned works are all alphabetical lists of words translated into HZJ by a photograph or an illustration. *Znak po znak* coursebooks are each accompanied by a DVD with video content, but it loads slowly and is somewhat difficult to navigate. HZJ used to have only one online dictionary, which is no longer fully usable, called *CroDeafWeb* – but the project has been abandoned and is no longer up to date with today's Internet browsers. The multilingual online dictionary *Spread the Sign* partnered up with a Croatian team in December 2015. To date they have added almost 10,000 entries into the dictionary out of the planned 15,000. A comparison of the number of entries in these works is given in Table 1.

In the following Sections 2.1, 2.2. and 2.3 we give an overview and a description of *Hrvatski znakovni jezik*, *CroDeafWeb* and *Spread the Sign* according to the abovementioned criteria. These three works have been chosen because they are not a part of textbooks, but instead are standalone dictionaries.

5 Eng. Croatian Sign Language.

6 Eng. Sign by Sign.

7 Eng. The Deaf and the Medical Sign Terminology: How to Communicate with a Deaf Patient.



Table 1: Approximate number of entries in various dictionaries of HZJ

Dictionary:	Number of entries:
<i>Znak po znak</i>	4,500
<i>Hrvatski znakovni jezik</i>	1,200
<i>CroDeafWeb</i>	500
<i>Gluhi i znakovno medicinsko nazivlje</i>	250
<i>Spread the Sign</i>	10,000

## 2.1 Hrvatski znakovni jezik

The second unaltered edition of this general traditional printed dictionary was published in 2015. The dictionary was created through an EU project “Poticanje zapošljavanja mladih jačanjem njihovih kompetencija za stjecanje boljeg položaja na tržištu rada”<sup>8</sup> to aid the education of future communication intermediaries in education for the deaf. It has been intended for hearing users as a general normative dictionary, with a primary function of helping in sign production. The dictionary is bilingual, monoscopal and monodirectional – the source language is Croatian and the target language is HZJ. The entries are listed in the alphabetical order of Croatian words. There are approximately 1,200 entries which is not enough for a comprehensive dictionary, especially considering that they have been chosen with no regard to the frequency of words nor signs. For example, one can find the word *žuboriti* (to murmur, to babble), but not *žvakati* (to chew). It is not known from where the data was collected, and there is no additional data within entries besides the translation. A word in Croatian is translated to HZJ by a photograph of a person signing. The editors are not professional lexicographers. As a first standalone HZJ dictionary of its kind this project is praiseworthy, and it was to be expected for a pioneer project to have some shortcomings. However, owing to the editors not being professional lexicographers, an inexperienced approach is unfortunately evident in all aspects of the dictionary. Nevertheless, due to the commendable work done by the editors by compiling the first standalone printed Croatian dictionary of sign language, it is now possible to analyze the pros and cons of the work, and use these insights to compile a new and improved dictionary of HZJ.

## 2.2 CroDeafWeb

*CroDeafWeb* was the first online dictionary of HZJ (<http://www.crodeafweb.org/rjecnik/index.html>) published around the year 2000 as a part of a same named portal (<http://crodeafweb.org/>). Unfortunately, it has not been updated for a long time, and the video content is no longer supported by today’s Internet browsers. The dictionary has a section where it uses gif format files, which still work but are not very detailed. Besides gifs, there is a short description of each sign’s elements, such as handshape and hand movement. It was probably intended for all groups of users as a normative dictionary with a primary function of helping in sign production. This dictionary seems to have been intended to be general, but it contains a special section Liturgy in Sign Language, and liturgical words are a big part of the dictionary. The dictionary is multilingual, monoscopal and monodirectional – the source languages are Croatian and English and the target languages are HZJ and Croatian. There is no search function. It is only possible to browse the entries, which is why they are listed in alphabetical order of Croatian and English words. There are also a few topics (e.g. food, days of the week) one can use to filter entries. The user interface design is simple, with a good

8 Eng. Promoting youth employment by strengthening their competencies to gain a better position in the labour market.

overview of the page. There are no options for typographic adaptations<sup>9</sup>. The home page contains some basic instructions on how to browse the dictionary, and the necessary system requirements needed to use it. The dictionary contains approximately 500 entries, which is again not sufficient for the average user. There are no restrictions prohibiting the users from saving the entries, but there is no explicit download button for this action. It is not known from where the data was collected and who the compilers were. There is no functioning extra content, but there are links to other sources that no longer work.

### 2.3 Spread the Sign

The online dictionary *Spread the Sign* (<https://www.spreadthesign.com/>) started in 2006 as a Swedish project that was initially funded by the European Commission and the Leonardo program under the slogan Life Long Learning. Today it is a general normative dictionary intended for all groups of users, with the function of helping in both sign reception and sign production. It contains 35 sign languages and their spoken counterparts, which makes it multilingual. The dictionary is monoscopical as it translates from a spoken into a sign language, but there are links within the entries that enable translation from one sign language into another. The dictionary is bidirectional inasmuch as it is intended for native users of any spoken or sign language. At the moment, there are over 380,000 signs from various sign languages – the goal is to have 15,000 entries in the dictionary for every sign language, which should satisfy the needs of the majority of users and thereby “make sign language available all over the world”.<sup>10</sup> Each language represented in the dictionary has a team in its country that works on compiling the dictionary and creating video content. These teams consist of sign language researchers and native speakers. The signs seem to be representative and authentic. The Croatian team joined the project in 2015, and have so far added 10,565 entries. Not all entries have gotten their HZJ translations, but it is possible to translate Croatian into other signed languages. Only the spoken languages are source languages, while the sign languages are target languages. It is not described what the first source for the entries list was, but we know other languages that were added over time followed the existing ones and added their translations, both in the spoken and the sign language. The only data all entries consist of is a gloss in a spoken language, a detailed video of signing and links to other sign languages that have added the same word and its translation of a sign. Some entries also have an illustration or a picture related to the meaning of the entry. Entries can be divided by topic such as: profanities, nouns, numbers, colors, music, business and so on. There is a special category with whole sentences ranging from *Are you deaf?* to *Do you want mustard, mayo, or both?* which are very useful for a user who is just starting to learn the sign language. All content is free, and even though there is no explicit download button within the entry, video can be saved to a user’s device. Users cannot create new content, but they can contact each country’s team to let them know if there are any mistakes in the translations. The content seems to be frequently added to and updated. The user interface is simply designed, and one can easily navigate between connected information, but there are no options of typographic adaptivity nor a user’s guide. The search function is monodirectional – users can only search from a spoken language to signed languages. The user has to first choose the language they want at the top of the page. The search result list contains all words that could match the searched word and the list of signed languages that the word can be translated to. The user can browse the content of the dictionary alphabetically and/or thematically. There is no extra content. This dictionary offers a lot to the average user, and with time might add a lot more information to its entries.

9 The deafblind use sign language too. Therefore, an e-dictionary should allow typographic adaptations (e.g. font size and type, contrast) to allow easier use for such users.

10 European Sign Language Centre (ESLC) Accessed at: <https://www.signlanguage.eu/en/about-us/> (10.3.2018.)

### 3 Model of an Online HZJ Dictionary

In this section we present the ideal online HZJ-Croatian dictionary following the evaluation criteria mentioned earlier. An ideal HZJ-Croatian dictionary is an online, general, descriptive, bidirectional and bilingual dictionary that takes the needs of the community into consideration. It is bicultural and intended for all groups of potential users. Its functions are directed to both text and sign reception and production. The ideal dictionary is based on a corpus that at the moment does not exist. Using such a corpus, the lexicographers can determine the authenticity and frequency of signs, and thereby choose the initial 5,000 entries that we see as the required minimum for the first publication of the dictionary. The corpus is the source for the multimedia. Each entry contains a video of the sign, a gloss, a textual description of the sign's elements, information on context or the use of the sign with examples, a definition of the meaning, ID number and the topic of the entry. The website clearly states how the content may be used and how it is protected. The ideal dictionary is free and available to all users. Distribution for non-commercial purposes is allowed, and so is downloading of the content. The main authors of the first edition are professional lexicographers. The users have a possibility to send feedback, corrections, and suggestions. Updates and revision of the existing content and adding of new content is done on a regular basis. The dictionary is easy to use and intuitive. The user interface design is simple, with options for typographic adaptivity. A user's guide, instructions for entry content, key to symbols, notations system and abbreviations are always accessible. All relevant content within the dictionary is connected by hyperlinks to make it easily and quickly accessible. The user has the option to search the dictionary by a Croatian word (spellcheck is an automatic part of the search tool) or by choosing elements of a sign and topic it might fall into. The list of search results contains the video of the sign, the gloss and the ID of the entry. The user is allowed to narrow down the search by adding new search criteria. It is possible to browse the entries alphabetically, thematically or by the sign elements. The necessary extra content are texts that describe HZJ, its grammar and the orthography used in the ideal dictionary.

In later editions the dictionary is expanded with less frequent entries, new entry content (e.g. geographical aspect), written form of signs using one of the notation systems, some more extra content such as words or signs of the day, educational games and other languages.

### 4 Conclusion

Croatian Sign Language (HZJ), like all sign languages and their spoken counterparts, has not yet been as thoroughly researched and described as Croatian has. There is a need for new learning resources in an online format rich with multimedia, as this is the only way to convey enough visual-spatial information of a sign language. Such a format also allows changes to be made if needed, easier additions to the dictionary and access for all potential users.

Sign languages that have built a corpus from which they can choose entries are very rare. Some existing corpora are Netherlands Sign Language (NGT) corpus, German Sign Language (DGS) corpus, Australian Sign Language corpus and a few others (Crasborn 2010). Having a corpus allows lexicographers to research the frequency of signs (which is not equal to the frequency of words in spoken languages), as well as research the use of each sign, and use the resulting discoveries to produce a better dictionary. HZJ has no existing corpus, and, to the best of our knowledge, no corpus is currently being compiled. Most of the lexicographical issues with sign languages can be handled by moving away from the traditional printed formats towards electronic dictionaries – images and videos can provide enough visual-spatial information, and such dictionaries are easier to use for all groups of users. By presenting the unsatisfactory current state of Croatian sign language

lexicography, we hereby issue a call for a much-needed corpus-based online dictionary of HZJ that will satisfy the needs of all user groups.

## References

- Capovilla, F. et al. (2003). Brazilian sign language lexicography and technology: Dictionary, digital encyclopedia, chereme-based sign retrieval, and quadriplegic deaf communication systems. In *Sign Language Studies*, 3(4), pp. 393-430.
- Crasborn, O. (2010). The Sign Linguistics Corpora Network: towards standards for signed language resources. *CroDeafWeb: rječnik hrvatskog znakovnog jezika*. Accessed at: <http://www.crodeafweb.org/rjecnik/index.html> [3.4.2017.]
- da Rocha Costa, A. C., & Dimuro, G. P. (2003). *SignWriting and SWML: Paving the way to sign language processing*. Atelier Traitement Automatique des Langues des Signes, TALN 2003.
- Filić, M. *Izražavanje količine u hrvatskom znakovnom jeziku*. Nacionalni repozitorij završnih radova (ZIR). Accessed at: <https://urn.nsk.hr/urn:nbn:hr:158:362319> [10.8.2017.]
- Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC* (Vol. 4).
- Hanke, T., Storz, J. (2008). iLex—A database tool for integrating sign language corpus linguistics and sign language lexicography. In *LREC 2008 Workshop Proceedings*. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Paris: ELRA pp. 64–67.
- Ristić, M., Baštijan, Z., Biškupić Andolšek, T. (Eds.). (2015). *Hrvatski znakovni jezik*. Zagreb: Hrvatski savez gluhih i nagluhih.
- Kavčić, D. (2012). *Hrvatski znakovni jezik: pregled opisanih jezičnih elemenata*. Diplomski rad. Filozofski fakultet Sveučilišta u Zagrebu. Accessed at: [www.darhiv.ffzg.unizg.hr/4454/](http://www.darhiv.ffzg.unizg.hr/4454/) [10.8.2017.]
- König, S., Reiner K., Langer, G. (2004). What's in a sign? Theoretical lessons from practical sign language lexicography. *Signs of the time*. In *Selected papers from TISLR*, pp. 379-404.
- Majetić, K., Bago, P. (2017). Proposing an Instrument for Evaluation of Online Dictionaries of Sign Languages. In *INFuture 2017 – Integrating ICT in Society*, pp. 189.
- Singleton, D. (2000). *Language and the lexicon. An introduction*. Arnold.
- Stokoe, W. C. (1960). Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. In *Studies in Linguistics: Occasional Papers* 8. Linstock Press.
- Šegota, I., Šendula-Jengiđ, V., Herega, D., Petaros, A., Conar, J. (2010). *Gluhi i znakovno medicinsko nazivlje: kako komunicirati s gluhim pacijentom*. Zagreb. Medicinska naklada.
- Tarczay, S., et al. (2006–2007). *Znak po znak 1, 2, 3: udžbenik za učenje hrvatskog znakovnog jezika*. Zagreb: Hrvatska udruga gluhoslijepih osoba „Dodir”.
- Zwitsersloot, I. (2010). Sign language lexicography in the early 21st century and a recently published dictionary of Sign Language of the Netherlands. In *International Journal of Lexicography* 23.4, pp. 443-476.
- Zwitsersloot, I., et al. (2013). Issues in sign language lexicography. In *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 259-283.

# Exploring the Frequency and the Type of Users' Digital Skills Using S.I.E.D.U.

**Stavroula Mavrommatidou**

*Democritus University of Thrace*

*E-mail: stavrmav@hotmail.com*

## Abstract

S.I.E.D.U. (Strategy Inventory for Electronic Dictionary Use) is a valid and reliable electronic instrument designed for assessing users' skills in electronic dictionary searches. It can be used for research purposes mainly for the detection of users' profiles in order to design appropriate intervention programs in classrooms. In the present paper, it has been used for collecting empirical data on users' dictionary skills, which is an important but poorly researched topic in language learning contexts. Seven hundred people (students from high schools and universities as well as teachers) participated in the investigation and completed the online questionnaire S.I.E.D.U., reflecting on their own digital dictionary use. It was found that not all users are familiar enough with the strategies required when using digital dictionaries and some of them lack the right skills to fully benefit from this useful source of information. In addition, there are differences in the skills applied by users depending on their level of education but not between university students in different study fields.

**Keywords:** digital lexicography, user skills, dictionary use

## 1 Introduction

Dictionaries – both printed and digital – are valuable language tools that satisfy different learner requirements and needs (Fuentes-Olivera 2009; Olmit 2010). They provide access to a vast amount of information about words and their use (Kobayashi 2008; Pousi 2010). Users can, for example, check the meaning, spelling, pronunciation, syntax, usage and etymology of a word, as well as find examples of use and synonyms or antonyms (Taylor & Chan 1994; Elola et al. 2008; Kuzmina & Rylova 2009). They may consult dictionaries to cover a linguistic information gap, to confirm lexical knowledge in cases of uncertainty and to learn what is considered correct (Lew 2010). This is because dictionaries often provide information about the language that cannot be found elsewhere (Koca et al. 2014), thus contributing to language acquisition (Walz 1990; Meitei et al. 2012), vocabulary enhancement (Scholfield 1987; Mdee 1997; Swanepoel 2000; Damascelli 2009) and text comprehension (Knight 1994; Fraser 1997; Watanabe 1997; Tono 2001; Constantinescu 2007; Omar & Mat Dahan 2011; Hamilton 2012; Shen 2013). Their usefulness in different (academic, professional and personal) contexts is therefore unquestionable (Prichard & Matsumoto 2011; Tüm 2012).

However, this does not mean that all users use dictionaries effectively. A number of studies (Christianson 1997; Chi 1998; Nesi 2000; Winkler 2001; Pousi 2010; Efthymiou 2013; Gavriilidou 2014) indicate that many lack the right skills, especially when using digital dictionaries. Compared to print dictionaries, the use of digital reference tools requires a higher level of knowledge and skills that need to be transferred to pupils. The vast content of linguistic sources, various retrieval options and multiple functions require complex reading skills and new strategies (Krajka 2007). It is not good enough to think that a student who often uses electronic reference tools is at the same time capable of doing it effectively (Wojtys 2009).



Therefore, the aim of the present research was to discover if pupils / students / teachers have similar skills when using online dictionaries, the type of strategies dictionary users in each one of the above categories seem to apply the most, and any possible differences between study fields.

## 2 Method

In order to check the strategies (e.g. lemmatization, search strategies, etc.) applied by users when consulting digital dictionaries in the field of digital lexicography, the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.) was used. S.I.E.D.U. is a valid and reliable tool of data collection (Gavriilidou & Mavrommatidou 2016). It is a digital self-report questionnaire, written in Greek and designed for users over 15 years of age. It contains 32 questions, concerning the following subscales:

- 1) Familiarity with different types of electronic dictionaries and the conditions of their use;
- 2) Strategies for lemmatization and acquaintance with dictionary conventions;
- 3) Navigation skills and
- 4) Look-up strategies in new electronic environments.

For a more detailed description of the way the instrument was constructed as well as its content, construct and discriminatory validity and reliability, see Gavriilidou and Mavrommatidou (2016).

S.I.E.D.U. was used in order to investigate the strategies users reported using when selecting a digital dictionary. 700 people completed the online questionnaire from February to December 2017 via Facebook. They were: 203 high school students (105 male and 98 female), 376 undergraduate students (96 male and 280 female) and 121 teachers (23 male and 98 female). They were all native speakers of Greek and their ages ranged from 15 to 54 years old. High school students came from different types of schools (ordinary, vocational, etc.), whereas undergraduate students and teachers studied/taught a number of many different subjects (mother tongue, foreign languages, math, economics, psychology, etc.) in various Greek cities (Thessaloniki, Komotini, Patras, Athens, and so on).

## 3 Results

IBM SPSS Statistics 23 was used for the analysis of results, and Welch's ANOVA was applied in order to investigate the differences people reported when using electronic dictionaries. Participants were grouped according to their level of education (High School Student, Undergraduate Student, Teacher) and the study fields (Science and Humanities).

The results examined by the level of education show statistically significant differences in the use of the electronic dictionaries in all four subscales (see Table 1).

Teachers and university students appear more familiar with the different types of electronic dictionaries and the conditions of their use compared to high school students. They also reported using the strategies for lemmatization and acquaintance with dictionary conventions as well as navigation skills to a significantly higher degree than pupils. The same is true in the case of look-up strategies in the new electronic environments, although all participants exhibited a low level of familiarity with these. In fact, teachers seem to use the most strategies of all, according to their statements.

Table 1: Differences in dictionary use strategies according to participants' level of education.

Scale	Level of education	<i>N</i>	Mean	Standard Deviation	ANOVA
Familiarity with different types of electronic dictionaries and the conditions of their use	HS	203	2.70	0.77	$F(4, 99,8) = 137.13$ $p < 0.001$
	US	376	3.47	0.58	
	T	121	3.70	0.53	
Strategies for lemmatization and acquaintance with dictionary conventions	HS	203	2.44	0.83	$F(4, 34,5) = 137.61$ $p < 0.001$
	US	376	3.01	0.71	
	T	121	3.01	0.78	
Navigation skills	HS	203	2.90	0.97	$F(4, 87,2) = 138.65$ $p < 0.001$
	US	376	3.73	0.88	
	T	121	4.17	0.82	
Look up strategies in new electronic environments	HS	203	2.30	0.83	$F(4, 87,2) = 138.65$ $p < 0.001$
	US	376	2.52	0.74	
	T	121	2.52	0.74	

As far as teaching subject is concerned,<sup>1</sup> the results do not show statistically significant differences in the use of the electronic dictionaries. Undergraduate students of humanities do not seem to have different skills compared to those who attend science courses (e.g. math, economics, physics, biology). The differences in all the skills applied (familiarity with different types of electronic dictionaries and the conditions of their use; strategies for lemmatization and acquaintance with dictionary conventions; navigation skills; look up strategies in new electronic environments) also seem to be rather small (see Table 2).

Table 2: Differences in dictionary use strategies according to participants' study field.

Scale	Study field	<i>N</i>	Mean	Standard Deviation	Independent-samples <i>t</i> -test
Familiarity with different types of electronic dictionaries and the conditions of their use	Science	84	3.45	0.56	$t(725) = -2.18$ $p = 0.03$
	Humanities	292	3.56	0.57	
Strategies for lemmatization and acquaintance with dictionary conventions	Science	84	2.95	0.79	$t(725) = -1.49$ $p = 0.16$
	Humanities	292	3.05	0.72	
Navigation skills	Science	84	3.92	0.82	$t(725) = 0.834$ $p = 0.41$
	Humanities	292	3.85	0.89	
Look up strategies in new electronic environments	Science	84	2.48	0.73	$t(725) = -0.775$ $p = 0.44$
	Humanities	292	2.54	0.74	

<sup>1</sup> In this measurement, the sample of high school students and teachers were excluded.

## 4 Discussion

According to the results, people do not all have the same skills or apply the same strategies when using digital dictionaries. The participants seemed to be more familiar with the first three types of strategies (especially with navigation skills, then with different types of electronic dictionaries and the conditions of their use, and lastly with strategies for lemmatization and acquaintance with dictionary conventions) than the fourth one (look-up strategies in new electronic environments). The respondents claimed that they are familiar with the different kinds of digital dictionaries (e.g. online, in DVD-ROM / CD-ROM or in a tablet), are able to use search engines or type specific URLs in order to find online lexicographical products, and can navigate easily between different parts of lexicographic data. Moreover, they prefer electronic dictionaries for their speed and ease of use, but they still use paper dictionaries in case the former have a low quality of information. On the other hand, they also claimed that they avoid online dictionaries available by subscription or dictionaries in a DVD-ROM or CD-ROM form. They also do not use the history menu, nor do they study the list of abbreviations or use complex search techniques (e.g. wildcards or phonological representations).

Competence seems to grow along with the level of education. Teachers thus seem to apply more strategies than younger students (of both high schools and universities), according to their statements. In contrast, the pupils reported the worst performance of all. Students, whatever subject they were studying, seemed to have average digital skills.

## 5 Conclusions

The present paper reports findings regarding users' strategies in electronic dictionary searches, as stated by 700 participants using the newly created online questionnaire S.I.E.D.U. The aim was check the type of strategies digital dictionary users seem to apply the most, and any possible differences between level of education and study fields. The results show that not all users are familiar enough with the strategies required when using digital dictionaries, confirming the literature review. In particular high school students lack the right skills to benefit from such digital resources, and are unable to use them in full. This may be contrary to our expectations that young people are really aware of technological achievements and computing applications in the field of language. Although they may have general knowledge concerning computer use, which allows them to find a lot of electronic lexicography products and navigate them relatively easily, they still need more complex search strategies in digital environments. Therefore, specially designed intervention programs aimed at teaching and practice of dictionary use in different subjects are definitely required at school.

## 6 Limitations of Research

Since S.I.E.D.U. is a self-report instrument, one cannot be really sure that the respondents' views, as expressed through the questionnaire, are their real and objective perceptions of the focal issues (Chamot 2004; Lew 2013). Users may not state what they do, but what they think they do, or what they think they ought to do (Hatherall 1984).

In addition, it is not certain that all participants define the categories in the same way, or that they answer honestly. Some users may be unwilling to answer specific questions concerning the frequency of dictionary use or the conditions and strategies of their lexicographic searches.

Finally, another problem is related to the composition of the sample. In fact, in the present investigation, there were far more women than men, as well as students of humanities than science. Therefore, more studies with bigger and more representative samples or with different research methods (e.g. observation studies) are needed. Investigations about the differences in the skills applied by users depending on their age or gender could also be useful.

## References

- Chamot, A. U. (2004). Issues in language learning strategy research and teaching. *Electronic Journal of Foreign Language Teaching*, 1(1), 12–25.
- Chi, M. L. A. (1998). Teaching dictionary skills in the classroom, Dictionary Use, in Fontenelle, T., Hilgsmann, P., Michiels, A., Moulin, A. & Theissen, S. (eds.), in: *Proceedings of the Eighth International Euralex Congress, Euralex 1998*, 565-577.
- Christianson, K. (1997). Dictionary use by EFL writers: what really happens? *Journal of Second Language Writing*, 6 (1), 2343.
- Constantinescu, A. I. (2007). Using Technology to Assist in Vocabulary Acquisition and Reading Comprehension. *The Internet TESL Journal*, 13 (2).
- Damascelli, A. T. (2009). Building a Bilingual Web-Glossary of Social Services Terms as Part of a Language Learning Environment, *eLEX2009 eLexicography in the 21st century eLexicography in the 21st century eLexicography in the 21st century: New challenges, new applications*. 22-24 October 2009, 47-48.
- Efthymiou, A. (2013). *Teaching the vocabulary in primary school*. Theory and Practice. Thessaloniki: Epikentro (In Greek).
- Elola, I., Rodríguez-García, V. & Winfrey, K. (2008). Dictionary use and vocabulary choices in L2 writing. *Estudios de Lingüística Inglesa Aplicada*, 8: 63-89. Accessed at: <http://institucional.us.es/revistas/elia/8/6.%20elola%20def.pdf> [10/2/2018].
- Fraser, C. A. (1997). *The impact of lexical processing strategy instruction on L2 readers' strategy use, reading rate, reading comprehension, and vocabulary learning*. Unpublished doctoral dissertation, University of Toronto (OISE), Toronto, Ontario.
- Fuertes-Olivera, P. A. (2009). The Function theory of lexicography and electronic dictionaries: Wiktionary as a prototype of collective multiple-language Internet dictionary. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.), *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographica Tools Tomorrow* (pp. 99-134). Bern: Peter Lang.
- Gavriilidou, Z. (2014). User's abilities and performance in dictionary look-up. Στο Lavidas, N. Alexiou, T & A. Sougari *Major Trends on Theoretical and Applied Linguistics* Vol. 2, 41-52.
- Gavriilidou, Z. & Mavrommatidou, St. (2016). Construction of a tool for the identification of electronic dictionary users' skills: test specification and content validity, *XVII Euralex International Congress, 6-10 September 2016*, Tbilisi, 168-178.
- Hamilton, H. (2012). The efficacy of dictionary use while reading for learning new words. *American Annals of the Deaf*, 157(4), 20.
- Hatherall, G. (1984). Studying dictionary use: Some findings and proposals. In: Hartmann, R.R. (ed.), *LEXeter '83 Proceedings: Papers from International Conference on Lexicography at Exeter, 9- 12 Sept. 1983, Lexicographica Series Maior 1*, 183-189. Tübingen: Niemeyer.
- Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal*, 78, 286–298.
- Kobayashi, C. (2008). *The Role of Pocket Electronic Dictionaries in EFL Learning*, 6, 2, 103-122.
- Koca, S., Pojani, V. & Jashari-Cicko, A. (2014). Dictionary use by EFL University Students a case study at Korca University, *Mediterranean Journal of Social Sciences*, Vol 5 No 19 August 2014, MCSER Publishing, Rome-Italy.
- Krajka, J. (2007). Online Lexicological Tools in ESP – Towards an Approach to Strategy Training, *Scripta Manent* 3(1), 3-19.
- Kuzmina, V. & Rylova, A. (2009). Software Demonstration. The ABBYY Lingvo Electronic Dictionary and the ABBYY Lingvo Content Dictionary Writing System as Lexicographic Tools, *eLexicography in the 21st century: New challenges, new applications*, Centre for English Corpus Linguistics Université catholique de Louvain, 22-24 October 2009, 131-133.

- Lew, R. (2010). Multimodal lexicography: The representation of meaning in electronic dictionaries. *Lexikos* 20: 290-306.
- Lew, R. (2013). From paper to electronic dictionaries: Evolving dictionary skills' in Kwary, Deny Arnos, Nur Wulan and Lilla Musyahda (eds.), *Lexicography and Dictionaries in the Information Age. Selected papers from the 8th ASIALEX international conference*. Surabaya: Airlangga University Press, 79-84.
- Mdee, J. S. (1997). Language Learners' Use of a Bilingual Dictionary: A Comparative Study of Dictionary Use and Needs, *Lexikos* 7, 94-106.
- Meitei, S. P., Ningombam, S. & Purkayastha, B. S. (2012). Word Search in a www Manipuri-English Electronic Dictionary, *International Journal of Computer Applications*, 55(3), 34-37.
- Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries*. (Lexicographica Series Maior 98.) Tübingen: Max Niemeyer Verlag.
- Olimat, S. (2010). *An Analytic Study of Four Dictionaries Used by Jordanian Translators*. Unpublished MA thesis, Faculty of Arts, Yarmouk University. Irbid, Jordan.
- Omar, C. A. M. B. C. & Mat Dahan, H. B. A. (2011). The Development of E-dictionary for the use with Maharah Al-Qiraah Texxtbook at a Matriculation Centre in a University in Malaysia, Tojet: *The Turkish Online Journal of Educational Technology*, 10(3), 255-264.
- Pousi, A. (2010). *Training in Dictionary Use: A teaching intervention in a 9th grade EFL classroom in Finland*, Bachelor's thesis, University of Jyväskylä, Department of Languages, 24.5.2010, 1-28.
- Prichard, C. & Matsumoto, Y. (2011). The Effect of Lexical Coverage and Dictionary Use on L2 Reading Comprehension, *The Reading Matrix*, 11 (3), 207-225.
- Scholfield, P. (1987). *Vocabulary Problems in Communication: What Determines the Learner's Choice of Strategy*, Bangor Teaching Resource Materials in Linguistics.
- Shen, Z. (2013). The Effects of Vocabulary Knowledge and Dictionary Use on EFL Reading Performance, *English Language Teaching*, 6 (6), 77-85.
- Swanepoel, P. (2000). Providing lexicographic support for SL vocabulary acquisition: What kind, under what conditions, for whom, and why? *Proceedings of Euralex 2000*, 403-417.
- Taylor, A. & Chan, A. (1994). Pocket Electronic Dictionaries and their Use, In Willy Martin et al. (eds), *Proceedings of the 6th Euralex International Congress*. Amsterdam: Vrije Universiteit, 598-605.
- Tono, Y. (2001). *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension*. Tübingen: Niemeyer.
- Tüm, G. (2012). Impact of Dictionary Type and Usage to Enhance Turkish Vocabulary in Teaching Turkish as a Foreign Language. *Turkish Studies International Periodical for the Languages, Literature and History of Turkish or Turkic*, 7(4), 3013-3023.
- Walz, J. (1990). The dictionary as a Secondary Source in Language Learning. *The French review* 64 (1), 79-94.
- Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, 19, 287-307.
- Winkler, B. (2001). Students Working with an English Learner's Dictionary on CD-ROM, *Information Technology and Multimedia in English Language Teaching Conference*: 227-254 Hong Kong.
- Wojtyś, E. (2009). Towards effective dictionary use by L2 learners: in search of new perspectives, *Lublin Studies in Modern Languages and literature*, Maria Curie-Skłodowska University, Lublin, Poland, 169-179.



# From Standalone Thesaurus to Integrated Related Words in *The Danish Dictionary*

**Sanni Nimb, Nicolai H. Sørensen, Thomas Troelsgård**

*Society for Danish Language and Literature*

*E-mail: sn@dsl.dk, nhs@dsl.dk, tt@dsl.dk*

## Abstract

This paper presents a method of integrating Danish standalone thesaurus data automatically into a monolingual dictionary of modern Danish (*Den Danske Ordbog*, ‘The Danish Dictionary’) and discusses the results, including some of the problematic cases. The method draws on the detailed semantic grouping with two types of keywords in a well-structured XML-manuscript of a recently published thesaurus of Danish (*Den Danske Begrebsordbog*, ‘The Danish Concept Dictionary’) and on the fact that the two resources are linked on sense level, allowing for the automatic identification of semantically related thesaurus extracts for any given sense in the dictionary. The paper also presents a study of similar integrations of thesaurus data in four online English dictionaries, namely the *Oxford English Dictionary*, the *MacMillan Dictionary*, the *Merriam-Webster English Dictionary* and the *Oxford Dictionaries: English*, which we carried out in order to compare the structure of the underlying English thesaurus data as well as the resulting dictionary presentations with the Danish case.

**Keywords:** thesaurus, dictionary, linked data, synonyms

## 1 Introduction

This paper presents and discusses the automatic extension of a monolingual Danish online dictionary (*Den Danske Ordbog*, henceforth the DDO dictionary) with more synonyms and semantically related words based on the automatic identification and extraction of data from a Danish thesaurus (*Den Danske Begrebsordbog*, henceforth the DDB thesaurus), a dictionary describing concepts of modern Danish published in print in 2014. The two dictionaries are both compiled at the Society for Danish Language and Literature (DSL) and closely related since the vocabulary of the DDB thesaurus is based on and linked to sense descriptions of the DDO dictionary (see Nimb et al. 2014). Both resources are continuously being extended with new words.

The underlying XML data of the DDB thesaurus manuscript is annotated with coarse-grained semantic types allowing for the identification of persons, animals, artifacts, acts, events, etc. These annotations have already proved to be useful when compiling different types of formal lexicons (Nimb & Pedersen 2012, Nimb et al. 2013, Nimb et al. 2017). The future plan is to publish an online DDB thesaurus based on the printed book where it is possible to browse through the hierarchy of named chapters and sections, and to study the full vocabulary of each section with direct access from the words to the definitions and entries in DDO. But in this paper we describe the first online dictionary use of the linked data the other way around: excerpts of words from the DDB thesaurus sections inserted into the entries of the online DDO dictionary in order to supply the user of DDO with more information on related words. This type of function in online dictionaries is not new, and in Section 2 we study how a number of different English dictionaries have incorporated related words from thesaurus data, before in Section 3 we turn to the detailed presentation of the structure and content of the DDB thesaurus. In Section 4 we present the method used to identify and extract the relevant thesaurus data. In Section 5 we discuss some of the problematic results and draw some conclusions. We start by giving an overall presentation of the DDB thesaurus and the task.

## 1.1 The DDB Thesaurus

The DDB thesaurus is not written bottom up with groups of synonymous words as the starting point, but top down with a thematic structure in the form of named chapters and sections as the starting point, inspired by Dornseiff (2004), and based on the vocabulary in the DDO dictionary. However, the XML-structure of the manuscript was from the very beginning intended to facilitate the transfer of semantically related data back to the same dictionary. Due to this, the thesaurus presents words in a semantic order to the lowest subgroup level in the structure, meaning that the nearest other word to both sides of a word in the manuscript is most likely to be either the most similar synonym or near-synonym. The printed manuscript is in many ways comparable to the well-known English work *Roget's Thesaurus* (2002). Like the DDB thesaurus, *Roget's Thesaurus* divides the vocabulary in a series of named chapters and sections which contain semantically ordered groups of words which are either synonyms, or co-hyponyms or otherwise semantically related, although always belonging to the same word class. Likewise, *Roget's Thesaurus* also marks certain words as keywords based on semantics (i.e. a hypernym or the most common word in a group of synonyms). To illustrate the semantic structure of the Danish DDB thesaurus, we transfer its structuring principles to a group from *Roget's Thesaurus*, namely to the group of soft drinks. The group is in *Roget's Thesaurus* initiated by the word *soft drink*, underlined as a keyword in the text with italic letters. Semicolons divide subgroups of (sometimes synonymous) words within the group. All words are semantically, not alphabetically ordered:

“*soft drink*, teetotal d., nonalcoholic beverage; water, drinking w., filtered w., eau potable, spring water, fountain; soda water, soda, cream s., soda fountain, siphon; table water, carbonated w., mineral w., Perrier (tdmk), tonic water, barley w., squash, low calorie drink, mixer; energy drink; iced drink, frappé; milk, milk shake; ginger beer, ginger ale, Coca Cola or Coke (tdmk); fizz, pop, lemonade, orangeade, bitter lemon; cordial, fruit juice, orange j., apple j., tomato j., vegetable j.; juice box; coconut milk; tea, iced t., lemon t., herbal t., char, pekoe, orange p., Indian t., China t., green t., black t., Russian t., herb t., maté, coffee, café au lait, café noir, black coffee, white coffee, decaffeinated coffee, decaf, Irish coffee, Turkish c., espresso, cappuccino, latte, cocoa”

Based on the structuring principles of the DDB thesaurus, the group would contain not only one, but two types of keywords: *soft drink* as a keyword at first level (in bold letters), and a number of words in the text marked as keywords at second level (in bold, italic letters), namely (at least) *water*, *soda water*, *tea* and *coffee*, maybe also *fruit juice*. Furthermore, there would be supplementary divisions and therefore more subgroups (e.g. between *maté* and *coffee*). We also find that the keywords as well as the subgroups (divided by ●) are more transparent in the Danish thesaurus. Here we present the *Roget's Thesaurus* data as they would look in the DDB thesaurus structure:

“**soft drink**, teetotal d., nonalcoholic beverage ● **water**, drinking w., filtered w., eau potable, spring water, fountain ● **soda water**, soda, cream s., soda fountain, siphon ● table water, carbonated w., mineral w., Perrier (tdmk), tonic water, barley w., squash, low calorie drink, mixer ● energy drink ● iced drink, frappé ● milk, milk shake ● ginger beer, ginger ale, Coca Cola or Coke (tdmk) ● fizz, pop, lemonade, orangeade, bitter lemon ● cordial, fruit juice, orange j., apple j., tomato j., vegetable j. ● juice box ● coconut milk ● **tea**, iced t., lemon t., herbal t., char, pekoe, orange p., Indian t., China t., green t., black t., Russian t., herb t., mate ● **coffee**, café au lait, café noir, black coffee, white coffee, decaffeinated coffee, decaf, Irish coffee, Turkish c., espresso, cappuccino, latte, cocoa”

Our goal is to be able to present the users of the DDO dictionary with a small extract of the most related words and expressions from the DDB thesaurus next to the sense definition, and eventually already existing synonyms (without having to activate a link), such as some of the other types of coffee

when the word *espresso* is looked up. But we also want to allow the user to be able to activate a link to boxes presenting a much larger variety of words and expressions from the DDB thesaurus, namely any noun which is related to *espresso* in the sections where the word (in that specific sense) occurs, in this case it would be a box with the entire group of *soft drinks*, since this is the keyword (explanatory headline) of the coffee keyword.

The challenge we deal with is that the automatic extraction must cover any type of word (e.g. keywords at first or second level, or words which are not a keyword) as well as any size of the word group – from one to maybe 30-40 words. While some words have no direct synonymous or near-synonymous neighbors (e.g. *energy drink*), others have far more than we want to show in the small extract directly in the entry (e.g. *espresso*). Furthermore, we have no indication in the DDB thesaurus of whether the subgroup (or whole group) consists of synonyms (like “*Coca Cola or Coke*”), or words which are semantically related in other ways, e.g. being co-hyponyms as in the case of soft drinks above, or just somehow thematically related. The thesaurus contains no information on direct synonymy, neither on co-hyponymy. Our task is furthermore complicated by the fact that some of the DDO sense descriptions, but far from all, already contain a few manually selected synonyms, near-synonyms and/or antonyms that we do not want to repeat.

Before we describe the thesaurus structure and the chosen transfer method in detail, we take a closer look into how other dictionaries have integrated thesaurus data, taking into consideration also which type of thesaurus data they had at their disposal in order to see whether or not it resembles the DDB thesaurus data. We chose to study a number of English dictionaries since the language is closely related to Danish.

## 2 Thesaurus Data Integrated in English Dictionaries

We studied four English dictionaries which have either integrated thesaurus data in the entries, or present links to thesaurus data, namely the *Oxford English Dictionary* (OED), *MacMillan Dictionary*, the *Merriam-Webster English* online dictionary, and *Oxford Dictionaries:English*<sup>1</sup>.

In the comprehensive *Oxford English Dictionary* (OED) there is a link from each sense description to thesaurus data taken from the standalone *Historical Thesaurus of the Oxford English Dictionary* but presented in OED style. The thesaurus is based on the information in the OED and linked to its senses. It organizes English words throughout the history into detailed hierarchies of meaning. In the OED the user is presented with a list of related words (including the headword itself) ordered historically, i.e. based on information on year of first recorded use with the oldest word first. The list is initiated by a headline presenting the hierarchy of chapters and section levels to which the word belongs in the historical thesaurus. For example, in the case of the adjective *high* in the sense ‘high-necked’, the user is presented by the headline ‘the world > textiles and clothing > clothing > types or styles of clothing > [adjective] < having specific parts < neckline’ to the list of adjectives: *high*, *low*, *low-necked*, *décolleté*, *semi-high* and *turtle-necked*. In the case of *happy* which has many synonyms, these are initiated by the headline ‘the mind > emotions > pleasure > happiness > happy’. The list of words is in this case long, and due to the historical order the most common words of modern English meaning ‘happy’ are spread in between rare synonyms and words which are not used any more.

In the *MacMillan Dictionary* there is again a link to a thesaurus presenting related words which are, also in this case, introduced by explanatory headlines. But the thesaurus behind the data consists in fact only of named groups of synonyms which are not organized in meaning hierarchies. But in

<sup>1</sup> We chose to leave out Dictionary.com which sometimes, but not always presents synonyms directly in the entries (e.g. for *hard* and *happy*, but not for *décolleté*, and, more surprisingly maybe, not for *stupid*). The dictionary links to more synonyms in Thesaurus.com, however the synonyms in the dictionary itself do not seem to come from Thesaurus.com.

contrast to the OED, the headline and a small excerpt of the data (three words) are presented directly in the dictionary entry itself. In the case of *décolleté*, the headline ‘Words used to describe clothes’, and the adjectives *A-line*, *backless*, *baggy* are presented, ordered alphabetically. Had the words been ordered on the basis of semantics, it would have been possible to present instead the most closely related words of *décolleté* (e.g. *low* with almost the same sense). The dictionary instead chose to simply present the first three words of the alphabetically ordered list. However, we do find cases in *MacMillan* where the synonyms are listed in semantic order, both in the integrated excerpt and when we click on the thesaurus link, e.g. in the case of *happy* where the adjectives ‘*happy*, *glad*, *alive*’, are presented directly in the entry, introduced by the headline ‘Feeling happy’.

**The Merriam-Webster English online dictionary**, which presents a semasiological dictionary and a thesaurus on the same homepage, contains one, maybe two precise (manually inserted) synonyms in the text with links to their sense description. The top lines of the thesaurus data are directly visible at the bottom of the whole entry (not at sense level), in the form of a few close synonyms with the headline ‘Synonyms’. When the link in the field is activated, the rest of the box becomes visible. The related words are presented in groups with the headlines ‘Synonyms’, ‘Antonyms’, ‘Near antonyms’, and ‘Related words’. In each of these, the word order is alphabetic, not semantic. The list of the synonyms of *happy* consists for example of *blissful*, *delighted*, *glad*, *gratified*, *joyful*, *joyous*, *pleased*, *satisfied*, *thankful*, *tickled*, thereafter a group of antonyms is presented before the near-synonyms are visible. Also in this case, the thesaurus data are not organized hierarchically based on meaning. As the only dictionary, *Merriam-Webster* gives a very detailed description of the relation between the different synonyms of the word and how they are used for different purposes at the bottom of the entry.

Finally, we take a look at the online **Oxford Dictionaries:English**. In the entry of a lemma, right below the definition, there is a link to synonyms. When it is activated, synonyms (or near-synonyms in a narrow sense) are presented in a box. Other types of semantically related words, such as co-hyponyms or antonyms, are not included in the presentation. Due to this, there is no data for *décolleté* – it has no synonyms – but a lot of data for *happy*: *contented*, *content*, *cheerful*, *cherry*, *merry*, *joyful*, *joyous*, *jolly*, *joking* and so on. There might also be more than one group of synonyms presented. In the case of *stupid*, for example, there are two: one initiated by the keyword *unintelligent*, another initiated by the keyword *foolish*. There is a link from the box to the thesaurus homepage itself, but like in the case of both *MacMillan Dictionary* and *Merriam-Webster*, the thesaurus data consists only of groups of related words which are not organized hierarchically, initiated by a bold headword (e.g. **contented**). Table 1 gives an overview of the different dictionaries, and compares the thesaurus data with the DDB thesaurus.

Two of the English dictionaries present a small extract of thesaurus data directly in the dictionary entry before linking to a larger group of related words. In one case the extract is introduced by a headline. Two dictionaries have at least to some degree ordered the thesaurus words semantically, but none as detailed as in the DDB thesaurus. Only one English dictionary contains manually selected synonyms as part of the dictionary sense description itself like the DDO dictionary, and only one, the OED, bases the integration on a thesaurus with meaning hierarchies similar to the ones we find in the DDB thesaurus, allowing for the automatic extraction of precise headlines of related words. In DDB however, we have no labels at the lowest grouping levels (for *décolleté* corresponding to ‘types or styles of clothing < having specific parts < neckline’). Two of the dictionaries, the *Merriam-Webster English Dictionary* and the *Oxford Dictionaries:English*, make a clear distinction between presenting synonyms (in a broad sense) and other related words, allowing them to give the thesaurus data in either separate groups, or to leave out completely words not being either a synonym or a closely related near-synonym. The data of the DDB thesaurus lacks the information which is needed to present synonyms and near-synonyms in a narrow sense, and less related words such as near-synonyms in a broad sense (for example co-hyponyms and thematically related words).



Table 1: Overview of compared thesaurus integrations.

Based on thesaurus with:	DDO	OED	MacMillan	Merriam-Webster	Eng. Oxford Dic
Meaning hierarchy	Yes	Yes	No	No	No
Distinction in thesaurus data between synonyms and other related words	No	No	No	Yes	Yes (only synonyms are shown)
Semantic order at lowest group level	Yes	No, historical	Sometimes	No, alphabetic	Sometimes
<b>The solution in the dictionary contains:</b>					
Small extract of directly shown data	Yes	No	Yes (with headlines)	Yes	No
Link to large extract of data in box	Yes	Yes	Yes	Yes	Yes
Large extract in box has explanatory headlines	Yes	Yes	Yes	No	No (but headword in bold)
<b>Already manually selected related words to be considered</b>	Yes	No	No	Yes	No

The solution we chose in the case of the integration of DDB thesaurus data into the DDO dictionary is most similar to the *MacMillan Dictionary*, although without an explanatory headline of the small, directly shown extract of closely related words. But in contrast to *MacMillan* we have to consider the already manually inserted synonyms, near-synonyms and antonyms in DDO: these must be left out from the small extract. Furthermore our headlines are most likely to be less precise, since they are transferred automatically from the title of the section in which the word group is only one of several groups. In *MacMillan*, each group of synonyms and near-synonyms has probably been assigned headlines manually. We will instead profit from the many keywords in the thesaurus data, which in most cases function quite well as headlines of the following group of words, as in the case of *Oxford Dictionaries:English*. In some cases, however, they do not, and we will discuss this in Section 5. In the next two sections we give a detailed presentation of the structure of the DDB thesaurus and the method of extracting data from the thesaurus to the dictionary DDO.

### 3 The Structure of the DDB Thesaurus

The DDB thesaurus is organized into 22 named chapters and 888 named sections. It contains approximately 200,000 words and expressions, almost all of which are linked to a specific sense in the DDO dictionary<sup>2</sup> by shared ID numbers, in the case of collocational expressions to at least one of the included word senses (see Nimb et al. 2014). The overall semantic grouping principles in each section of the thesaurus are based on the distinctions between 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> order entities (Lyons 1977) and between semantic types and relations (Pustejovsky 1995). In each of the 888 sections a number of subgroups of words, typically sharing the same semantic type, are listed one after the other in five word class groups ('noun', 'verb', 'adjective', 'adverb' and 'other'). The semantic order of the words in a subgroup is based on linguistic characteristics such as prototypicality, frequency, broad versus narrow meaning, style, etc., based on the descriptions in DDO as well as the subjective judgment of

<sup>2</sup> DDO contains approximately 140,000 sense descriptions



the lexicographer who weighs up the different characteristics against one another when the order is established. The subgroups might consist of co-hyponyms or maybe just thematically related words, not only synonyms. Some of the subgroups are initiated by a keyword, as described in Section 1.1. First level keywords indicate large shifts in meaning between the semantic types in the subgroups, for instance between artifacts and persons. Second level keywords indicate meaning shifts within the same semantic type, e.g. between different types of artifacts (like *tea* and *coffee* as described above). First level keywords function as a kind of ‘headline’ until the next one, while second level keywords do the same, but only until the next keyword, no matter whether it is a first or second level one. In Figure 2 we illustrate the structure in a formal way. The first word, ‘A’, functions as the headline of all the words until ‘X’. The words in the first group, ‘A, b, c, d’ are more closely related to one another than they are to the words in the next group, ‘e, f, g, h’. Likewise, ‘j’ and ‘k’ are more closely related to ‘i’ and ‘m’ in their own group than to ‘g’ and ‘h’ in the proceeding group, although still closer to ‘g’ and ‘h’ than to ‘s’ and ‘t’ in the preceding group, which starts with a keyword, ‘r’.

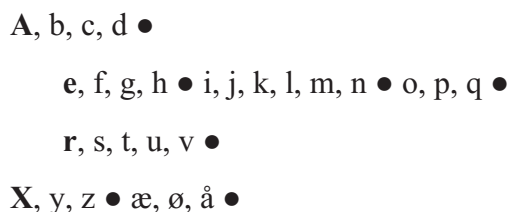


Fig. 2: The structure of the DDB thesaurus data. Each letter represents a word.

During the dictionary-making process of the thesaurus we found the structure very flexible. It allowed us to group words without having to specify the degree of synonymy, since this demands a rather detailed study of the use of the words. Instead we aimed at including as many words as possible from the DDO dictionary, i.e. also those which do not have any real synonyms. We focused on the communicative purposes of the word order, trying to make a fluent ‘text’ with as slight a change in meaning as possible from one word to the next. Whenever there was a big meaning change, the word was changed into a keyword. The drawback of this flexible structure is that we cannot identify the synonyms and closest near-synonyms from groups of other related words in the manuscript, and we are therefore not able to present automatically extracted very precise synonym groups in the same way as e.g. *Merriam-Webster English* online dictionary. The related vocabulary that we extract from the DDB thesaurus is most likely to consist of either synonyms, near-synonyms or co-hyponyms, but might also contain a broader variety of related words than any of the English dictionaries do.

#### 4 Extraction of Data from the Thesaurus to the DDO Dictionary

In order to automatically identify exactly which DDB thesaurus data to extract, the semantic relevance of the surrounding words in a section where the DDO sense in question is represented is dynamically calculated. The calculation is based on the semantic order of the words as well as on keywords and subgroup structure. We benefit from the fact that the closest words in the DDB thesaurus structure within the same subgroup are most likely to be also the semantically closest word, and that the first keyword to the left is almost certainly closely related, since it was chosen as a kind of headline of the word.

Initially we determine the ‘best’ section in the case of multiple occurrences in the DDB thesaurus, based on the number of words within the scope of the first keyword to the left: the higher the number,

the better (i.e. the more likely to contain most synonyms and near-synonyms of the word sense). We then select one to six words from the immediate surroundings of our headword (first the first one to the left, then the first one to the right, then the second one to the left and so forth), include the key-word to the left but exclude the headword itself and eventually manually inserted DDO synonyms. We present the words as a list integrated directly in the DDO entry, in the DDO style. Secondly, we present a link which gives access to an even larger extract of words, presented in a box initiated by the corresponding section name in DDB and with the search word highlighted in red. In the case of multiple occurrences of the specific sense of the search word, data from different DDB sections is presented in separate boxes which are listed according to their section number in the DDB thesaurus. See Figure 3.

**slentre** verbum  
BØJNING -r, -de, -t  
UDTALE ['slendɐ]  
OPRINDELSE svensk dialekt *släntra*, nedertysk *slentern* • dannet til roden i jysk *slante* 'dingle, drive omkring'

**Betydninger**

spadsere i et langsomt og afslappet tempo  
ORD I NÆRHEDEN **NYT** bevæge sig langsomt | gå | gå langsomt | lunte | trisse  
| tulle | tusse ...vis mere

GRAMMATIK NOGEN slentrer (+RETNING/+STED) HJÆLPEVERBUM være og have  
vi ville slentre stille og roligt ind mod centrum og således nyde den første dag i juni fuldtud JeHoNi89

**BEVÆGELSE**  
gå, være til fods, være på gåben, spadsere, promenere, **slentre**, trippe af sted, trave, marchere, føre sig, skride af sted, stoltse, spankulere omkring, svanse • bevæge sig langsomt, trisse, luffe, daffe, jkke, vade rundt, sjokke, sjoske, sjaske, smatte  
fra Den Danske Begrebsordbog, kapitel 8

**LANGSOM BEVÆGELSE**  
bevæge sig langsomt, snegle sig af sted, krybe, gå, gå langsomt, lunte, **slentre**, trisse, tulle, tusse, daske, daffe, dappe, dalre, drysse, luffe, sjokke • vakle af sted, famle sig frem • trille af sted • slæbe sig af sted, halte, slæbe på benene, slæbe på fødderne/benene, hinke, humpe, vralte, rokke  
fra Den Danske Begrebsordbog, kapitel 8

**FØRNOJELSER OG FRITIDSAKTIVITETER**  
gå, gå lange ture, trave, gennemtrave, gennemvandre, lufte (sin) hund, promenere, spadsere, **slentre**  
fra Den Danske Begrebsordbog, kapitel 17

Fig. 3: To the left, the verb *slentre* ('stroll around (in a relaxed manner)') in the DDO dictionary, extended with the direct presentation of seven near synonyms from DDB, including a link '*vis mere*' ('show more'). To the right, the link has been activated, showing three boxes with more related words, based on the three occurrences of the sense in the DDB thesaurus: 1) words from the section "*Bevægelse*" ('movement'), 2) words from the section "*Langsom bevægelse*" ('slow movement'), and 3) words from the section "*Fornøjelser og fritidsaktiviteter*" ('leisure').

The extract from each box is calculated dynamically, based on an algorithm telling us how far to the left and right of the search word we should go in DDB if we intend to present the words from the same word class which are most similar in meaning. Keywords on the first level always constitute the borderline, indicating a shift to a new semantic type. Keywords on the second level, including the words within their scope, are sometimes, but not always, included in the extract, depending on the type of search word, which may itself be a keyword on the first or the second level. The boxes include synonyms already presented in the DDO entry.

## 5 Results and Conclusion

In most cases, the method results in very useful extracts of related words, as for example seen in Figure 3, where we get three boxes of words altogether reflecting very well the different aspects of the verb *slentre*. We have received a lot of positive user feedback since the thesaurus function was released at the beginning of January 2018. But there are also some problems, especially when it comes to the small automatically inserted extract. One challenge was to decide whether to include the headword in the extract or not, and we discussed at length how to define the small extract (how many words to show, how to find the best box to extract them from, etc.). When comparing to the English dictionaries which mostly leave out the small extract when a few synonyms have already

been described manually, it is maybe worth reconsidering whether to leave it out, and link directly to the boxes.

The many cases of collocations in the DDB thesaurus also cause problematic results. As an example, the expression *glad dreng* ('happy young man') in the DDB thesaurus is linked to the adjective *glad* ('happy') in the XML-structure, not to the noun *dreng* ('young man') in the DDO dictionary. This results in a box in the entry of *glad* with words that are in fact near synonyms of 'young man' (next to three boxes presenting very useful related words of *glad*, we should mention). Negated senses in the thesaurus also cause problems. The DDO dictionary sometimes describes the 'positive' sense of a word which rarely occurs without a negated context, then mentioning this as a constructional comment. But in the thesaurus the word is presented negated, since it is hard to understand it without the negated context. This leads of course to the extraction of antonyms instead of synonyms. We solved these problems by presenting the expression (e.g. *glad dreng*) in the box headline, but in some cases we might consider changing the underlying data

The groups in the DDB thesaurus which consist of co-hyponyms or thematically related words do not always give good results when they are presented as related words in the DDO dictionary. This was foreseen to be a problem since we lack information on whether the group consists of co-hyponyms or rather of synonyms. Especially when keywords were introduced in the thesaurus in very large word groups, not because they were good headlines, but simply in order to facilitate the look up-process from the index in the printed book, we have a problem of odd 'headlines' of a group of words. This problem cannot be solved without going through all the cases manually.

Finally, we would like to present the boxes not in the numeric order of sections in the thesaurus, but rather from a calculation of semantic relevance, for example by giving prominence to sections where the search word itself is a keyword at level one, or where it has the highest number of direct near-synonyms in the DDB thesaurus. We plan to improve the presentation order as part of the hopefully future project of presenting the entire DDB thesaurus manuscript, including the hierarchies of meaning, as an online thesaurus at the DSL dictionary site [www.ordnet.dk](http://www.ordnet.dk).

## References

### Online dictionaries

*Macmillan English Dictionary Online*. Accessed at: <http://www.macmillandictionary.com> [27/03/2018].

*Merriam-Webster* Accessed at: <http://www.merriam-webster.com> [27/03/2018].

*Oxford English Dictionary (OED)* Accessed at: <http://www.oed.com> [27/03/2018].

*The Historical Thesaurus of English* Accessed via: <http://www.oed.com> [27/03/2018].

*Dictionary.com* Accessed at: <http://http://www.dictionary.com> [27/03/2018].

*Oxford Dictionaries:English* . Accessed at: <https://en.oxforddictionaries.com/english> [27/03/2018].

*The dictionary DDO: Den Danske Ordbog* (the Danish Dictionary) Accessed at: <https://www.ordnet.dk/ddo>.

Dornseiff, Franz (2004). *Der deutsche Wortschatz nach Sachgruppen*, 8. Auflage, Berlin/New York: Walter de Gruyter.

Lyons, John (1977). *Semantics. Volumes 1-2*, Cambridge, University Press.

Nimb, Sanni, Anna Braasch, Sussi Olsen, Bolette Sandford Pedersen, Anders Søgaaard (2017). From Thesaurus to Framenet. In: (Eds. Kosem, I., Tiberius C., Jakubíček, M., Kallas, J., Krek, S., Baisa, V.): *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, Leiden, the Netherlands, 19–21 September 2017*, s. 1-22.

Nimb, Sanni, Lars Trap-Jensen, Henrik Lorentzen (2014). The Danish Thesaurus: Problems and Perspectives. In: Andrea Abel, Chiara Vettori & Natascia Ralli (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: EURAC Research, s. 191-199.

- Nimb, Sanni, Henrik Lorentzen, Liisa Theilgaard, Thomas Troelsgård, Lars Trap-Jensen (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab, Copenhagen.
- Nimb, Sanni, Bolette S. Pedersen, Anna Braasch, Nicolai H. Sørensen & Thomas Troelsgård (2013). Enriching a wordnet from a thesaurus. In: *Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics (NODALIDA)*. Linköping Electronic Conference Proceedings; Volume 85.
- Nimb, Sanni, Bolette Sandford Pedersen (2012). Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri. In: *LREC 2012 Proceedings*. Istanbul, Turkey.
- Pustejovsky, James (1995) *The Generative Lexicon*, MIT Press, Cambridge, Massachusetts.
- Roget, Peter Mark (2002). Roget's Thesaurus, 150th anniversary edition edited by George Davidson 2002. London: Penguin.





# Exploratory and Text Searching Support in the *Dictionary of the Spanish Language*

**Jordi Porta-Zamorano**

*Centro de Estudios de la Real Academia Española*

*E-mail: porta@rae.es*

## Abstract

Online dictionaries try to include search capabilities to meet most users' needs. Although users are not always aware of how to effectively use dictionaries, sometimes it is the interface that does not facilitate a friendly access to the dictionary information. This work aims at lowering the barrier in supporting onomasiological and semasiological advanced searches to the *Diccionario de la lengua española* (DLE) by combining text searches and faceted navigation into a user-friendly dictionary interface, allowing even non-experts to move through the dictionary in a natural and flexible manner. However, since the DLE is an electronic version of a printed dictionary, it contains related and unrelated abbreviations condensing different information that have to be properly converted into the set the facets and values provided by the search system.

**Keywords:** Dictionary interfaces, dictionary searching, online dictionaries

## 1 Introduction

The *Diccionario de la lengua española* (DLE) is the 23<sup>rd</sup> edition of the monolingual general Spanish dictionary produced and published by the *Real Academia Española*<sup>1</sup> (RAE) and the *Asociación de Academias de la Lengua Española*<sup>2</sup> (ASALE). Its basic online edition<sup>3</sup> receives sixty million look-ups per month on average, and provides several search facilities such as lemma and multi-word autocompletion, inflected forms look-up and linguistically-motivated approximate searches implementing an orthophonographic variation model and linking some derivative forms not explicitly registered into the dictionary with its stems. Searches by prefix, suffix, infix, and anagrams are also offered. In addition to these search capabilities, definitions and examples have been lemmatized to provide textual navigation. However, the mentioned search capabilities work in one direction, namely, from words to definitions, supporting only semasiological searches. This paper presents the Advanced DLE, a new interface and search engine to the DLE allowing textual searches and faceted navigation, which can be found within the Enclave RAE Platform<sup>4</sup>.

## 2 Textual Search and Faceted Navigation

Faceted navigation is arguably the most significant innovation in search patterns of the last few decades, which has become nearly ubiquitous in e-commerce (Tunkelang 2009). However, the most common design in dictionary interfaces makes a distinction between basic and advanced searches with a parametric design. In such parametric search interfaces, the user selects all the information in

1 <http://www.rae.es>

2 <http://www.asale.org>

3 <http://dle.rae.es>

4 <https://enclave.rae.es>

one shot, using *a priori* filters, sometimes combining them using Boolean expressions, and reaching a dead end when selecting unsatisfiable combinations of constraints. By contrast, faceted navigation addresses the weaknesses of conventional parametric approaches by offering a progressive query refinement which also allows users to explore and discover new information.

The Advanced DLE interface allows users to start a query by typing one or more terms into a search box and then narrow the search by iteratively selecting facet values or a textual zone of the microstructure where they want to focus their search. Facet values and text zones give also counts, giving users an overview of the distribution of the selected subset of senses. These counts are dynamically updated with every selection. Figure 1 shows a search where *herramienta* (tool) and *madera* (wood) were typed and noun category was selected. Note that there are nine senses matching these criteria, and that search terms appear only within their definition.

Alternatively, users can start searches by selecting facet values without typing any term. The results area of the interface shows the alphabetically ordered list of all the senses matching the selected criteria. Initially, the results list contains all the dictionary senses. Selected facet values or the textual zone can be deselected at any time widening the results list. Figure 2 shows all verbal senses in the field of information technologies coming from French.

Selections are interpreted using the Boolean model, which is not exposed to users. Selecting values from different facets produces an AND between facets and selecting a textual zone produces an AND across facets.

The screenshot shows the Advanced DLE interface with the search box containing 'herramienta madera' and a 'Buscar' button. On the left, under 'Facetas seleccionadas', 'Categoría' is set to 'sustantivo (9)'. Under 'Facetas disponibles', 'Género', 'Lengua', 'Tecnicismo', and 'Tema' are listed with expand/collapse icons. On the right, under 'Usos en textos', 'Definiciones (9)' are listed. A 'Descargar HTML' button is present. The results list shows three entries: 'ahuecador, ra.', 'azuela.', and 'escoplo.', each followed by a numbered list of definitions. The definitions for 'ahuecador, ra.' and 'azuela.' mention 'madera' (wood). The definition for 'escoplo.' mentions 'madera' and 'hierro' (iron).

Figure 1: Nouns containing in the definition the terms *herramienta* (tool) and *madera* (wood).

The screenshot shows the Advanced DLE interface with the search box containing 'informática' and a 'Buscar' button. On the left, under 'Facetas seleccionadas', 'Tecnicismo' is set to 'informática (3)', 'Categoría' is set to 'verbo (3)', and 'Lengua' is set to 'francés (3)'. Under 'Facetas disponibles', 'Tema' and 'Tipo' are listed with expand/collapse icons. On the right, under 'Usos en textos', 'Definiciones (3)' are listed. A 'Descargar HTML' button is present. The results list shows three entries: 'editar.', 'ensamblar.', and 'instalar.', each followed by a numbered list of definitions. The definitions for 'editar.' and 'ensamblar.' mention 'informática' (informatics). The definition for 'instalar.' mentions 'informática' and 'disco duro' (hard disk).

Figure 2: Verbal senses from French in the field of information technologies.

Textual zones can be referred and combined in the search box using a syntax similar to that of Google Advanced Search. Every textual zone is defined as an “advanced operator” and alternative values can be expressed by means of the *O* (OR) operator. Using these operators, one can search words defining plants and fruits, beginning with *al-*, and coming from Arabic, just by typing *definición:planta O fruta O fruto etimología:árabe lema:al\** into the search box.

### 3 From Abbreviations to Facets

Sense facets are an orthogonal set of categories representing the information conveyed mainly by abbreviations and tags of diverse types (etymological, grammatical, geographic, register, semantic, etc.). This information is coded within senses or inherited from its containing entry. There are a total of seventeen facets and 449 values related to senses.

However, because the DLE is an electronic version of a printed dictionary, senses contain related and unrelated abbreviations condensing different information. Unfortunately, abbreviations and facets do not always maintain a straightforward correspondence, and facets extracted from sets of abbreviations must be bundled.

#### **batracio.**

Del lat. cient. *Batrachium*, y este del gr. βατράχειος *batrácheios* 'propio de las ranas', der. de βάτραχος *bátrachos* 'rana'.

1. adj., Zool. anfibio (ll vertebrado). U. m. c. s. m., y era u. en pl. como taxón.

Figure 3: Entry for *batracio* (batrachian).

As an example of these correspondences, a sense like the one in *batracio* (batrachian), shown in Figure 3, contains the following grammatical, domain and usage abbreviations:

- adj. (adjective)
- Zool. (Zoology)
- U. m. c. s. m. y era u. en pl. como taxón (most commonly used as a masculine noun and was used in plural as a taxon)

which generates the following facets bundles:

- {category: adjective, domain: zoology}
- {category: noun, gender: masculine, domain: zoology}
- {category: noun, number: plural, domain: zoology, usage: obsolete}

This bundling scheme avoids retrieving the sense 1 in *batracio* when selecting obsolete adjectives, since this information does not co-occur in the same bundle.

### 4 Indexing Textual Zones

Textual zones correspond to four different structural parts of a sense in the DLE: headword, etymology, definition, and examples. The first two zones are inherited from the containing entry. Selecting a textual zone restrict the results list to senses containing the search terms in that specific zone. Figure 4 shows how many senses have textual zones containing the term *diccionario* (dictionary): there are twenty-one senses having *diccionario* in *Definiciones* (definitions), three in *Ejemplos* (examples), two in *Lemas* (headwords) and two in *Etimologías* (etymologies).



Figure 4: Textual zones containing the term *diccionario* (dictionary).

Headwords such as *actor*<sup>l</sup>, *triz* are converted to *actor* (actor) and *actriz* (actress), and text in definitions and examples are tokenized before indexing. For etymologies, all the abbreviations are expanded and the information about its language family is inserted into the textual index. By doing so, a user can find senses with lemmas coming from the Latin word *ferrum* or demonyms, just by searching place names. Users can also find senses with lemmas from any Indo-European language in its etymology, even if it is not explicitly mentioned, just searching with wildcards the term *indoeurope\** (Indo-European).

## 5 Search Engine Implementation

The search engine has been implemented on top of SWI-Prolog (Wielemaker et al. 2012). Prolog programs describe relations, defined by means of clauses, and computations are initiated by running a query over these relations. Prolog is particularly well-suited for in-memory databases or declarative knowledge-based applications with inference capabilities.

Facets bundles and inverted indices for textual searches are represented in RDF (Resource Description Framework). SWI-Prolog integrates core packages for efficient main-memory RDF storage and querying (Wielemaker et al. 2003). However, faceted search is computationally demanding, and every time a user makes a selection, facet values, zones and counts have to be recomputed. For this reason, some simple queries involving the selection of facet values with the higher number of senses are precomputed and memoized when the server is started up.

A parser that builds a term representing queries using the advanced operators described at the end of Section 2 has been implemented with a simple definite clause grammar. These query terms are evaluated making intensive use of Prolog higher-order predicates. In addition, SWI-Prolog provides packages for indexing XML and gives support to concurrent HTTP requests and JSON. These packages have been used to provide a RESTful web service access to the dictionary (Wielemaker 2014). The complete backend has been implemented in a compact Prolog program of about six hundred lines of code.

## 6 Conclusions and Future Work

Preliminary evaluation from a reduced group of language specialized users revealed a high degree of satisfaction with the new search and exploration possibilities offered in the interface of the Advanced DLE. However, for some queries, users could prefer counts or results to be referenced to entries

instead of senses. The possibility to offer both counts and to group senses could satisfy users with different points of view. Other ways of querying the dictionary, using a form instead of advanced operators, seem to be a natural extension to the interface. Finally, some users have suggested new facets for recreational uses, such as the number of letters in the lemma.

## References

- RAE & ASALE. (2014). *Diccionario de la lengua española*. Espasa, Madrid, 23th ed.
- Tunkelang, D. (2009). *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.
- Wielemaker, J. (2014). SWI-Prolog version 7 extensions. *Proceedings of the Workshop on Implementation of Constraint and Logic Programming Systems and Logic-based Methods in Programming Environments (CICLOPS-WLPE'14)*.
- Wielemaker, J., Schreiber, G. & Wielinga, B. (2003). Prolog-based infrastructure for RDF: performance and scalability. *Proceedings of the 12<sup>th</sup> International Semantic Web Conference (ISWC'03)*, pp. 644-658.
- Wielemaker, J., Schrijvers, T., Triska, M. & Lager, T. (2012). SWI-Prolog. *Theory and Practice of Logic Programming* 12(1-2), pp. 67–96.





# Interactive Visualization of Dialectal Lexis Perspective of Research Using the Example of Georgian Electronic Dialect Atlas

*Marine Beridze<sup>1</sup>, Zakharia Pourtskhvanidze<sup>2</sup>, Lia Bakuradze<sup>1</sup>, David Nadaraia<sup>1</sup>*

*<sup>1</sup>Javakhishvili State University, <sup>2</sup>Goethe-University Frankfurt/M*

*E-mail: marineberidze@yahoo.com, pourtskhvanidze@em.uni-frankfurt.de, lia.bakuradze@tsu.ge, david.nadaraia@gmail.com*

## Abstract

This article presents a report of the results on the current situation in the development of two projects. These are (1) “The Large Georgian Dialect Lexicographic Database and the Georgian Electronic Dialect Atlas” and (2) “A Georgian Language Island in a Trans-Ethnic Area (GLITEA)”<sup>1</sup>. In the first project, the lexicographical component of the Georgian dialect corpus will be expanded and visualized cartographically. The second project examines the dialect of the Georgian – Fereydanian – spoken in Iran by the descendants of about 100 thousand Georgians, who were forcibly evacuated from east Georgia to Iran by Shah Abbas I in the period 1614 to 1616. The dialect is a typical case of a language island and offers the possibility for diverse linguistic research into language history, language contacts and language migration.

**Keywords:** dialectology, Linguistic Geography, dialectological lexicography, canonical visualization of linguistic data.

## 1 Introduction

The eighteen dialects of Georgian, three of which are spoken outside of Georgia, have been the subject of scientific research for about a hundred years. In the course of this period, the dialects were described by empirical and field research methods at the levels of grammar and vocabulary. In the 1930s, several documentations of the dialects were carried out by field research (Beridze V. 1938). As early as 1956, the Chrestomathy of the Georgian Dialects with Dictionaries (Dzidziguri (1956) was founded and published, then the Chrestomathy of the Georgian Dialects (Gigineishvili 1961). These created a whole series of monographs of the grammatical structure of the dialects (e.g. Jorbenadze 1988, 1989; 1991;1995;1998; Dzotsenidze 1974; Nizharadze 1975; Glonti 1975; Gachechiladze 1976; Meskhishvili 1981; Martirosov 1985; Gambashidze 1988; Chincharauli 2005; Tsotsanidze 2012). In recent years, the research groups have also published essays on the morphology of Georgian dialects (Gogolashvili 2017).

In this research tradition the dialectal phenomena were always considered within the framework of an adopted dialect standard and historically developed geographical boundaries. The migrations or the language contact phenomena were not taken into account at all. In the 1980s the basic idea of the Dialectal Atlas of the Georgian was conceived. In this context, a questionnaire was standardized and used in field research. In 2003 the project The Linguistic Portrait of Georgia was started. This project foresees the comprehensive documentation of the linguistic situation in Georgia and continues to this day. The Georgian Dialect Corpus is a result of this project, and forms the modern methodological

<sup>1</sup> Supported by Shota Rustaveli National Science Foundation (SRNSF) [grant numbers 217008/217438].

approach to the study of the vernaculars. The structure of the corpus and the associated lexicographical database contains information on the geographical distribution of the data. Their visualization is the current phase of the project.

## 2 The Specifics of the Geography of Georgian Dialects – Historical and Linguistic Tradition

Georgia is composed of historically formed provinces, distinguished by peculiar cultural, ethnographic and linguistic characteristics (Figure 1, Figure 2.), with historical information about the dialectal diversity of Georgians available in the literature (Jorbenadze 1989; Sardjveladze 1975).



Figure 1. Ethno-culturally defined provinces of Georgia.



Figure 2. Geographical distribution areas of the dialects.

The Georgian dialectology, as a subject, was based on the historically and scientifically developed principle of the ethno-cultural classification of the country. This means that the geographical distribution areas of the dialects coincided precisely with the historically accepted limits of the distribution of ethno-culturally defined provinces of the country

## 3 Dialects in the Context of Migration

The migration processes in Georgia were constantly taking place with varying intensity. These were both ecologically and economically dependent and forced migrations through wars and raids. The traces of migration are still visible in the onomastics of the various localities. There were some great migration waves that have significantly changed the dialectological image.

- Massive expulsion in the 17<sup>th</sup> century from eastern Georgia to Iran near Esfahan by Shah Abbas and the emergence of the Fereydanian.
- Territorial redistribution in border areas to Azerbaijan in the 20<sup>th</sup> century and Turkey in the 19<sup>th</sup> century. Internal migrations for economic, environmental or legal reasons, occurred in the first half of the 20<sup>th</sup> century.
- Internal migrations resulted in compact settlements, but also dispersed branches.

If one considers the idea of the historical boundaries of the ethno-cultural provinces of Georgia in the context of migrations, then the picture shifts. The rigid borders are softened and small dialectal islands are created elsewhere. The dialect islands have their own language development, independent of the “mother dialect”.

## 4 The Status of the Current Dialectal Geography

The Georgian dialects, which through migrations created a new geographical image, developed in Turkey, Azerbaijan and Iran surrounded by completely different cultures, ethnicities and languages. This composition allows the research field to be considered as a large language laboratory by examining the dynamics of dialectal change, as well as language contact phenomena. The same applies to the dialectal islands within Georgia. For example, the Imeretian dialect, which is historically located in western Georgia, is also compactly represented in other areas of Georgia.

1. In Kakhetia (Lagodekhi, east Georgia) (Figure 4). Created by economic migration around 1905.
2. In Samtskhe-Javakheti (south Georgia) (Figure 3). Created by the tight resettlement under the Stalinist repression.
3. In Marneuli (south-east Georgia). Created by the so-called planned resettlement.

On the other hand, if you look at the Kakhetian dialect that is historically and solely located in eastern Georgia, you will find that there are at least five compact dialectal islands in this area: Imeretian, Ratchan, Pshavian, Khevsurian, Tushetian. A similar situation is also seen in Samtskhe-Javakheti. These dialectal “insertions” have different ages and degrees of isolation. Accordingly, they show different dynamics of language development.

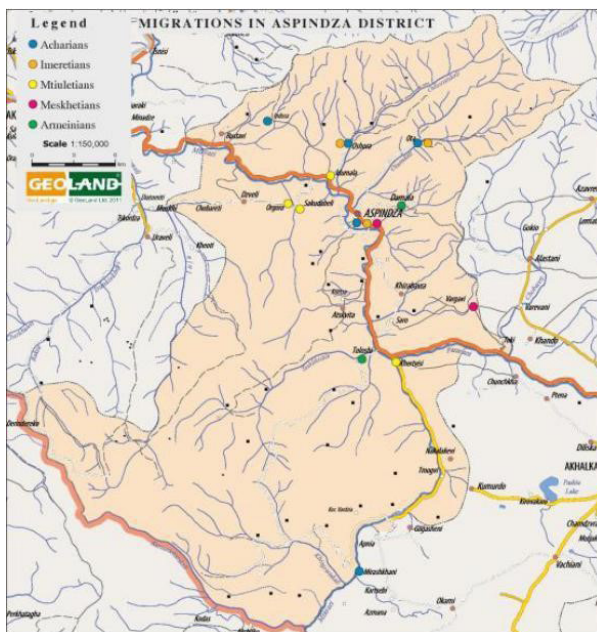


Figure 3. Dialectal islands in Aspinza.



Figure 4. Language islands in Kahetia.

The dialectal distribution of the lexeme ‘child’ *bavšvi* / *bovši* / *balgi* / *boši* illustrates this observation (Figure 5).

According the fixed ethnographic boundaries the dialectal distribution of the lexeme ‘child’ is as follows:

*balgi* / *bavšvi* - Mtiulets-Gudamaqrian, Pshavian, Tushetian, Khevsurian, Mokhevian (*bavšvi* is secondary lexical unit).

- *bavšvi* / *balgi* - Kakhetian, Kartlian (*balgi* is secondary lexical unit due to migrations from mountains).
- *bavšvi* - Javakhian.



- *bovši* - Imeretian.
- *boši* - Lechkhumian, Rachan.
- *bağvi* - Adjarian.
- *bağana* / *bağane* - Gurian.

On a strict search of the dialectal form *bavšvi* in the corpus texts, we will find exclusively Javakhian texts, but if we search the same lexeme based on place of elicitation, then we get at list four dialectal variants of *bavšvi* at the territory of Javakheti (Figure 6): *bağvi* / *balgi* / *bovši* / *boši*.



Figure 5. The dialectal distribution of the lexeme ‘child’.

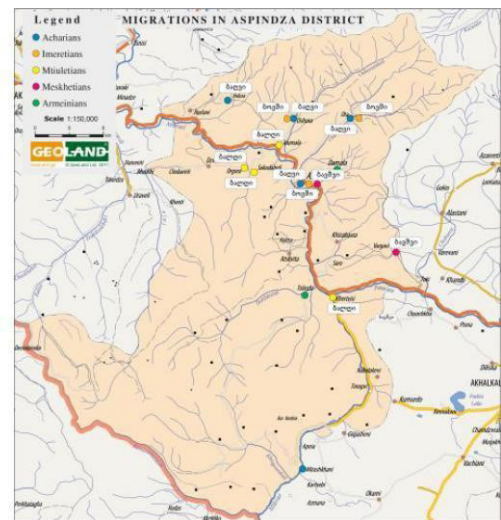


Figure 6. Dialectal variants of the lexeme ‘child’ at the territory of Aspindza (Samckhe-Javakheti)

## 5 Basic Question of the Canonical Visualization of Dialects

With regard to the conception visualization, we operate with the terms “operative imagery” and “recognizable vision” (both terms from Krämer 2009:94-123). For this we included the idea of cognition-winning visualization. Our goal is not simply to describe the migration processes, but to understand the reasons for migration in the geographical context. To achieve this goal, the documented linguistic data is linked to the geographical information of the migration. In the end, it should be visualized not only with the migration of the ethnic groups, but especially the migration of the dialects. The added value of this method lies in the justification of the geo-referentiality for the Georgian dialects. The expectation is to get a new perspective and thus new knowledge about the known language data and processes. On the one hand, it is the simple digital mapping of migration processes wherein the final product in the form of an interactive map has the explanative effect.

## 6 Traditional Solutions and New Tasks

The idea for the Georgian dialect atlas came for the first time from one of the founding fathers of the Georgian dialectology, Varlam Topuria, in the middle of the last century. In the 1980s questionnaires were drawn up and standardized in order to operate uniform field research and so create the language atlases of the individual linguistic regions (Kiziria 1984). The resulting questionnaire included about 1,000 questions on the basic levels of grammar. The work was interrupted in the 1990s and only



continued in 2006. However, the methodological battery of research was extended with digital instruments and the earlier, analog approach was replaced. Large language databases and complex search systems currently determine the documentation of Georgian dialects and their exploration. The use of digital geo-referentiality contours and the creation of dialectal maps offers additional ways of obtained knowledge from this research.

## 7 The Project: Linguistic Portrait of Georgia: Features and Details

As part of the project, the old, analog-documented data were systematized and digitized. Records received on magnetic tapes were transferred and secured in the new data carriers. In addition, new data was also collected and archived in a standardized process. The foundation of the dialect corpus was thus laid. At present, the Georgian Dialect Corpus is characterized by the follow statistical features:

- The lexicographic base was made using documented data in approximately 800 loci.
- The size of the corpus is 2,041,830 tokens.
- The number of the types is 460,631 word forms.
- The number of the texts is 3,356.
- The number of the documented dialects is 18.
- The number of the annotated tokens is 425,631
- The word lists in the testing procedures contain 346,670 word forms.
- The number of all lexicographic articles is 102,638.
- The number of the published lexicographic articles is 54,000.

The dialect corpus and lexicographic base have a particular system of annotation resp. tagging. The common hierarchical structure for the marking of the grammatical and lexical features was constructed. The first stage contains the POS tagging and some parameters (like word fragments, affixes and so on), which are necessary in the tagging process. The second stage marks prepositions, particles of word formation and enclitics. In the last stage the marking system analyzes the semantic features like borrowing, terminology or idiomatic expressions. The annotation tag set is based on the Leipzig Glossing Rules and contains additional specific tags like ‘Fpseudo’ for ‘pseudo standard language’, ‘Fragment’ for a ‘word fragment’ and ‘NonGeo’ for a ‘borrowed word’.

## 8 Dialectography of the Language Islands

The central point of the current phase of the project is the dialectography of the language islands. There is a special scientific interest in phenomena such as linguistic innovation, sub-dialectal forms, pseudo-literary lexemes, borrowings of all kinds, semantic shifts, idiomatic printouts obtained in the use of missing words, and so on. The Georgian Dialect Corpus incorporates the lexicon of the Georgian dialects and Laz in Turkey (corresponding to over nine and six thousand words and articles, respectively), in Iran – Fereydanian (with over six thousand words and articles) and in Azerbaijan – Ingilo (with over 11 thousand words and articles). In field research on the language islands, a little-known phenomenon of the transformation of the language assistants (informants) was observed with regard to independent scientific research on their own dialects. Over time, a special perception of one’s own language isolation is created, and an analytical-structural approach is established. It is thus necessary to later get a naïve linguist to help with the elicitation of the island language. The corpus contains the editing tool “lexicographical editing”, which allows the processing of a lexical registration at different levels (Beridze M. 2017). This is especially important in the description of

language islands, because there are many words initiated by the contact to the standard language of the motherland. A particular difficulty is the description of the internal migrations on the language island, when no certificates exist. An attempt to study the internal differentiation of the Fereydanian dialect was done with the application of dialectometry. The empirical data elicited from the seven Fereydanian villages resulted in a Levenshtein distance matrix (Table 1), which was visualized (plotted) differently.

Table 1. Levenshtein Distance Matrix

	DASH-KASAN	AGCHE	BOIN	NEHZA-D_A	SIBAK	FEREY-DUN_S	MIAN-DASHT
DASHKASAN	0.0	2.142	2.054	2.666	2.405	2.192	2.644
AGCHE	2.142	0.0	2.888	2.567	1.916	1.607	2.533
BOIN	2.054	2.888	0.0	2.964	2.733	3.0	2.666
NEHZAD_A	2.666	2.567	2.964	0.0	1.677	2.3	2.914
SIBAK	2.405	1.916	2.73	1.677	0.0	2.325	2.882
FEREYDUN_S	2.192	1.607	3.0	2.3	2.325	0.0	2.469
MIANDASHT	2.64	2.533	2.666	2.914	2.882	2.469	0.0

The tree diagrams of the language distances were based on the corresponding geographical data (mapped polygons) and the tendency of the internal grouping was designed (Figure 7).

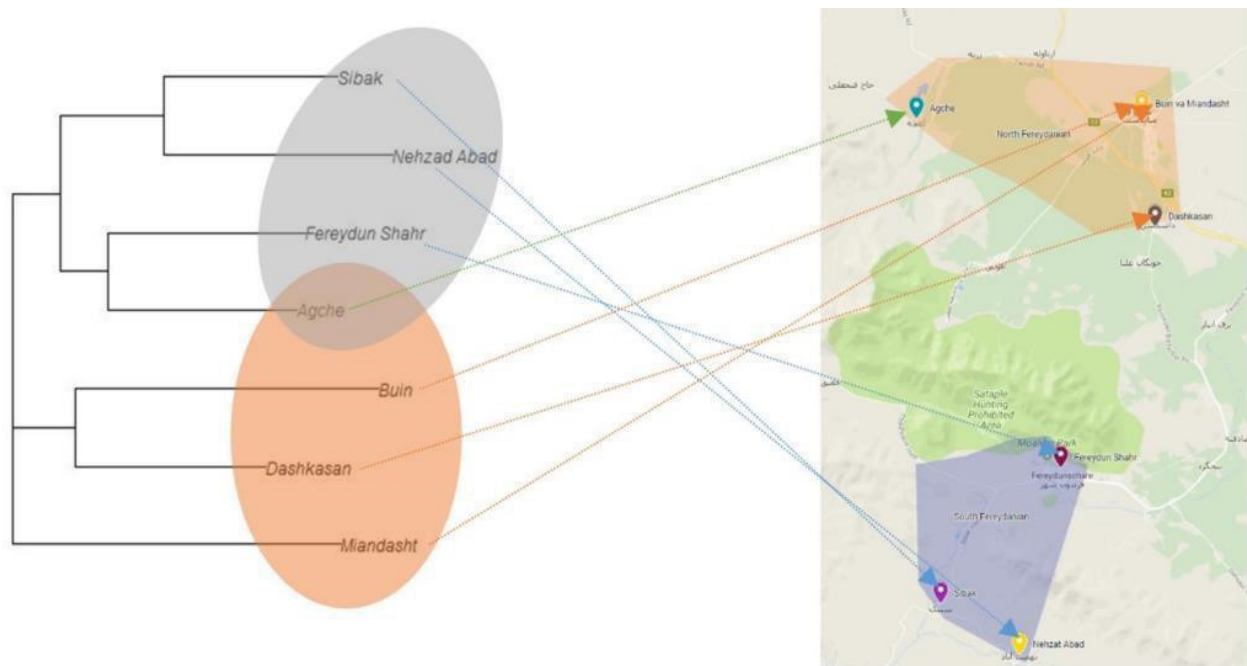


Figure 7. Geographically compact parts of the language islands also show linguistic proximity to each other, whereby a settlement (Agche) occupies the middle position, which has to do with its history of origin by immigrants from the south-Fereydan villages.

In the traditional dialectology in relation to the Fereydan, it was always assumed that the Georgians forced into Iran in the 17<sup>th</sup> century were the speakers of a certain dialect, called Kakhetian, and therefore the current linguistic situation on the language island should prove a common variant. This assumption was based mainly on the historical scientific evidence. The results of the dialectometry

and the different visualizations have clearly shown that the Fereydanian language island is not a monolithic language unit, but is experiencing internal migrations and shifts in language development. However, the empirically justified dialectal diversity of this language island is based on the quantifiable linguistic method of the dialectometry

## 9 Aggregate for Geo-referential Visualization

A central point of the projects is the establishment of the digital dialectal atlas of the Georgian. The atlas is based on the data from the Georgian Dialect Corpus and is designed using an aggregate of geo-referential visualization, and with the use of this in both projects a complex link between the corpus and the possibilities of digital cartography is understood. The text database, the lexicographical database and the geo-referencing algorithm interlock in the way that the geographical distribution of dialectal phenomena is more precisely and spatially represented empirically.

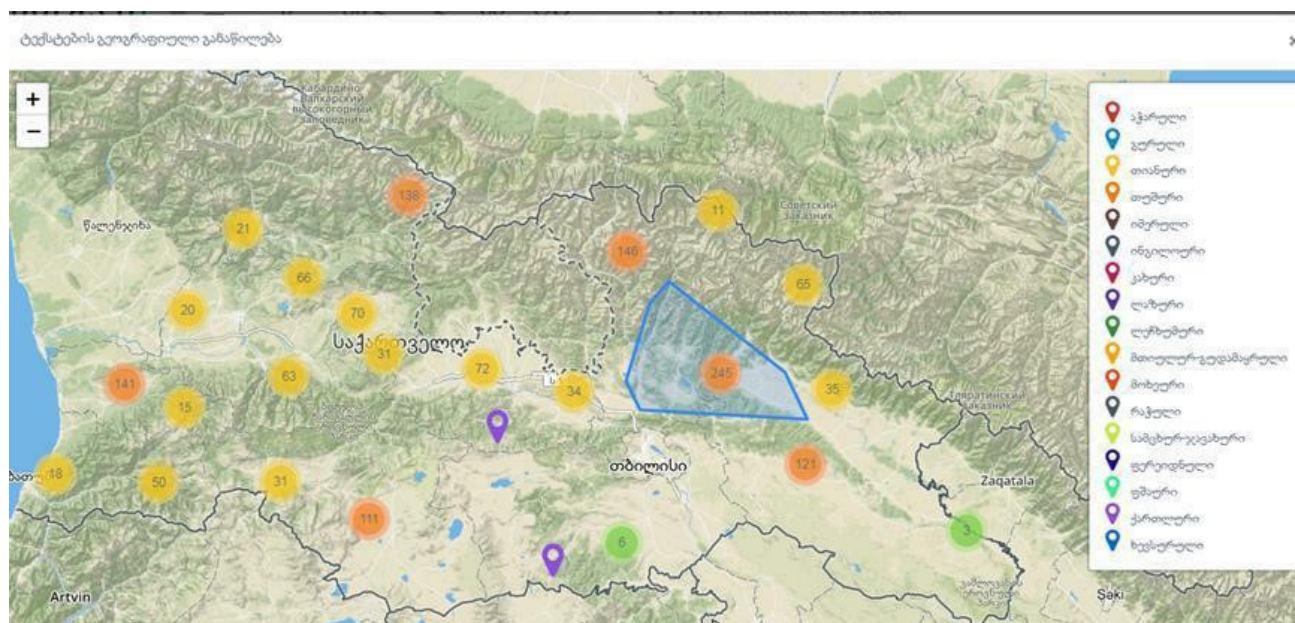


Figure 8. The points with numbers show the number of documented texts and the geographical areas can be mapped as polygons

The focus of the aggregate is the database of digitized dialectal texts, which are provided with detailed metadata. All texts archived in the Georgian Dialect Corpus contain a reference to the geographical location where they were documented. When the text database is tokenized, the meta-information of the entire text is projected onto the individual tokens. The metadata acts as the parent information that is inherited by the child tokens. At the end of this process, each token gets an additional geo-referential write-up that can be visualized. With the automatic lexicon entry of a token, the geo-referential attribution is included and is part of the lexicon article. Thus, the text database guarantees the exact location of tokens from the texts and lexical entries in the lexicographical database. The freely usable geodata for visualization are taken from OpenStreetMap and Geojson. By linking the geo-referential metadata of individual tokens with the information from the OpenStreetMap, a digital atlas is created that visualizes the distribution of dialectal phenomena accordingly. The collections were developed in the lexicographical database, and were created independently of the texts. These collections come from the research tradition of the middle and end of the last century, and are not always uniform in terms of meta-information. In most cases, however, the assessment of the location is the correct one. The entire database is annotated on several levels that are hierarchically structured. At the first level, POS tagging takes place. At the second level, the specification of the grammatical information

includes categorical properties of the analyzed elements. The further level of annotation sets the semantic properties of the tokens. The lexical features are supplemented with additional information and corresponding tags.

## 10 Prospect

The main challenge for the future remains the construction of a sub-corpus, which contains only the language data of the migrated dialects. This means the geo-referential documentation of the Georgian dialects in Turkey, Iran and Azerbaijan, and the corresponding material is currently being prepared. The visualization of the data in the context of the geo-referentiality presents us with the task of constructing a corresponding multi-view-mask, which will be web-based.

## References

- Beridze, M. et al. (2015) Dialect Dictionaries in the Georgian Dialect Corpus, Logic, Language, and Computation/ XIV Springer. Pp. 82–96.
- Beridze, M. et. al. (2017) Georgian Dialect Corpus: Linguistic and Encyclopedic Information in Online Dictionaries. In *Journal of Linguistics/Jazykovedný časopis. The Journal of Ludovít Štúr Institute of Linguistics, SAV. N.68/2*. Pp. 109-121.
- Beridze, M. et al. (2009) The Corpus of Georgian Dialects, In *Proceedings of the NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia. Tribun*. Pp. 25-35.
- Beridze, M. et al. (2015) The Georgian Dialect Corpus: Problems and prospects. In *Proceedings of the conference on Historical Corpora Challenges and Perspectives, Frankfurt*. Pp. 323–333.
- Beridze, M. et al. (2011) Dictionary as a textual component of Corpus (Georgian Dialect Corpus). In *Proceedings of the conference on corpus linguistics, St. Petersburg*. Pp. 92-97.
- Beridze, M. et al. (2014) Lexicographical concept of Georgian Dialect Corpus and problems of morphological analysis.). In *Proceedings of the conference on Applied Linguistics in Science and Education, Knijnii dom, St. Peterburg*. Pp. 91-94.
- Beridze, M. et al. (2016) Lexicographic Potential of the Georgian Dialect Corpus. In *Proceedings of the XVII EURALEX International Congress, Lexicography and Linguistic Diversity*. Pp. 300-309.
- Beridze, V. (1938) Vocabulary of the Kartvelian Languages, I, 164 p.
- Chincharauli, Al. (2005) *Khevsurian Dictionary*, 1177 p.
- Dzidziguri, S. (1956) *Chrestomty of the Georgian Dialects with Dictionaries*. Tbilisi, 401 p.
- Dzotsenidze, K. (1974) *Upper Imeretian Dictionary*, 645 p.
- Explanatory dictionary of the Georgian language (1950-1964) 8 vols. Tbilisi: Georgian Academy of Sciences.
- Gachechiladze, P. (1976) *Lexical material of the Imeretian dialect*, 182 p.
- Gambashidze, R. (1988) *Dictionary of Ingiloan dialect of Georgian Language*, 629 p.
- Gigineishvili, I. et al. (1961) *Georgian dialectology*, I, 732 p.
- Glonti, Al. (1975) *Dictionary of Georgian dialects*, 411 p.
- Gogolashvili. G. et al. (2016) *Morphology of the Contemporary Georgian Language: Dialects*, II, 916 p.
- Jorbenadze, B. (1989) *Georgian Dialectology*, I, 636 p.
- Jorbenadze, B. (1998) *Georgian Dialectology*, II, 675 p.
- Jorbenadze, B. (1991) *The Kartvelian Languages and Dialects*, 272 p.
- Jorbenadze, B. (1995) *Dialects of the Kartvelian Languages*, 448 p.
- Kiziria, A. et.al. (1984) *Questionnaire for Dialect Atlas Material*, 80 p.
- Krämer, S. (2009) Operative Bildlichkeit. Von der Grammatologie zu einer “Diagrammatologi? Reflexion über erkennendes Sehen. In *Martina Heßler and Dieter Mersch (Eds.), Logik des Bildlichen. Zu Kritik der ikonischen Vernunft, Bielefeld: transcript, 2009*. Pp. 94-123.



- Martirosov, A. (1985) The Main Issues of the Study of Georgian Dialect Vocabulary and Compilation of Dictionaries, In *Ibero-Caucasian linguistics, XXIII*. Pp. 139-148.
- Meskhishvili, M. et al. (1981) *Dictionary of Kartlian Dialect*, 551 p.
- Nizharadze, S. (1975) *Adjarian dialect*, 231 p.
- Sarjvelasze, Z. (1975) *The issues of Georgian literary language history*, 271 p.
- Speelman, D. and D. Geeraerts. (2008) 'The role of concept characteristics in lexical dialectometry', In *International Journal of Humanities and Arts Computing 2 (1-2)*. Pp. 221-42.
- Szmrecsanyi, B. (2008) 'Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects', In *International Journal of Humanities and Arts Computing 2 (1-2)*. Pp. 279-96.
- Szmrecsanyi, B. (2010) *The Morphosyntax of BrE Dialects in a Corpus-based Dialectometrical Perspective: Feature Extraction, Coding Protocols, Projections to Geography, Summary Statistics*. Freiburg: University of Freiburg. URN: urn:nbn:de:bsz:25-opus-73209. Available online at: <http://www.freidok.uni-freiburg.de/Volltexte/7320/>
- Tsotsanidze, G. (2012) *Dictionary of Tushian Dialect*, 319 p.
- Trudgill, P. (1974) 'Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography', In *Language in Society 3 (2)*. Pp. 215-46.
- Viereck, W., H. Ramisch, H. Händler, P. Hoffmann and W. Putschke (1991) *The Computer Developed Linguistic Atlas of England*. Tübingen: Niemeyer.
- Georgian National Corpus      <http://gnc.gov.ge>
- Georgian Dialect Corpus      <http://corpora.co>





# The Dictionary of the Serbian Academy: from the Text to the Lexical Database

**Ranka Stanković<sup>1</sup>, Rada Stijović<sup>2</sup>, Duško Vitas<sup>1</sup>, Cvetana Krstev<sup>1</sup>, Olga Sabo<sup>2</sup>**

<sup>1</sup>University of Belgrade, <sup>2</sup>Institute for Serbian Language, Serbian Academy of Sciences and Arts

E-mail: ranka.stankovic@rgf.bg.ac.rs, rada.stijovic@isj.sanu.ac.rs, vitas@matf.bg.ac.rs, cvetana@matf.bg.ac.rs, olga011@yahoo.com

## Abstract

In this paper we discuss the project of digitization of the *Dictionary of the Serbo-Croatian Standard and Vernacular Language*. Scanning and character recognition were a particular challenge, since various non-standard character set encoding was used in the course of the almost 60-year long production of the dictionary. The first aim of the project was to formalize the micro-structure of the dictionary articles in order to parse the digitized text of and transform it into structured data stored in relational lexical database. This approach is compatible with several standard structured forms and ontologies (TEI, LMF, Ontolex, LexInfo). A lexical database model was designed in compliance with these structured forms, following mostly the *lemon* model. Mapping of the lexical entry markers to LexInfo and TEI enabled export of the lexical data to the mentioned formats. A software solution for the dictionary text analysis, parsing and lexical database population was developed and tested on the first and the last published volumes of the dictionary (which contain 27,141 articles in total). An evaluation of the results shows that the developed model and software solution can be successfully used for the other volumes as well.

**Keywords:** computer lexicography, lexical database, language resources, dictionary, Serbian language

## 1 Introduction

The first volume of the *Dictionary of the Serbo-Croatian Standard and Vernacular Language* (referred to as the *Dictionary of Serbian Academy* or *DSA*), prepared and compiled by the Institute for the Serbian Language of the Serbian Academy of Sciences and Arts, was published in 1959 in paper form. Out of 35 planned volumes, 19 volumes have been published with the 20<sup>th</sup> volume to be released soon. The material used for the dictionary covers written resources of the standard Serbo-Croatian language from the beginning of the 19<sup>th</sup> century to the present day, as well as about 300 word collections (provincial expressions, dialectical variations, etc.) of all Shtokavian dialects. The paper version of the dictionary has a complex microstructure that was designed at the time of the release of the first volume. Later, it was supplemented and described in a handbook.<sup>1</sup> The text itself is in basic Cyrillic alphabet, but in addition to this it contains accented vocals and various Church Slavonic, Latin and Greek characters.

It was first suggested that the production of the dictionary should be modernized many years ago (Sabo & Vitas 1989). However, only in recent years were these ideas revitalized, and various possibilities of updating the work on this vocabulary have since been considered (Vitas & Krstev, 2015; Ivanović et al. 2016). The digitization (which is also the topic of the present paper) of the published volumes and raw materials (lexicographic leaflets) began in 2016, and the first use of the two volumes that were

1 Упутство за обраду Речника, Београд: Институт за српск(охрватск)и језик САНУ (рукопис), 1959. и (допуњено) 2017 [*A Handbook for Dictionary Processing*, Belgrade: Institute for Serbo(-Croatian) language SASA (manuscript), 1959 and (supplement) 2017].

digitized and published was reported in Stijović et al. (2017), while the application for management of lexicographic leaflets is described in Stijović (2018). Out of 19 volumes, two were available as MS Word files, two as PDF files, and the others only in paper form. Unfortunately, neither MS Word nor PDF files could be used without further preprocessing, since non-Unicode character sets were used with non-standard accents and other forms of character encoding. All three available formats needed additional transformation: conversion, transliteration (where the Latin alphabet was used instead of Cyrillic) and extensive manual postediting. After the implementation of OCR, dictionary articles were corrected and formatted (bold, italic) manually to obtain the identical layout and formatting as in the printed version. The final correction of all 19 digitized volumes is nearly finished.

In this paper we will present the work that has been done so far in transforming the digitized text of the *Dictionary* into various standard structured formats and into a lexical database with the first aim to speed up the linear production process of the dictionary. This work makes it possible to use the lexical base of the dictionary for research purposes and for the production of various derived lexicographic products.

## 2 Related Work

Digital dictionaries ceased to be a novelty a long time ago. The majority of new dictionaries are produced (and in some cases exist only) in digital form. However, many significant lexicographic works that were produced in the past now need to be transformed into this format. Ever since some initial retro-digitization projects, such as the transformation of the *Oxford English Dictionary* (Berg et al., 1988), the transformation from an unstructured to a structured text was recognized as the main task of such endeavors. To do this, the text of a dictionary has to be parsed and the structure of the articles has to be formalized. For the representation of this formal structure markup languages are used, preferably standard ones that support interchange and merging (Lemnitzer et al., 2009). At this point, retro-digitized and digital-born dictionaries meet, since both types should preferably use the same or compatible formal structure and markup language.<sup>2</sup> This development led to further linking of lexical data and their integration with semantic resources, such as ontologies (McCrae et al., 2011).

The *DSA* is rather special compared to similar dictionaries for other languages: its significant part has already been compiled and published, but still needs to be digitized; on the other hand, the dictionary is not finished, and a lot of work remains to be done. These two tasks need to be synchronized in order to obtain a homogenous work as the final product. Moreover, although the digitization of the *Dictionary* is currently lagging, we would like to catch up on the lost time by using up-to-date technology.

## 3 Model of Lexical Database

### 3.1 Formalization of the structure of dictionary articles

The first phase of the conversion of the *DSA* from the text form (unstructured text) into the lexical base (structured text) consisted of a thorough analysis of formatting conventions that were used for typesetting dictionary entries as well as of the identification of triggers (such as special words, abbreviations or punctuation marks) used to introduce specific information. Conventions and triggers were used to identify basic information presented in the dictionary. This analysis enabled us to recognize the following entry structure:

<sup>2</sup> For instance, Ahačič (2015) presents a Slovenian Dictionary Portal that collects information from 22 dictionaries, dating from the 16<sup>th</sup> century to the present day. Some of these dictionaries were transformed into XML format, while other were developed in it.

- 1) headword group;
  1. headword lemma;
  2. grammatical data;
  3. related words (lexical entries);
- 2) grammatical data;
- 3) etymology – an element from the closed set of abbreviations for the names of languages is used to introduce the information on etymology, for instance *грч.* for *грчки* ‘Greek’ or *фр.* for *француски* ‘French’;
- 4) sense – the individual meanings of a headword are marked with Arabic numerals or, if the meanings are close, with lowercase letters. If the headword is a verb, its non-reflexive and reflexive forms are marked by Roman numerals I and II.
  1. terminological markers – predefined terminological abbreviations are used to indicate the scientific domain in which a particular sense of a lemma is used (for instance, *геол.* for *геологија* ‘geology’ or *фил.* for *филозофија* ‘philosophy’);
  2. the linguistic and stylistic tags – qualifiers of predefined linguistic and stylistic values (for instance, *покр.* for *покрајински* ‘provincial’, *арх.* for *архаичан* ‘archaic’, *неј.* for *нежоративан* ‘pejorative’);
  3. related words (lexical entries) – references to other lexical entries (for instance, preferred forms) are introduced after appropriate abbreviations: *исп.* for *испореди* ‘compare’;
  4. definitions are descriptive or referential (*в.* for *види* ‘see’), in rare cases synonyms;
  5. definitions are supplemented by the lists of:
    1. synonyms (after abbreviation *син.* for *синоним* ‘synonym’);
    2. antonyms (after abbreviation *супр.* for *супротан* ‘antonym’);
    3. related words;
  6. examples:
    1. the text of the example
    2. the bibliographic reference (in parenthesis);
- 5) multiword expressions (syntagmatic and phraseological – they are listed in the separate paragraph beginning with the abbreviation *Изр.* for *Израз* ‘phrase’; and
- 6) proverbs – they are also listed in the separate paragraph, beginning with the abbreviation *НПосл.* for *Народна пословица* ‘vernacular proverb’.

Some elements of this structure may appear at different positions and levels, such as ‘related words’. Grammatical data can contain various information, depending on the lemma’s part-of-speech: some may refer to the lemma, others to the headword. Both high- and low-level elements are optional and repeatable, except for the headword group itself. The typographic conventions and triggers as applied to nouns are summarized in a simplified way in Table 1, while the graphical outline of the basic structure of an article in the *Dictionary* is presented in Figure 1. The same entry has been taken as an example in Table 1 and Figure 2, which illustrates the result of the parsing process.

### 3.2 Dictionary markers

Beside various semantic, accentual and grammatical (phonetic, morphological and, more recently, syntactic) information, the *DSA* also includes indications of the normative, functional, stylistic and socio-historical status of the lexical entries, as well as their spatial and temporal scope and domain of use. Apart from other lexicographical and technical procedures, various markers are used to denote the status of the lexemes and to detail the rules of their use. They are placed at different positions in the articles of the *Dictionary* following strict rules. A total of 371 such markers, in abbreviated forms, is used, and they are all listed at the beginning of each printed volume. For the purpose of the

digitization, a meticulous systematization of all such markers, in terms of the information they convey and their position in the articles of the *Dictionary*, was performed.

A feature structure, as a general-purpose data structure which identifies and groups together individual features, each of which associates a name with one or more values, was mapped with the help of the aforementioned abbreviations used in the dictionary. The existing 371 abbreviations (markers) were mapped as data category values with 30 data categories, and further grouped in data-category sets. Feature structures represent the interrelations among various pieces of information and provide a metalanguage for the generic representation of the analyses and interpretations.

Table 1: The typographic conventions and triggers as applied to nouns.

Element		Example	Typography	Trigger begin	Trigger end
Headword group					
	lemma	палеоџен	<nl> bold		comma or trigger begin
	gramm. data	-a		hyphen	
	lemma	палеоџен	<nl> bold	и	comma or trigger begin
	gramm. data	-ена		hyphen	
gramm. data		м		item in a list	
Etymology		palaiós kainós		Open parenthesis + item in a list, e.g. „грч.“	closing parenthesis
Sense				1, 2, 3 or а, б, в or I, II or trigger begin	
	terminological markers	геол.		item in a list	trigger begin
	linguistic/markers	/		item in a list	trigger begin
	related words	/		Some punctuation marks; item in a list	trigger begin
	definition	прва, најстарија епоха палеогена.	italic		trigger begin
	synonyms, antonyms, related	/		item in a list	
Example	example text	Формације геолошке се даље дијеле...		dash	
	Bibliographic references	Д-П1, 17		Open parentheses	Closing parenthesis
MWE		/		Изр.	
Proverbs.		/		НПосл.	



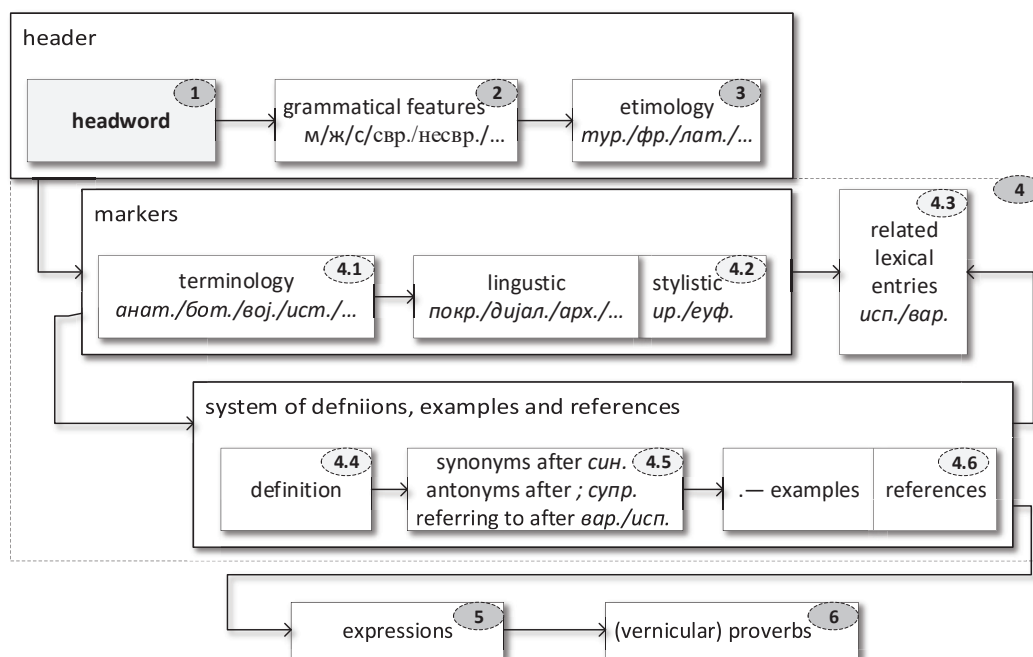


Figure 1: The microstructure of dictionary articles.

#### 4 The transformation from the dictionary article text form to the lexical database

The guidelines for dictionary writing were used to define the rules for the segmentation of the dictionary articles, the pattern recognition, and the alignment of the recognized markers with the predefined categories, as described in the previous section. The dictionary article units that were recognized were marked with XML tags, in accordance with the predefined scheme and imported into the relational database. The model of the lexical database was inspired by LMF (Calzolari et al. 2013), TEI<sup>3</sup> and *lemon* (McCrae et al. 2011). The current trends seem to be consistent with the idea of an ecosystem, where different standards can coexist and mutually enrich each other (). We have experimented with partial transformations of our XML documents to these models in order to find the most suitable solution.

MS Word documents formatted identically as the print versions were input to the automatic segmentation procedure. One thus prepared dictionary article is presented on the left side of Figure 2, while the result of the segmentation of the same dictionary article is presented on the right. One of the additional functions of the developed software is the consistency check that reports any occurrence that does not comply with the syntactic rules for the predefined article structure. This information is then used for the manual postediting of the digitized text.

Within this research, the partial alignment of the XML tag set, defined for the DSA with the TEI dictionary module, was made. For instance, the `<gen>` element is used to denote the grammatical gender, while `<usg type = "dom">` refers to the terminological field, and `<def>` to the definition. Citations are tagged with `<cit>`, and bibliographic references with `<bibl>`; in the content of these elements some additional phrases are tagged, such as the names of locations (using the `<placeName>` tag) or authors (using the `<author>` tag). Similarly, comparison and partial alignment of the DSA tag

3 <http://www.tei-c.org/>

set was done with Ontolex<sup>4</sup> and LexInfo<sup>5</sup>, but a more precise and detailed alignment is envisaged. The dictionary article from Figure 2 is represented in the TEI compliant form in Figure 3.

<p><b>пӑлеоцӑн</b>, -а и <b>палеоцӑн</b>, -ена м (грч. palaiós kainós) геол. <i>прва, најстарија епоха палеогена</i>. — Формације [геолошке] се даље дијеле на ... епохе. Тако се ... терцијар [састоји] од пет: палеоцена, еоцена, олигоцена, миоцена и плиоцена (Д–П 1, 17). (Калм. Р. 1, 81; Р. МС).</p>	<b>пӑлеоцӑн</b> -а <b>палеоцӑн</b> -ена м				
	грч.	palaiós kainós	геол.		
	<i>прва, најстарија епоха палеогена.</i>				
	Формације [геолошке] се даље дијеле на ... епохе. Тако се ... терцијар [састоји] од пет: палеоцена, еоцена, олигоцена, миоцена и плиоцена				
		Д–П 1, 17	Калм. Р. 1, 81; Р. МС		

Figure 2: An example of the automatic dictionary article segmentation.

The information that was registered in the dictionary explicitly, such as the domain, the stylistic use, the etymology, and so on, was mapped with the help of the data categories and their values; however, some grammatical information was not explicitly encoded. For example, part-of-speech (POS) is rarely encoded explicitly, but rather through an indirect indicator. For instance, the gender mark (*м* masculine, *ж* feminine, *с* neuter) indicates the grammatical category of nouns, the aspect type mark (*свр.* for *свршен* ‘perfective’, *несвр.* for *несвршен* ‘imperfective’) is the indicator of verbs, while adjectives are given in all three grammatical genders in the nominative singular (e.g. *активан*, *-вна*, *-вно* ‘active’). The set of rules was produced that calculates the POS where it is not explicitly mentioned, and their application yielded correct POS information for more than 95% of all lexical entries.

```
<entry n="3971">
  <form type="lemma"><orth> пӑлеоцӑн </orth></form>
  <form type="inflected">-а</form>
  <form type="lemma"><orth> палеоцӑн </orth></form>
  <form type="inflected">-а</form>
  <gramGrp><gen>м</gen></gramGrp>
  <sense><etym><lang>грч.</lang> palaiós kainós </etym><usg type="dom"> геол.</usg>
    <def> прва, најстарија епоха палеогена.</def>
    <cit> Формације [геолошке] се даље дијеле на ... епохе. Тако се ... терцијар [састоји] од пет: палеоцена,
еоцена, олигоцена, миоцена и плиоцена <bibl>( Д–П 1, 17).</bibl>( Калм. Р. 1, 81; Р. МС) </cit>
  </sense>
</entry>.
```

Figure 3: The example of a dictionary article using TEI compliant tagging.

Our main goal is to produce a central lexical database that will enable multiuser management of the lexical data and provide access to the content of the volumes that were already published. For the development of the lexical database model for the *DSA*, a similar approach was used as in Stanković et al. (2018) for the Serbian morphological electronic dictionary. The main class, in the core of this dictionary model, is *LexicalEntry*, representing a headword of the dictionary article, which encompasses the set of senses that are associated with this headword. The *LexicalRelation* class relates lexical variants (for instance, *пӑлеоцӑн* and *палеоцӑн* ‘paleocen’ in our example), full forms and their abbreviations (*дп* for *доктор* ‘doctor’), orthographic variants and different pronunciations (Ekavian *sneg* and Ijekavian *snijeg* ‘snow’). *LexicalSense* is used to represent a particular sense of a lexical entry, and to

4 [https://www.w3.org/community/ontolex/wiki/Final\\_Model\\_Specification](https://www.w3.org/community/ontolex/wiki/Final_Model_Specification)

5 <http://www.lexinfo.net/ontology/2.0/lexinfo>

link a lexical entry with a set of senses. Each sense can be related to its own set of markers (outlined in Section 3.2) through the *SenseProperties*. These markers are controlled by the internal thesaurus of data categories, as outlined in Stijović and Stanković (2017).

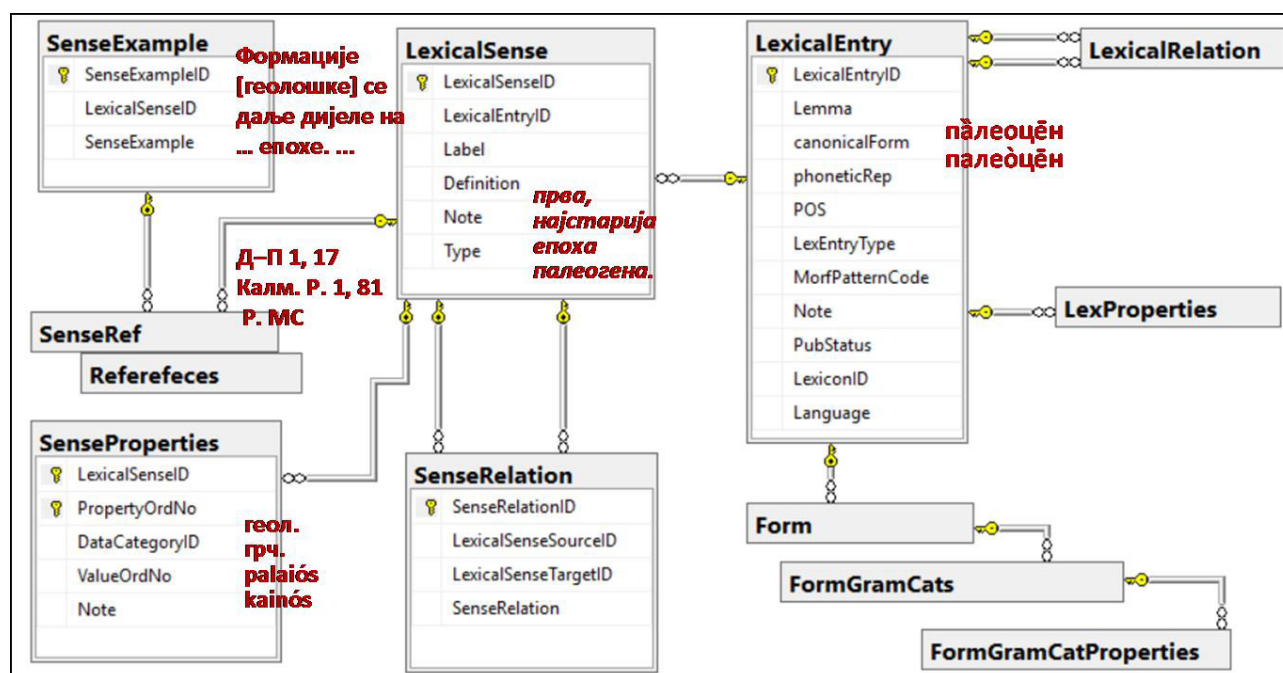


Figure 4: Database model with mapped dictionary article segments.

For languages with complex morphology, such as Serbian, information about inflection is very important. In our model we used the class *Forms* to store the information related to the inflected forms, while the grammatical categories are assigned to the *LexProperties*. At this moment, only information found in the *DSA* is stored in the table *Forms* – sometimes this information represents a selection of inflected forms, sometimes only inflectional endings. The *SenseRelation* is used to connect various senses of lexical entries, while the *SenseRef* and *SenseExample* contain information about provenance and use. The class *References* contains metadata about the bibliographic information from the dictionary corpus. The set of markers is partially aligned with the TEI elements (and attributes) and *LexInfo* in order to relate the lexical data to other resources and provide automatic production of the dictionary in different forms and formats. Figure 3 illustrates a part of the database model with a few examples of structured data.

## 5 Results and discussion

The procedure that was presented in this paper was applied to the digitization of the 1<sup>st</sup> (1959) and 19<sup>th</sup> (2014) volumes of the *DSA*. As a result, the structured documents were obtained in standardized formats and they were subsequently loaded into the lexical database. The automatic procedure recognized, structured, annotated and stored in the lexical database 15,988 dictionary articles from the 1<sup>st</sup> volume and 11,153 from the 19<sup>th</sup>.

Two processed volumes were compared regarding the POS of headwords, and the absolute and relative frequencies were compared: 10,633 lexical entries were recognized as nouns in the 1<sup>st</sup> volume (66.2% of the total number of entries) and 7,808 (69.7%) in the 19<sup>th</sup>, 1,364 (8.5%) entries were recognized as verbs in the 1<sup>st</sup> volume and 1,192 (10.6%) in the 19<sup>th</sup>, while 2,654 (16.5%) entries were

recognized as adjectives in the 1<sup>st</sup> volume and 1,391 (12,4%) in the 19<sup>th</sup> volume. A small number of entries was not labelled with POS: 607 (3.8%) and 521 (4.7%) in 1<sup>st</sup> and 19<sup>th</sup> volumes, respectively, due to the lack of information in the *Dictionary* itself or the lack of appropriate rules; the improvement of the set of rules for POS detection is underway (Stijović & Stanković, 2017). The 1<sup>st</sup> volume has more dictionary articles than the 19<sup>th</sup> (15,988 vs. 11,153), but at the same time less tokens (1,722,483 vs. 1,987,504) and formal words (551,030 vs. 672,004). The 1<sup>st</sup> volume refers to 2,105 different sources with 7,127 examples, while the 19<sup>th</sup> refers to 6,037 different sources with 28,725 examples.<sup>6</sup> Some other comparative differences between 1<sup>st</sup> and 19<sup>th</sup> volume are presented in Figure 5.

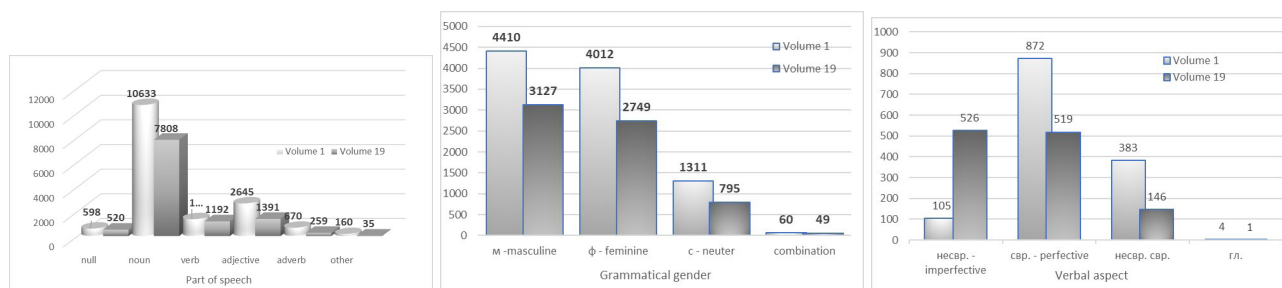


Figure 5: The comparison of two volumes: by part-of-Speech (left), nouns by grammatical gender (middle), and verbs by aspect (right) by lexical entries types from data in lexical database.

## 6 Conclusion

In this paper we discussed the challenges and results of the digitization of the *DSA*. The results presented in this paper are restricted to the processing of the 1<sup>st</sup> and 19<sup>th</sup> volumes, but the digitization of all previously published volumes is in progress. So far, the *Dictionary* was produced linearly. In the future, however, the linear processing of entries may be abandoned, as this could accelerate the production process. The supplements to the already published volumes could also be produced. Once the *DSA* is fully populated, users with different levels of accessibility will be able to search through its lexical database. It is also envisaged for the basic data will be open to the general public.

## References

- Ahačič, K., Ledinek, N., & Perdih, A. (2015). Fran: The Next Generation Slovenian Dictionary Portal. In *Natural Language Processing, Corpus Linguistics, Lexicography. Eight International Conference Bratislava, Slovakia*, pp. 21-22.
- Berg, D. L., Gonnet, G. H., & Tompa, F. W. (1988). *The New Oxford English Dictionary Project at the University of Waterloo* (pp. 2-7). UW Centre for the New Oxford English Dictionary.
- Calzolari, N., Monachini, M., Soria, C. (2013). LMF – Historical Context and Perspectives, in: LMF Lexical Markup Framework, Eds: G. Francopoulo, P. Paroubek, John Wiley & Sons, Inc.
- Ivanović, N., Jakić, M., Ristić, S. (2016). Građa Rečnika SANU – potrebe i mogućnosti digitalizacije u svetlu savremenih pristupa, u: S. Ristić i dr. (ed.), *Leksikologija i leksikografija u svetlu savremenih pristupa*, Beograd: Institut za srpski jezik SANU, pp. 133–154. [The material of the Dictionary of the SANU - the needs and possibilities of digitization in the light of contemporary approaches (in Cyrillic)].
- Lemnitzer, L., Romary, L., & Witt, A. (2009). Representing human and machine dictionaries in Markup languages. In *arXiv preprint arXiv:0912.2881*.

<sup>6</sup> The 19<sup>th</sup> volume contains fewer articles but they have a more complex semantic structure, which accounts for more meanings, and consequently more examples.

- McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with *lemon*. In *Extended Semantic Web Conference* Springer, Berlin, Heidelberg, pp. 245-259.
- Monachini, M. & Khan, A. F. (2018). Towards the Construction of a Lexical Data and Technology Ecosystem: The Experience of ILC-CNR. In *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, Eds: Ilan Kernerman, Simon Krek, Miyazaki, Japan, pp. 52-54.
- Sabo, O., Vitas, D. (1998). Mogućnost osavremenjivanja izrade rečnika na primeru Rečnika srpskohrvatskog književnog i narodnog jezika SANU i Instituta za srpskohrvatski jezik. In *IV međunarodni naučni skup „Računarska obrada jezičkih podataka”*, Portorož: Institut Jožef Stefan, pp. 375–384. [Possibility for modernizing the development of the dictionary on the example of the Dictionary of the Serbo-Croatian literary and vernacular language SASA and the Institute for Serbo-Croatian]
- Stanković, R., Krstev, C., Lazić, B., Škorić, M. (2018) Electronic Dictionaries – from File System to *lemon* Based Lexical Database, In *6<sup>th</sup> Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science* (in print)
- Stijović, R., Sabo, O., Stanković, R. (2017). Rečnik SANU kao baza terminoloških rečnika (na primeru Rečnika kulinarstva) u: Piper, P., Jovanović, V. (eds.), *Slovenska terminologija danas*, Beograd: Srpska akademija nauka i umetnosti, 2017, pp. 229–241. [The dictionary of the SASA as a basis of terminology dictionaries (on the example of the Culinary Dictionary) (in Cyrillic)]
- Stijović, R. (2018). Građa Rečnika SANU – blago koje treba sačuvati (o digitalizaciji listića), In *Naš jezik XLVI-II/3–4, Beograd: Institut za srpski jezik SANU*, pp. 201–207. [The structure of the Dictionary of the SANU - the goods to be preserved (on the digitization of the leaflets) (in Cyrillic)]
- Stijović, R., Stanković, R. (2017) Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU, In *47. Naučni sastanak slavista u Vukove dane, Beograd* (Međunarodni slavistički centar. Filološki fakultet) (in print). [Digital edition of the SASA Dictionary: a formal description of the microstructure of the SASA Dictionary (in Cyrillic)]
- Vitas D., Krstev C. (2015) Nacrt za informatizovani rečnik srpskog jezika, In *Naučni sastanak slavista u Vukove dane - Srpski jezik i njegovi resursi: teorija, opis i primene, Vol. 44/3*, Međunarodni slavistički centar, Beograd, pp. 105-116. [Blueprint for the computerized dictionary of the Serbian language (in Cyrillic)]

## Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003, #178003 and #178009.





# SOFTWARE DEMONSTRATIONS



# An Overview of FieldWorks and Related Programs for Collaborative Lexicography and Publishing Online or as a Mobile App

**David Baines**

*SIL International*

*E-mail: david\_baines@sil.org*

## Abstract

The FieldWorks ecosystem provides open-source tools for linguists whether working alone or in distributed teams.

FieldWorks is a comprehensive tool for managing linguistic data. It has an extensive selection of fields for each lexical entry and areas for storing grammatical data and interlinear texts. The bulk editing tools can save hours of work by operating on many entries at once. FieldWorks can be used to create mono- or multi-lingual dictionaries and has excellent support for complex scripts. Comprehensive help and resources are available within the tool, which is designed for trained linguists.

Language Forge is an online dictionary creation tool that allows collaborators to browse, comment or contribute to a lexicon. The project manager can control the roles for each team member. Language Forge shares the FieldWorks data allowing users of either tool to modify a shared lexicon. Language Forge can be used with minimal training as it exposes only a small subset of the FieldWorks data.

Webonary is an online platform for publishing dictionaries and their reversal index. The linguist can update the data on Webonary from within FieldWorks as often as desired. Dictionary App Builder facilitates the creation of Android and iOS apps from the FieldWorks data.

**Keywords:** FieldWorks, Language Forge, collaboration, multilingual, complex scripts, Online publishing, Webonary, mobile publishing, Dictionary App Builder

## 1 Introduction

In this paper I will briefly discuss the capabilities of FieldWorks (FLE<sub>x</sub>) with a particular emphasis on possibilities it provides for collaboration, both with other linguists and with other software. There are two broad categories of software that work with FLE<sub>x</sub>; software that enables collaboration on the same data and software that prepares the data for publication in various media. Key limitations of FLE<sub>x</sub> will also be considered. The aim of this paper is to enable the reader to know whether (or not) FieldWorks and its ecosystem will be of use to them. Recommendations are given about how to investigate further should this paper be insufficient to determine the suitability of FLE<sub>x</sub> for a particular project.

## 2 FieldWorks

### 2.1 Introduction and a Little History

FieldWorks was once a suite of programs developed by SIL International as a repository for the academic data that a field-based linguist would want to store about a single language and culture.

Language Explorer was the linguistic component and Data Notebook managed anthropological notes. The Data Notebook functions have been incorporated into Language Explorer, and “FieldWorks Language Explorer” (FLEX) is currently the only program under development. Writing about the first release in November 2006, Moe (2008) describes its purpose as follows: “Language Explorer is designed to create and manage a dictionary, create and maintain a text corpus, interlinearise texts, and study morphology”. The website for FLEX describes its functions in this way:

Fieldworks Language Explorer (FLEX) enables linguists to be highly productive when building a lexicon and interlinearising texts. Powerful bulk editing tools can save hours of work. Fieldworks allows control of which fields and entries show up in a dictionary publication. Through Pathway, beautiful dictionaries can be exported easily. Send/Receive Project allows users to collaborate with colleagues located anywhere (2018).

## **2.2 FieldWorks: Data**

FieldWorks is built on a complex data model which includes hundreds of possible fields. Dictionary entries can be described and annotated in great detail. There is also the facility to add custom fields to any project which needs to store further information. Custom fields can be added at the level of entries, senses, examples and allomorphs. One great advantage of a system that is built on a database structure is the ability to maintain referential integrity. FLEX will manage the lexical relationships between entries, so, for example, they can only be added if both entries exist in the data. Similarly before the deletion of a related item, FLEX will show a warning to alert the user. Should the entry be deleted then the lexical relationship is automatically deleted from the referent. FLEX maintains a clear distinction between the data and the presentation of the data. This allows the data to be presented in many different forms through many different media. Comprehensive dictionary formatting options are available which use HTML and CSS to format the output ready for draft printing. For even finer control over the format of the output, other programs such as Pathway can be used.

## **2.3 FieldWorks: Collaboration**

Version 8 of FLEX, released in April 2014, added the ability for multiple users to collaborate on a shared dataset. A project repository may be stored online, on a network drive or even on a USB stick. Online hosting of repositories is available at [language depot.org](http://language depot.org).

[LanguageDepot] is provided as a service to language communities by the Language Software Group of the Linguistics Institute at Payap University, Thailand. It is for teams collaborating with FLEX, WeSay and Language Forge (2018).

## **2.4 FieldWorks: Writing System and Complex Script Support**

Every piece of textual data must be expressed in one language or another, and knowledge of the language is vital to understand its meaning. You may have seen the T-shirt with the slogan “There are only 10 types of people in the world, those who understand binary and those who don’t.” The code switching from English to ‘Binary’ isn’t obvious when both languages share the same script and that can cause confusion or a loss of information. A single language can be represented with multiple scripts, for example English can be represented in Latin, IPA, Morse code or Braille. Most programs rely on the operator to know or to recognize both the script and the language of each piece of data. However, a linguist creating an orthography for an unwritten language needs to process vernacular data in closely related writing systems, such as phonetic and phonemic data both expressed in IPA. In some cases the data is identical, and without making the writing system explicit it is impossible to



know exactly what the data means. To address these potential problems FieldWorks explicitly stores the language and script information as metadata on each piece of data entered. FieldWorks supports the vast majority scripts including right to left and complex scripts such as Arabic, Devanagari and Tamil and even the sloping Arabic script: Nastaliq.

## **2.5 FieldWorks: Rapid Word Collection**

Gathering words for a dictionary in a minority language would often take a single field-based linguist many years. In order to accelerate the work the Rapid Word Collection method was developed. This involves people from the language community working together over the course of a two week workshop. During the workshop thousands of words and senses are gathered by the participants. The method was used by speakers of Gusilay in the town of Thionck-Essyl in the south of Senegal. They collected a total of 12,485 words in 11 days, a fairly typical result for a RWC workshop. FLEEx includes a tool to facilitate the Rapid Word Collection method.

While the text corpus method can produce similar results it can't be used for a language where a sufficient text corpus does not yet exist. In many cases FLEEx has been used to facilitate and inform the creation of a writing system for a minority language community. However, since such people have only recently begun to write their own language, few texts are available. For many minority languages it will be many years before a text corpus exists that is sufficiently large for a text-corpus approach.

## **2.6 FieldWorks: Interlinear Texts and Discourse Analysis**

Two parsers are included in FLEEx, one is XAmple and the other is the phonological rule-based parser HermitCrab.NET. Once the lexicon is sufficiently complete and the grammatical rules and information have been supplied, these parsers can be used to facilitate the analysis and interlinearising of texts. FieldWorks is able to analyse a text and provide guesses at the most likely morphemes and translation equivalents. This can greatly accelerate the work of interlinearising. FLEEx also contains a tool to facilitate discourse analysis of longer texts that show discourse features.

## **2.7 FieldWorks: Limitations and Contingencies**

### **2.7.1 Collaboration**

The send/receive system within FLEEx enables asynchronous collaboration. If two team members edit the same item of data and then send/receive, FLEEx will show that the edits are in conflict and allow the users to resolve it. Good inter-team communication reduces the number of conflicts, as does the practice of using send/receive before and after each session of work. If the project is stored on languagedepot then team members need an internet connection to be able to send and receive, although the bulk of the work in FLEEx can be done while offline.

### **2.7.2 OS support**

FieldWorks is now released for Windows and Linux, but there are no plans to support any other platform. However, while it won't run natively on a Mac it runs well on Windows in Parallels or VirtualBox.

### **2.7.3 Training Requirements**

Users need a good understanding of linguistics to be able to make full use of the program. However there are comprehensive resources available from the help menu that describe how to use the tools for

lexicography and interlinearization. The use of the parsers is also described in some detail. A series of training videos have been produced that should be of great help to new users of FLEx, and a few universities are now teaching FLEx as part of a Field Methods class in their undergraduate linguistics course.

#### 2.7.4 Designed for Manual Analysis

FieldWorks is designed primarily to assist linguists in their own analysis of a language. It isn't designed for corpus linguistics or automatic analysis of large quantities of text, nor is it particularly helpful for comparing words across multiple languages.

#### 2.7.5 Migrating Data from Other Tools

Data can be imported into FLEx from Standard Format Marker (SFM) files used by Toolbox and Shoebox. The task of importing data may be fairly easy if the data is consistent and simple. Often the task is complicated by the need to make the data consistent and explicit before importing. Some data formats, such as CSV, can fairly easily be transformed into SFM, but others may require significant work or specific skills in order to convert them.

The import process makes a series of XML files, each one created by processing the previous one with an XSLT. Advanced users may use one of those XML formats in order to import their data into FLEx. In theory an XSLT could be produced that would convert an existing XML or TEI file into an XML format that could be imported into FLEx.

Expert help is available for the process of importing data into FLEx from SIL International's Dictionary and Lexicography Services.

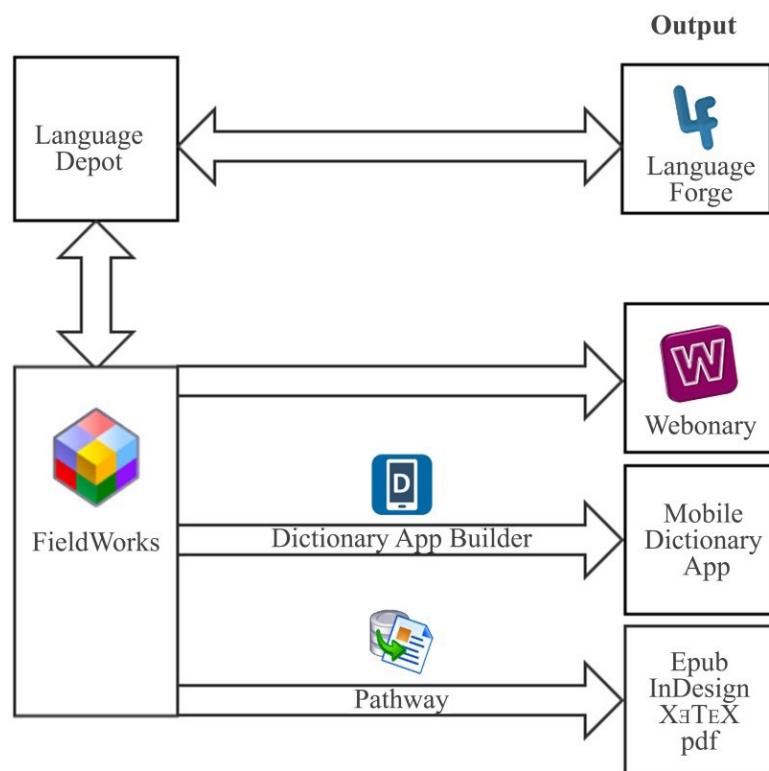


Figure 1: FieldWorks ecosystem overview.

### 3 Language Forge for online dictionary creation.

A visual representation of the interactions between FieldWorks and related programs is provided in Figure 1. This shows that the FLEEx data can be stored in a LanguageDepot repository. Language Forge provides an online interface to the data so that the same set of data can be edited with either FLEEx or Language Forge. As it operates within a browser, Language Forge can be used on most operating systems and devices, including tablets and smartphones. Language Forge may be particularly useful when many people from a language community wish to create a dictionary in their own language. One advantage that it has over FLEEx is its real-time collaboration feature. “Members of the project will see entries updated as it happens by others in the same project” (Language Forge website 2018). Another advantage is that the interface is much simpler than that of FieldWorks, and it is suitable for use by those without linguistic training. Language Forge presents a only subset of the lexical data that is available in FLEEx.

### 4 Webonary for Online Dictionary Publishing

Webonary offers online dictionary publication, particularly for minority language dictionaries. Webonary currently has dictionaries in 119 different languages, from 38 countries. “Webonary gives language groups the ability to publish bilingual or multilingual dictionaries on the web with a minimum of technical help” (2018). In one sense the task of creating a dictionary is never ending, so a small team with many other demands on their time may never get around to sharing the results of their work-to-date. Webonary lowers the technical barriers to publishing a dictionary online, and thus language communities that wish to create a dictionary for their own use, or are documenting their own language, can easily publish the results of their efforts. For language communities with no dictionary at all, even a draft dictionary is useful. The site therefore encourages editors to publish early and update often, and this is easy to do from within FieldWorks. A graphic indicates the publication-status on a scale from “Rough-draft” through to “Formally published”. Sharing the data early in the project provides the possibility of gaining input from the readers of the dictionary in the form of comments on the site. Useful feedback may be gained to improve existing entries and add new ones.

### 5 Dictionary App Builder for Smartphone Dictionary Apps

Dictionary App Builder (DAB) makes it possible to create a mobile app using data exported from FLEEx. No programming is necessary in order to create the apps; however the program does have multiple dependencies making the installation of DAB more complex than it is for most programs. The Mac OS version can create iOS and Android apps, while the Windows and Linux versions can only create Android apps. DAB allows control over many aspects of the apps it produces, including the choice of colors, fonts, and icons to be used. “The apps do not require an internet connection. All content can be packaged together for offline use and distribution, or audio can be made available online for download when needed” (Dictionary App Builder website 2018). Dictionary App Builder also contains an interface for localization, so that it is very easy for the editors to provide apps with an interface in any language. Apps can be published through the Google or Apple stores. Users of the Android apps created with DAB can share the app, complete with its data, with other Android users. This isn’t possible on iOS, as Apple prevents side-loading of apps.

## 6 Pathway for Typeset Output

Pathway provides a way to transform the dictionary data, exported from FieldWorks, into other formats. These include word processing formats such as DOCX and ODF. For higher-end typesetting, InDesign and X<sub>Y</sub>TEX output formats are available. These options allow greater control over the style and formatting of the dictionary prior to publication.

X<sub>Y</sub>TEX is an extension of TEX that integrates TEX's typesetting capabilities with (a) the Unicode text encoding standard (supporting most of the world's scripts) and (b) modern font technologies (TrueType and OpenType) and text layout services (X<sub>Y</sub>TEX website 2018).

## 7 Conclusion

FieldWorks is useful for teams and individual linguists manually analyzing a language. It supports the vast majority of complex scripts, and maintains detailed metadata about the language and script of each data point. Team members can collaborate by synchronizing the data regularly with a server. Language Forge provides real-time collaboration on a sub-set of the data. Many publishing options are provided: through Dictionary App Builder for mobile apps, Webonary for online, and Pathway for print and Epub publication. There are active user communities where answers to specific questions or advice about specific use-cases can be obtained. Finally, I would encourage experimentation with the programs if there is doubt as to their suitability for a given project.

## References

- Dictionary App Builder. Accessed at: <https://software.sil.org/dictionaryappbuilder/> [07/03/2018]  
FieldWorks. Accessed at: <https://software.sil.org/fieldworks/> [07/03/2018].  
LanguageDepot. Accessed at: <https://public.languagedepot.org/> (login required) [07/03/2018]  
Language Forge. Accessed at: <https://Language Forge.org/> [07/03/2018]  
Moe, R. (2008). FieldWorks Language Explorer 1.0 *SIL Forum for Language Fieldwork 2008-011*, pp.1-4. Dallas, USA.  
Webonary. Accessed at: <https://www.webonary.org/> [09/03/2018]  
Pathway. Accessed at: <https://software.sil.org/pathway/> [09/03/2018]  
X<sub>Y</sub>TEX. Accessed at: <http://xetex.sourceforge.net/> [25/03/2018]  
Rapid Word Collection. Accessed at: <http://rapidwords.net/> [14/6/2018]

# Wortschatz und Kollokationen in „Allgemeine Reisebedingungen“. Eine intralinguale und interlinguale Studie zum fachsprachlich-lexikographischen Projekt „Tourlex“.

**Carolina Flinz<sup>1</sup>, Rainer Perkuhn<sup>2</sup>**

<sup>1</sup>Università di Pisa, <sup>2</sup>Institut für Deutsche Sprache, Mannheim

E-mail: c.flinz@ec.unipi.it, perkuhn@ids-mannheim.de

## Abstract

Zur Vorbereitung eines zweisprachigen Fachwörterbuchs zur Tourismusfachsprache werden korpuslinguistische Verfahren eingesetzt, um Auffälligkeiten in der jeweiligen Fachsprache im Vergleich zum allgemeinen sprachlichen Gebrauch aufzuspüren. Neben den hervorstechenden Elementen des Vokabulars, den Schlüsselwörtern als potentiellen Stichwörtern, geht es vor allem um sprach- und fachsprachspezifische typische Formulierungen und deren Übersetzungsäquivalente. Für die gemeinsame, interlinguale Betrachtung des Sprachenpaars Deutsch-Italienisch wurde ein kleines Fachsprachenkorpus aufgebaut und innerhalb der Sketch Engine-Umgebung unter Zuhilfenahme der darin integrierten Referenzkorpora ausgewertet. Für eine weitere intralinguale Untersuchung der deutschsprachigen Komponente wurde auf das Deutsche Referenzkorpus DeReKo und weitere, intern zu Verfügung stehende Instrumente des Instituts für Deutsche Sprache zurückgegriffen. Neben üblichen Verfahren der quantitativen Ein- oder Mehrwortbewertung wird ein Ansatz ergänzend getestet, der der dünnen Datengrundlage im fachsprachlichen Bereich Rechnung trägt: Diese ergibt sich nicht nur aus der Korpusgröße, sondern auch daraus, dass bestimmte feste Floskeln (wie ‚eine Reiserücktrittsversicherung abschließen‘) selten rekurrent, vielmehr eher nur einmal pro Text verwendet werden. Auch wenn dieser Ansatz aufgrund infrastruktureller Artefakte in Einzelfällen an seine Grenzen stößt, die hier selbstkritisch nicht verschwiegen werden sollen, so zeigt sich doch an vielen Stellen auch das große Potential.

Abschließend wird beispielhaft illustriert, wie Evidenzen dieser und der anderen korpuslinguistischen Auswertungen lexikographisch umgesetzt wurden.

**Keywords:** Korpuslinguistik, Fachlexikographie, intralingual, interlingual

## 1 Einleitung

Korpuslinguistische Analysen haben sich sowohl für intralinguale als auch für interlinguale Studien als sehr positiv erwiesen. Ausgehend von der Sprachoberfläche lenken sie die Aufmerksamkeit auf Sprachgebrauchsmuster, die typisch für bestimmte Diskurse sind (vgl. u.a. Bubenhofer 2009; Bubenhofer et al. 2015; Bubenhofer/Scharloth 2013; Felder u. a. 2011, Flinz in Vrb.). Sie ermöglichen die Identifizierung von möglichen Übersetzungskandidaten und äquivalenten Kollokationen (Bubenhofer/Rossi in Vrb.) und können auch für lexikographische und translatorische Zwecke äußerst fruchtbar sein.

Ziel dieses Aufsatzes ist es zu zeigen, wie die Ergebnisse von intra- und interlingualen Studien konkret für die Realisierung eines zweisprachigen deutsch-italienischen Internetwörterbuchs zur Tourismusfachsprache<sup>1</sup> verwendet worden sind, insbesondere welche Auswirkungen sie auf die Auswahl

1 Das Wörterbuch ‚Tourlex‘, das sich sowohl an Lernende (Studenten der Tourismuswissenschaft) als auch an sonstige Angestellte im Tourismusbereich richtet, wird als Pilotprojekt vorerst Stichwörter aus einem besonders komplizierten touristischen Bereich („Allgemeine Reisebedingungen“) enthalten und einen besonderen Fokus auf Kollokationen und usuelle Wortverbindungen legen,



der provisorischen Stichwortliste und auf die Redaktion der Angaben im Wörterbuchartikel gehabt haben. Nach Beschreibung der Forschungsfragen werden sowohl die Korpora (das Fachsprachenkorpus „Allgemeine Reisebedingungen/Indicazioni contrattuali di viaggio“ und die Referenzkorpora) als auch die Methodologie vorgestellt. Ergebnisse der Analyse bezüglich ausgewählter Schlüsselwörter und auf die Textsorte projizierte Kookkurrenzen werden diskutiert. Abschließend werden ausgewählte Beispiele zu fachsprachlichen Kollokationen und ihre fremdsprachlichen Realisierungen anhand eines Lemmas des Wörterbuches<sup>2</sup> vorgestellt.

## 2 Forschungserkenntnisse und Forschungsfragen

Korpuslinguistische Methoden haben sich in den letzten Jahren für unterschiedlichen Disziplinen, u.a. für die Diskurslinguistik und die Lexikographie als sehr fruchtbar erwiesen.

In der Diskurslinguistik<sup>3</sup> haben korpuslinguistische Analysen die Aufmerksamkeit auf Phänomene auf der sprachlichen Oberfläche (Linke/Feilke 2009: 7) gelenkt und die pragmatische Komponente hervorgehoben. Aus den Daten können semantische, grammatische und pragmatische Informationen entnommen werden: Während sich Keyword-Analysen insbesondere für die Identifizierung von Schlüsselwörtern eignen und inhaltliche Aspekte hervorheben, geben Kollokations- und Mehrwortanalysen Informationen zum syntagmatischen Sprachgebrauch wieder. Aus der statistischen und maschinellen Bewertung von Wortformen, Lexemen und/oder Wortarten etc. findet ein neuer Zugang zu Sprache statt: Rekurrente Eigenschaften können fokussiert werden und typische Gebrauchsmuster eines Diskurses im Vergleich zu anderen, auch im interlingualen Vergleich, identifiziert werden. Die bisherigen Studien fokussieren jedoch meist den politischen Diskurs (u.a. Bubenhofer et al. 2015; Bubenhofer/Rossi in Vrb), während der touristische Diskurs nur in vereinzelten Arbeiten (Flinz 2018) Berücksichtigung gefunden hat.

Korpusbezogene Ansätze haben auch der Lexikographie einen starken Impuls gegeben (vgl. u.a. Engelberg/Lemnitzer 2009; Wiegand 1998). Korpora haben den lexikographischen Prozess verändert (vgl. u.a. Klosa 2016), vor allem haben sie sich als sehr hilfreich für die Lemmaauswahl und für die Herausfilterung der Informationen im Angabenteil des Wörterbuchartikels (Geyken/Lemnitzer 2012) erwiesen. Korpusanalysen können u.a. benutzt werden, um das Regelhafte vom Idiosynkratischen zu trennen, um Neologismen und Okkasionalismen zu identifizieren und zu beschreiben, um Anglizismen und Entlehnungen herauszufiltern, um Kookkurrenzen und Kollokationen aufzuspüren, um feste Redewendungen oder auch Partikelverwendung zu untersuchen etc.<sup>4</sup> Korpora geben Auskunft über verfestigte Gebräuche und dienen als Vorlagen für den etablierten Gebrauch eines Wortes (Lemnitzer/Zinsmeister 2015: 173).

Mit unserem Beitrag möchten wir zeigen, wie die Fachlexikographie von der Korpuslinguistik und der Diskurslinguistik erfolgreich profitieren kann und wie intra- und interlinguale Studien konkret in der lexikographischen Praxis eingesetzt werden können, in diesem Fall für die Realisierung eines zweisprachigen Fachwörterbuches zur Tourismussprache. Untersuchungsgegenstand werden Kollokationen sein, die sowohl in der Diskurslinguistik als auch in der Lexikographie bevorzugte

die Fremdsprachenlernern große Probleme bereiten. Das Wörterbuch, das in Bearbeitung an der Universität Mannheim ist, wurde dank eines Humboldtstipendiums für erfahrene Wissenschaftler am Lehrstuhl von Frau Prof. Storrer ermöglicht.

2 Für die Erstellung von Tourlex wurde das Mediawiki-System ausgewählt, das als System von Wiktionary und Wikipedia weit verbreitet und bekannt ist.

3 Pionierarbeiten waren diejenigen von Bubenhofer (2009) und Felder et al. 2011.

4 Einen Überblick über die Anwendung der Korpuslinguistik zu den obengenannten lexikographischen Zwecken bieten Lemnitzer/Zinsmeister (2015: 165-187). Die wichtigsten lexikographischen Ressourcen, die davon profitiert haben, werden dort erwähnt und beschrieben.

Analysekategorien sind, und wir werden uns der empirischen Auffassung folgend<sup>5</sup> an folgende Definition anlehnen: Wir verstehen als Kollokationen zwei oder mehr Wörter (Kollokationspartner), die überzufällig häufig, d.h. häufiger als eine zufällige Verteilung erwarten ließe, benachbart sind (Belica/Perkuhn 2015: 218).

Wie werden mit unseren Korpusanalysen folgende Forschungsfragen angehen: Welche sind die typischen Kombinationen und Vorkommenskontexte von bestimmten lexikalischen Zeichen im Fachdiskurs? Welche Unterschiede können zur Standardsprache festgestellt werden? Kann das auffällige Muster in Sprache A mit jenem in Sprache B verglichen werden? Welches ist sein Äquivalent in Sprache B?

Wir werden also zeigen, wie die Auffindung der Äquivalente nicht mehr nur in den Händen des Lexikographen liegt, sondern vom jeweiligen Diskurs herausgefiltert werden kann, da auch dort seine Relevanz entsteht (Foucault 1981).

### 3 Methodologie

Für die durchgeführte Untersuchung haben wir ein Fachsprachenvergleichskorpus<sup>6</sup> (ca. 40.000 Token pro Sprache) aufgebaut. Dabei handelt es sich um ein deutsch-italienisches Spezialkorpus<sup>7</sup>, das Texte aus der touristischen Teiltexsorte „Allgemeine Reisebedingungen“ und „Condizioni generali di contratto di vendita di pacchetto turistico“ enthält<sup>8</sup>. Es wurde mithilfe des Werkzeugs Sketch Engine aufbereitet (Kilgariff et al. 2004)<sup>9</sup>. Folgende Tabelle (Tabelle 1) liefert einen Überblick über die Zusammensetzung des Korpus:

Tabelle 1: Token und Words des deutsch-italienischen Vergleichskorpus

	Deutsches Fachsprachenkorpus	Italienisches Fachsprachenkorpus
Token	47.467	46.811
Words	38.590	39.412

Das Korpus wurde benutzt, um die vorläufige Lemmakandidatenliste (anhand der absoluten und relativen Häufigkeiten auf der Basis der integrierten Referenzkorpora German Web 2013 und Italian Web 2016) zu erstellen. Diese wurde mithilfe der Angaben aus dem Abgleich der intralingualen Analyse (s.u.) bestätigt und verfeinert. Zurzeit besteht die Lemmakandidatenliste aus 175 Stichwörtern (ausschließlich Substantive<sup>10</sup>). Die jeweiligen Fachsprachenkorpora wurden des Weiteren auf typische Wortverbindungen der ausgewählten Lemmata unter Zuhilfenahme der Informationen der Sketch Engine-Analysen gesichtet. Da keine anderen lexikographischen Ressourcen dieser Art für

5 Der empirische Begriff lässt sich auf Firth (1957: 194) zurückführen. Vgl. u.a. auch Evert (2009: 1213) der zwischen Kollokationen und Mehrwortverbindungen unterscheidet, und Steyer (2013: 76), die den Terminus ‚usuelle Wortverbindungen‘ einführt.

6 Unter Vergleichskorpora (oder auch *comparable corpora* genannt) sind Korpora zu verstehen, die Texte in mehreren Sprachen zu vergleichbaren Diskursen erfassen, die aber keine Übersetzung voneinander sind (Lemnitzer/Zinsmeister 2015: 138). Dazu vgl. auch Laffling 1992: 20; Prinsloo 2013: 1346.

7 Das ad hoc erstellte Vergleichskorpus ist ein dynamisches Korpus, das erweitert und ausgebaut werden kann.

8 Die Texte stammen alle aus dem Jahr 2017 und wurden deutschen und italienischen Reisekatalogen entnommen. Folgende Reiseveranstalter wurden berücksichtigt: Dertour, FTI Touristik, Interchalet, Kiwi Tours, Novasol, Olimar, Piccolonia, Reisegeier, Camino-Reisen, Wolters Reisen, Tui für das deutsche Korpus und Alpitour, Columbia Turismo, Eden Viaggi, Italia Tourism Online, Itermar, King Holidays, Settemari, Veratour, I Viaggi del Turchese, Viaggi dell’Elefante für das italienische.

9 Es wurden Metadaten erstellt und, da für die Veröffentlichung des Wörterbuches das Korpus zur Verfügung gestellt werden soll, wurde ein Prozedere eingeleitet, um die Nutzungsrechte zu klären.

10 Die Bearbeitung der Lemmata folgt Modulen, die sich nach Wortarten und Frequenzschichten richten.

das Sprachenpaar Deutsch-Italienisch vorhanden sind und da viele dieser Fachwörter (vor allem in diesen speziellen Verwendungen) nicht in den allgemeinen zweisprachigen Wörterbüchern beschrieben sind, wurde das Gesamtfachsprachenkorpus auch dazu verwendet, die italienischen Äquivalente zu bestimmen.

In einem weiteren Schritt wurden die Konkordanzen zusätzlich in das Tool Lexpan (Lexical Pattern Analyzer, vgl. Steyer 2013: 110f) hochgeladen, um die syntagmatischen Strukturen aufgrund von Festigkeit, Varianz, Slotbesetzungen und kontextuellen Einbettungsmustern zu untersuchen. Die daraus resultierenden Daten wurden als Grundlage für die interlinguale Analyse benutzt.

Für die Vorbereitung der intralingualen Analysen im Vergleich zum Deutschen Referenzkorpus DeReKo wurde größtenteils auf IDS-interne Werkzeuge für die folgenden Berechnungen zurückgegriffen. Neben Häufigkeitslisten der Wortformen der einzelnen deutschsprachigen Reisebedingungen wurde auch eine Liste für die Gesamtmenge dieser Texte bestimmt. Dazu wurde dasselbe Verfahren eingesetzt, über das auch eine Häufigkeitsliste des Archivs DeReKo vorbereitet worden war. Für das Gesamtvokabular der Reisebedingungen wurden dann auf der Basis dieser beiden letzten Listen verschiedene Assoziationsmaße (wie  $\chi^2$  und LLR, vgl. Dunning 1993) berechnet. Als sehr einfaches Maß für die Streuung wurde ebenfalls festgehalten, in wie vielen Reisebedingungen eine Wortform belegt ist (entspricht in diesem Fall der Texthäufigkeit). Je nach Kombination dieser Angaben kann die Information nach verschiedenen Gesichtspunkten sortiert werden. Ein sehr hohes Assoziationsmaß bei gleichzeitig geringer Häufigkeit in DeReKo kann ein Hinweis auf ein Artefakt des Verfahrens oder sogar der Quellen sein (wie z.B. der Tippfehler ‚Bestimmu9‘ auf der Webseite [www.urscher-reisen.de](http://www.urscher-reisen.de)). Hohe Assoziationsmaße (insbesondere LLR) bei gleichzeitig großer Streuung sind durchaus gute Kandidaten für Schlüsselwörter des Texttyps Reisebedingungen, somit plausible Stichwörter für das Lexikon. Hohe Assoziationsmaße mit geringer (bis zu minimaler) Streuung weisen auf idiosynkratische Formulierungen einzelner Anbieter hin, wie u.a. deren Eigennamen oder konstruierte Bezeichnungen für Teile ihres Angebots. In Einzelfällen kann es aber auch sein, dass man anerkennen muss, dass gewisse Sparten nur von wenigen Anbietern besetzt und trotzdem relevant sein können wie z.B. ‚Gruppenreisen‘, die nur ein Anbieter unserer Auswahl im Repertoire hat.

Da in dem Arbeitsumfeld der intralingualen Analyse kein Apparat zur Verfügung steht, um die intern-fachsprachlichen Kollokationen zu ermitteln, haben wir ein Verfahren angepasst und komplementär eingesetzt, das sich in ähnlicher Weise bereits für zwei andere Studien bewährt hat (Vorbereitung von Hinger 2009 und Perkuhn et al. 2015). Dazu definieren wir die Menge der Wortformen, die in den Reisebedingungen vorkommen, als das Vokabular dieser Fachsprache. Basierend auf der CCDB-Sammlung von typischen allgemeinsprachlichen Formulierungen (Belica 2007) werden die Verbindungen herausgefiltert, die nur (bzw. bis zu einem gewissen Grad) aus Wörtern bestehen, die in diesem Fachsprachenwortschatz enthalten sind. Der Einfachheit halber ist dieses Verfahren zunächst auf der Wortformenebene operationalisiert. Das Ergebnis betrachten wir als „Kollokationschatz“ (DeReKoll), eine Sammlung von typischen Formulierungen, die sich durch ihre Usualität im allgemeinen Sprachgebrauch legitimieren, sich grundsätzlich aber auch durch das Material der Fachsprache bilden ließen.

## 4 Ergebnisse und Diskussion

Die Ergebnisse der Analyse haben interessante Fälle zum Vorschein gebracht. Mit den unterschiedlichen Arbeitsumgebungen konnten nicht nur die spezifischen Kollokationen des Spezialkorpus herausgefiltert werden, weiterhin konnten sie der Standardsprache gegenübergestellt werden und es

ließen sich italienische Entsprechungen identifizieren. Wir werden uns nun auf ein ausgewähltes Beispiel (das Lexem *Rücktritt*) konzentrieren, es kommentieren und am Ende die lexikographische Darstellung präsentieren.

Ein Ausschnitt des Ergebnisses der interlingualen Untersuchung des Lexems *Rücktritt* und seiner bevorzugten Verbindungen im Spezialkorpus ist in folgender Tabelle dargestellt (Tabelle 2):

Tabelle 2: Bevorzugte Wortverbindungen des Lexems *Rücktritt* im deutsch-italienischen Vergleichskorpus und entsprechende Übersetzungen im Deutschen

	Bevorzugte Verbindungen im deutschen Korpus	Bevorzugte Verbindungen im italienischen Korpus	Übersetzung der italienischen Wortverbindungen (CF)
a.	Rücktritt des Kunden/Reisegasts	recesso del turista/del consumatore	Rücktritt des Reisenden/des Verbrauchers
b.	bei spätem/späterem Rücktritt		
c.	zum Rücktritt berechtigt sein		
d.	nach Rücktritt des Reisegastes		
e.	im Fall des Rücktritts des Veranstalters	in caso di recesso da parte dell'organizzatore/turista	im Falle des Rücktritts des Veranstalters/des Reisenden
f.		in caso di recesso	im Falle von Rücktritt
g.	einen (kostenlosen/kostenfreien/unentgeltlichen) Rücktritt anbieten	proporre il recesso; proporre il recesso senza pagare penali)	einen Rücktritt anbieten; einen Rücktritt ohne Vertragsstrafe anbieten
h.	den Rücktritt (schriftlich) erklären	esercitare il recesso per iscritto; operare recesso per iscritto	den Rücktritt schriftlich erklären
i.	Rücktritt vom Reisevertrag	recesso dal contratto/dai servizi	Rücktritt vom Reisevertrag/von den Leistungen
j.		per recesso sino/fino a x giorni per il recesso operato	für den Rücktritt bis zu x Tage; für den erklärten Rücktritt

Für alle denkbaren Konstellationen ließen sich Beispiele finden: (1) Muster, die typisch für das deutsche Spezialkorpus sind, die aber auch eine italienische Entsprechung haben (wie a., e., g., h. und i.); (2) Muster, die keine Entsprechungen im Vergleichskorpus haben (vgl. b., c., d.); (3) Muster, die im italienischen Korpus vorkommen, aber keine deutsche Entsprechung haben (wie f. und j.).

Bei der intralingualen Untersuchung mit DeReKo, speziell mit der DeReKoll-Perspektive, zeigen viele Beispiele eindrucksvolle Ergebnisse, insbesondere Lexeme, die eng (und vorrangig) mit dem Themengebiet verknüpft sind (z.B. *Buchung*). Mit dem Lexem *Rücktritt* haben wir selbstkritisch ein Beispiel ausgewählt, an dem wir trotz verhaltenem Optimismus auch die Grenzen des DeReKoll-Ansatzes illustrieren wollen. Schaut man auf die syntagmatischen Muster, die wir über unsere Methode aus den Angaben der Kookkurrenzdatenbank CCDB als Tourlex-DeReKoll-Kollokationsschatz herausgefiltert haben, ist das Bild übermäßig stark geprägt durch die thematische Zusammensetzung des zugrundeliegenden Korpus. Nahezu alle Muster sind geprägt durch Rücktritte bestimmter Personen von Positionen in Politik oder Sport. Die Angaben zu konkreten Fällen (Eigennamen oder Amtsbezeichnungen) lassen sich zwar gut erkennen, da das Material als nicht zum Tourlex-Wortschatz gehörig entsprechend markiert ist. Aber auch die übrigen Muster deuten über Personalpronomen in die gleiche Richtung. Hinweise auf Reisekontexte lassen sich nicht finden, auch die flektierten Formen von ‚erklären‘ und ‚anbieten‘ sind mit direkten Rollenverbindungen verknüpft (vgl. Tabelle 3).



Tabelle n.3: Bevorzugte Verbindungen von *Rücktritt* in der CCDB

Lexem	kanonisierte Form
Rücktritt	seinen Rücktritt [...] erklären
	zum Rücktritt [...] aufgefordert werden
	seinen sofortigen Rücktritt verlangen
	[... den] Rücktritt des ... fordern
	ihren seinen Rücktritt [...] bekannt geben
	... seinen Rücktritt [...] anbieten

Nichtsdestotrotz steckt natürlich auch in diesen Verwendungsweisen im Kern dieselbe Usualität in der Substantiv-Verb-Verbindung wie bei den reisevertragstechnischen Realisierungen. Dass wir so wenige Hinweise auf Reisekontexte gefunden haben, lässt sich auch nicht eindimensional mit der Zusammensetzung der Daten erklären. Im Gesamtarchiv ist Rücktritt ein hochfrequentes Wort, das in den allermeisten Fällen in den o.g. rollen-/amtgebundenen Kontexten verwendet wird. Vor allem der vordere Bereich des Kookkurrenzprofils wird deshalb von Verwendungen in den genannten Domänen dominiert. Um dem entgegenzuwirken wurde für die CCDB im Jahr 2007 ein virtuelles Korpus definiert, das diese Effekte aber nicht gänzlich dämpfen konnte. In der CCDB umfassen Kookkurrenzprofile darüber hinaus aus technischen Gründen maximal 253 Einträge, was im Fall von *Rücktritt* zu einem Schnitt bei dem LLR-Wert 393 geführt hat. Zieht man zum Vergleich das Kookkurrenzprofil von *Reisevertrag* heran, findet man dort die Wortverbindung ‚Rücktritt vom Reisevertrag‘ (mit LLR 55) gebucht. Selbst bei einer Kookkurrenzanalyse mit Cosmas II im aktuellen Gesamt-Datenbestand lässt sich die Verbindung ermitteln, allerdings, da hier durch die Nicht-Ausgewogenheit der Themen der o.g. Effekt wirkt, erst auf Rang 2123.

Bei diesem konkreten Beispiel haben wir es tatsächlich mit dem Zusammenkommen verschiedener widriger Umstände zu tun: Ein hochfrequentes Wort, das gerade gerne in einer Verwendungstypik in Formulierungen vorkommt in Texten zu Themen, die außerordentlich stark im Archiv vertreten sind. Ob es sinnvoll ist, die Vorkommen in den anderen, uns interessierenden Texten durch eine gezielte Zusammenmischung der Datengrundlage zu betonen, hat seine Grenzen darin, dass wir nicht verzerren wollen, welche Formulierungen allgemeinsprachlich typisch sind. Dass die Kookkurrenzprofile eigentlich nicht abgeschnitten werden sollten, ist selbstverständlich. Eine zusätzliche, methodisch sehr interessante Option wäre aber auch, für die Rollenbezeichnungen und -inhaber „Oberbegriffe“ setzen zu können, die dann entweder zu einer Instanz (quasi eine Art Lemma) zusammengefasst oder auch ausgeblendet werden könnten, sodass wesentlich kompaktere Profile entstehen könnten.

Um zu illustrieren, dass der Ansatz bereits in der jetzigen Form interessante Hinweise liefern kann, sollen einige Beispiele folgen, bei denen die Stichwörter domäneneingeschränkter belegt sind. Es sind nicht ganz zufällig eher fachsprachliche Wortbildungsprodukte, die durchaus auch weniger frequent sind, wie u.a. ‚den Reisevertrag [...] zu] kündigen‘, ‚vom Reisevertrag zurücktreten‘, ‚Reisevertrag wegen höherer Gewalt‘, ‚Abschluss einer Reiserücktrittsversicherung‘, ‚Schadenersatzansprüche [nach X Jahren] verjähren‘, ‚Verjährungsfrist [für] Schadenersatzansprüche‘, ‚nicht Bestandteil ... Reisevertrags‘, ‚einen Reisevertrag/eine Reiserücktrittsversicherung abschließen‘, ‚eine Reiserücktrittsversicherung abschließen‘, ‚Schadenersatzansprüche [gegen ...] geltend machen‘, ‚Schadenersatzansprüche [...] stellen/anmelden‘.

Allein aus diesem kleinen Ausschnitt der Sammlung wird ersichtlich, welche Fülle von Material auch diese Herangehensweise hervorbringt. Als nächster Schritt stünde allerdings noch an, zu überprüfen, ob diese Formulierungen – in welcher syntaktischen Realisierung auch immer – tatsächlich in den Fachtexten realisiert sind.



## 5 Fazit

Die Ergebnisse der übergreifenden Analysen haben Auswirkungen auf die syntagmatischen Angaben und auf die Belege im Wörterbuchartikel gehabt, denn nur auf ihrer Basis konnte die Usualität der Wortverbindungen der unterschiedlichen Lexeme (in diesem Fall *Rücktritt*) im Fachsprachenkorpus determiniert werden. Zusätzlich wurden hiermit auch die äquivalenten Verbindungen in italienischer Sprache und auch die Unterschiede zu den nicht domänenspezifischen Verbindungen festgestellt.

Als Exemplifizierung der Analyse möchten wir abschließend den Screenshot des Lemmas *Rücktritt* in Tourlex vorstellen:

### Rücktritt

#### Rücktritt, der

Inhaltsverzeichnis [Verbergen]
1 Grammatica
2 Ortografia
3 Fonetica
4 Traduzione/i
5 Combinazioni tipiche
6 Sinonimi
7 Derivazione e/o composizione
8 Parole collegate
9 Link

#### Grammatica [Bearbeiten]

Sostantivo (maschile)

	Singolare	Plurale
Nominativo	der Rücktritt	die Rücktritte
Genitivo	des Rücktritt(e)s	der Rücktritte
Dativo	dem Rücktritt	den Rücktritten
Accusativo	den Rücktritt	die Rücktritte

#### Ortografia [Bearbeiten]

Suddivisione in sillabe: Rück-tritt

#### Fonetica [Bearbeiten]

Pronuncia: der Rücktritt

Trascrizione fonetica: /rʏkˌtʁɪt/

#### Traduzione/i [Bearbeiten]

recesso (m)

#### Combinazioni tipiche [Bearbeiten]

##### • Aggettivo - Sostantivo:

kostenloser/unentgeltlicher Rücktritt: recesso senza penali; recesso senza pagare penali

##### Esempi:

Gegebenenfalls werden wir Ihnen eine kostenlose Umbuchung auf ein anderes Ferienobjekt von NOVASOL oder - falls kein gleichwertiger Ersatz existiert - einen kostenlosen Rücktritt anbieten. (Novasol 2017)

##### • Sostantivo - Verbo:

einen kostenlosen/unentgeltlichen Rücktritt anbieten: proporre un recesso senza penali/pagare penali

##### Esempi:

Gegebenenfalls werden wir Ihnen eine kostenlose Umbuchung auf ein anderes Ferienobjekt von NOVASOL oder - falls kein gleichwertiger Ersatz existiert - einen kostenlosen Rücktritt anbieten. (Novasol 2017)

den Rücktritt (schriftlich) erklären: comunicare il recesso (per iscritto)

##### Esempi:

Es wird empfohlen, den Rücktritt schriftlich zu erklären. (Dertour 2017)

##### • Altre:

Rücktritt vom Reisevertrag: recesso dal contratto

##### Esempi:

Spätere Änderungen sowie Änderungen über den Geltungszeitraum der der Buchung zugrunde liegenden Katalogausschreibung hinaus können nur nach Rücktritt vom Reisevertrag zu den Bedingungen gemäß Ziffer 7.5 bei gleichzeitiger Neuanmeldung vorgenommen werden. (Wolters 2017)

Im Falle des Rücktritts durch X: In caso di recesso da parte del turista

##### Esempi:

Im Fall des Rücktritts durch Olimar nach Ziffer 8.3 ist der Kunde berechtigt, die Teilnahme an einer mindestens gleichwertigen anderen Reise zu verlangen. (Olimar 2017)

In caso di recesso dal contratto da parte del Turista prima della partenza al di fuori dei casi elencati ai precedenti commi del presente articolo e nel caso previsto dall'articolo 5, secondo comma, sarà addebitata una penale (Alpitour 2017)

den Rücktritt gegenüber x erklären: invocare il recesso nei confronti di x

##### Esempi:

„der Rücktritt ist gegenüber Piccolonia unter der in diesen Bedingungen angegebenen Anschrift zu erklären. (Piccolonia 2017)

zum kostenlosen Rücktritt berechtigt sein: poter recedere senza pagare penali

##### Esempi:

„so sind Sie deshalb nicht zum kostenfreien Rücktritt vom Reisevertrag berechtigt. (Kiwi Tours 2017)

Il turista può recedere dal contratto, senza pagare penali, nelle seguenti ipotesi. (Viaggi dell'Elefante 2017)

Frist zur Ausübung des Rücktritts: termine ultimo per esercitare il diritto di recesso

##### Esempi:

In diesem Falle ist die Zahlung erst dann fällig, wenn die Frist zur Ausübung des Rücktritts - rechts abgelaufen ist. (FTI 2017)

der späteste Zeitpunkt des Rücktritts: termine ultimo per esercitare il diritto di recesso

##### Esempi:

Der späteste Zeitpunkt des Rücktritts durch Piccolonia muss in der konkreten Reiseausschreibung oder, bei einheitlichen Regelungen für alle Reisen oder bestimmte Arten von Reisen, in einem allgemeinen Kataloghinweis oder einer allgemeinen Leistungsbeschreibung angegeben sein. (Piccolonia 2017)

Rücktritt und Nichtantritt: recesso e mancata partenza

Rücktritt und Kündigung: recesso e disdetta

##### Esempi:

Rücktritt und Kündigung durch Camino-Reisen. (Urscher 2017)

#### Sinonimi [Bearbeiten]

#### Derivazione e/o composizione [Bearbeiten]

Reiserücktritt

#### Parole collegate [Bearbeiten]

Vertrag

#### Link [Bearbeiten]

canoo.net<sup>9</sup>

DWDS<sup>9</sup>

OWID<sup>9</sup>

Bild 1: Screenshots des Lexems *Rücktritt* in Tourlex  
(<https://wiki.uni-mannheim.de/tourlex/index.php?title=Rücktritt>)

## Literatur

- Belica, C. (2007). *Kookkurrenzdatenbank CCDB - V3*. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. <http://corpora.ids-mannheim.de/ccdb/>.
- Belica, C., Perkuhn, R. (2015). Feste Wortgruppen/Phraseologie I: Kollokationen und syntagmatische Muster. In U. Haß, P. Storjohann (Hrsg.) *Handbuch „Wort und Wortschatz“*. (= Handbücher Sprachwissen 3). Berlin/Boston: de Gruyter. S. 201-225.

- Bubenhof, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin, New York: de Gruyter (Sprache und Wissen).
- Bubenhof, N. et al. (2015). Rhizome digital: Datengeleitete Methoden für alte und neue Fragestellungen in der Diskursanalyse. In *Zeitschrift für Diskursforschung, Sonderheft Diskurs, Interpretation, Hermeneutik* 1, S. 144–172.
- Bubenhof, N., Rossi, M. (in Vrb.). Die Migrationsdiskurse in Italien und der Deutschschweiz im korpuslinguistischen Vergleich. In R. Goranka, E. Schaefroth (Hrsg.) *Methoden der vergleichenden Diskurslinguistik. Germanistisch-romanistische Beiträge zur Methodenreflexion und Forschungspraxis* ---
- Bubenhof, N., Scharloth, J. (2013). Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren. In I. Warnke, U. Meinhof, M. Reisigl, Martin (Hrsg.) *Diskurslinguistik im Spannungsfeld von Deskription und Kritik*. Berlin: Akademie-Verlag (Diskursmuster – Discourse Patterns), S. 147–168.
- Institut für Deutsche Sprache (2017). Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-I (Release vom 08.03.2017). Mannheim: Institut für Deutsche Sprache. PID: 10932/00-0373-23CD-C58F-FF01-3.
- Dunning, Ted (1993): Accurate methods for statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61–74.
- Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. 4. überarb. u. erw. Aufl. Tübingen: Stauffenburg.
- Evert, S. (2009). 58. corpora and collocations. In A. Lüdeling, M. Kytö (eds.) *Corpus Linguistics*. Berlin/New York: de Gruyter, S. 1212–1248.
- Felder, E. et al. (2011). *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin/New York: de Gruyter.
- Firth, J.R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis*. Philological Society. Oxford: Blackwell, S. 1-32.
- Flinz, C. (2018). Off the beaten track oder Massentourismus? Eine kontrastive Untersuchung deutscher und italienischer Orientierungstexte in Mallorca-Reiseführern. In: U. Schaffers, S. Neuhaus, H. Diekmannshenke (Hrsg.) (Off) *The beaten track? Normierungen und Kanonisierungen des Reisens*. Königshausen & Neumann. 2018, S. 51-67.
- Flinz, C. (in Vrb.). Persuasionstrategien in deutschen rechtsorientierten Zeitungen. Eine korpuslinguistische Studie. In: F. Ricci Garotti, M. Moroni (Hrsg.) *Sprache und Persuasion*. Sonderheft der Zeitschrift *Linguistik Online*.
- Foucault, M. (1981). *Archäologie des Wissens*. 10. Frankfurt am Main: Suhrkamp.
- Geyken, A., Lemnitzer, L. (2012) Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In *Proceedings of EURALEX 2012*. Oslo, S. 362–366.
- Heringer, H.J. (2009). *Valenzchunks. Empirisch fundiertes Lernmaterial*. München: Iudicium.
- Kilgariff, A. et al. (2004). The Sketch Engine. In G. Williams, S. Vessier (Hg) *Proceedings of the 11 th Euralex International Congress*, Lorient, France, July 6-10. Bd. 1. S. 105-115.
- Klosa, A. (2016): Der lexikographische Prozess im Projekt elexiko. In: V. Hildenbrandt, A. Klosa (Hg.): *Lexikographische Prozesse bei Internetwörterbüchern*. Mannheim: Institut für Deutsche Sprache. (OPAL 1/2016). S. 29-39.
- Laffling, J. (1992). On constructing a Transfer Dictionary for Man and Machine. In: *Target*, 4, S. 17-31.
- Lemnitzer, L., Zinsmeister, H. (2015): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.
- Linke, A., Feilke, H. (Hg.) (2009): *Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt*. Berlin/New York: Niemeyer.
- Perkuhn, R., Belica, C. et al. (2015). Valenz und Kookkurrenz. In: M.J. Domínguez Vázquez, L.M. Eichinger. (Hrsg.) *Valenz im Fokus. Grammatische und lexikographische Studien*. Festschrift für Jacqueline Kubczak. Mannheim: Institut für Deutsche Sprache, S. 175-196.
- Prinsloo, D. J. (2013). The utilization of bilingual corpora for the creation of bilingual dictionaries. In R.H. Gouws, et al.: *An International Encyclopedia of Lexicography*. Berlin/Boston: De Gruyter, S. 1344-1356.
- Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr Verlag.
- Wiegand, H.E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1. Teilband. Berlin/New York: de Gruyter.

# Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages

*Mika Hämäläinen, Jack Rueter*

*Department of Digital Humanities, University of Helsinki*

*E-mail: mika.hamalainen@helsinki.fi, jack.rueter@helsinki.fi*

## Abstract

We present our ongoing development of a synchronized XML-MediaWiki dictionary to solve the problem of XML dictionaries in the context of small Uralic languages. XML is good at representing structured data, but it does not fare well in a situation where multiple users are editing the dictionary simultaneously. Furthermore, XML is overly complicated for non-technical users due to its strict syntax that has to be maintained valid at all times. Our system solves these problems by making a synchronized editing of the same dictionary data possible both in a MediaWiki environment and XML files in an easy fashion. In addition, we describe how the dictionary knowledge in the MediaWiki-based dictionary can be enhanced by an additional Semantic MediaWiki layer for more effective searches in the data. In addition, an API access to the lexical information in the dictionary and morphological tools in the form of an open source Python library is presented.

**Keywords:** online dictionary, collaborative editing of XML, Semantic MediaWiki dictionary

## 1 Introduction

In this paper, we present advances in the development of our open-source synchronized XML-MediaWiki dictionary environment<sup>1</sup> (Rueter & Hämäläinen 2017). The dictionary data consists of multiple XML dictionaries for small Uralic languages<sup>2</sup> following the same XML structure. XML dictionaries are used on the Giellatekno infrastructure (Trosterud, Moshagen & Pirinen 2013) for many distinct facets of linguistic research such as Intelligent Computer-Assisted Language Learning (ICALL) (Antonsen et al. 2014), FST generation for morphological analyzers and spellcheckers.

XML is a great format for storing structural data, such as information usually stored in a dictionary. It does, however, have some drawbacks, such as editing XML data in a collaborative fashion is a challenging task. This is even more so in the case of non-technical people native in endangered Uralic languages. In order to enable them to produce and correct dictionary resources, a simplified way to edit XML data is needed.

We have thus developed a MediaWiki-based online dictionary system, the purpose of which is to make it possible to edit structural dictionary data collaboratively with a simplified interface. The dictionary works in such a way that we can get the edits instantly in our XML formalism, and edits made directly in the XMLs are also updated to the MediaWiki.

<sup>1</sup> Available on <https://sanat.csc.fi/>

<sup>2</sup> The languages currently supported in the dictionary are Skolt Sami, Ingrian, Meadow Mari, Votic, Olonets-Karelian, Erzya, Moksha, Hill Mari, Udmurt, Tundra Nenets and Komi-Permyak

## 2 Related Work

This section presents some of the previous research done in the context of online dictionaries. The previous work ranges from theoretical takes on online dictionaries to actual online systems implemented for the task. In a meta-analysis of studies on the usage of electronic dictionaries (Töpel 2014), several advantages in electronic dictionaries were identified. A positive impact on speed, performance, ease of use, vocabulary retention and satisfaction were reported in a dictionary use situation.

The XML structures of this project are compatible with and, where possible, identical to those used by the dictionaries in the Giellatekno infrastructure, where local enhancement provides the availability of special glyphs for assistance in individual language input, links to corpora search in Giellatekno-hosted Korp, as well as grammatical links for the enlightenment of the lay user<sup>3</sup>. Whereas the Giellatekno dictionaries provide for dictionary users without specific keyboards for the individual languages, we require our users to have keyboards of their own<sup>4</sup>. Pointer data in our XML and MediaWiki interfaces allow us to open individual page links to the etymological database for Sami languages (Álgu-tietokanta 2002).

Uralic language databases are the target of continuous development in Estonia. This can be observed in the outline of Estonian and Uralic language archive materials in Tallinn and Tartu (Viikberg 2008), and subsequent mention of work on Estonian-Mari and Estonian-Erzya dictionaries at EKI (Eesti Keele Instituut [Estonian Language Institute]) in EELEX (Tender et al. 2017). Similar bilingual dictionary development with audio resources are described for Võro-Estonian (Männamaa & Iva 2015).

The Dictionary of Old Norse Prose (ONP) (Johannsson & Battista 2017) has implemented multiple search and presentation features. It strives towards an online tool with enhanced corpus search and allows for presentation of manuscript and archive materials, as well as individuated download possibilities. In our project, however, we retain a light structure with synchronic editing for XML and MediaWiki. The XMLs contain a set of hand-selected example sentences from corpora to be displayed to the user in the online dictionary. However, our system has not been linked to full corpora for example sentence extraction.

The role of e-lexicography is growing. Not only is the detail required for the conversion from printed dictionaries to digital format being examined, but investigations are also being made of the feasible saturation of data presentation. E-lexicography allows for the introduction of new tools, and is seen as an opportunity to provide direct data extraction from various data sources (Bothma, Gouws & Prinsloo 2017). Our MediaWiki presentation involves three dimensions of linking. It includes links to external datasets (etymology and audio), other languages in the internal dataset (definitions, etymology), and dictionary internal links between articles (compound word constituents and derivation stems). We also generate regular paradigmatic tables for viewing while retaining a view of lemma, native definition, translation and morphologically important category information on the screen.

3 Giellatekno online morphologically savvy dictionaries with click-in-text readers and possible Korp links are available at: <http://sanit.oahpa.no/> (North Sami), <http://baakoeh.oahpa.no/> (South Sami), <http://saanih.oahpa.no/> (Inari Sami), <http://saan.oahpa.no/> (Skolt Sami), <http://sanat.oahpa.no/> (Northern Balto-Finnic languages), <http://sonad.oahpa.no/> (Southern Balto-Finnic languages), <http://valks.oahpa.no/> (Mordvin languages), <http://muter.oahpa.no/> (Mari languages), <http://kyv.oahpa.no/> (Permic languages), and <http://vada.oahpa.no/> (Nenets).

4 The necessary keyboards for most Uralic languages are produced for Windows, Mac and Android and available at <http://divvun.no/> for Saamic languages, and analogical keyboards for other languages can be generated directly in the Giellatekno infrastructure.

### 3 The XML Dictionaries

The XML dictionaries draw upon the goal of minimizing data redundancy in different branches of an extended infrastructure at Giellatekno (Trosterud, Moshagen & Pirinen 2013). Original parallel sources existing for online morphologically savvy translation dictionaries, on the one hand, and minimal sized ICALL dictionaries, on the other, have been integrated with lemma:stem pair data utilized in transducer production. Subsequently, other research data has been incorporated into the XML structure as well, such as audio pointers, and etymological as well as derivational information partially inherited from previous language projects. Thus, while the dictionaries can be used through XSL transformation to provide code output (lexc) for the construction of transducers used in finite-state morphological analyzers and spell checkers, they also serve as extensive databases for other research projects. The distinction between source and target languages is maintained utilizing ISO 639-3 three-letter codes, which can be attested in the XML root element as well as the translation group <tg/>, example group <xg/>, etymon and cognate elements.

The translation dictionaries were originally set up as source-to-target, bi-lingual dictionaries. In word entries with broader semantic coverage, granularity has been introduced. This allows for multiple translations in the instance of semantically close definitions <t/>, and separate meaning groups <mg/> for distinct senses of a word. Contextual usage is demonstrated in Figure 1, which shows translation groups within the semantically appropriate meaning group. The Giellatekno dictionaries based on this XML structure are available and undergoing continuous development within the Giellatekno infrastructure.

```

2  <e>
3    <rev-sort_key>issaa</rev-sort_key>
4    <lg>
5      <l pos="N">aassi</l>
6      <etymology>
7        <etymon algu_lekseemi_id="82224" id="246450" xml:lang="sms">aassi</etymon>
8      </etymology>
9      <stg>
10       <st Contlex="N_PRSPRC-VVKK-I">aassi</st>
11     </stg>
12     <inc-audio>
13       <c name="ID_Audio">3341</c>
14     </inc-audio>
15     <comp drv="V»N" type="Der">
16       <comp drv="" ord="E2" pos="Suf">Der/NomAg</comp>
17     </comp>
18   </lg>
19   <mg relId="0">
20     <tg xml:lang="eng">
21       <t pos="N">resident</t>
22     </tg>
23     <tg xml:lang="rus">
24       <t pos="N">житель</t>
25     </tg>
26     <xg>
27       <x src="JS2 06749_2az 0:02:36">To'b lij poostai päi'kk, jeä'la ni keäk jeänaš aazzi.</x>
28       <xt xml:lang="fin">Se on syrjäseutua, ei ole juuri ketään asukkaita.</xt>
29     </xg>
30   </mg>
31 </e>

```

Figure 1: XML entry

Optional enhancement of the underlying lemma (e/lg/l), stem (e/lg/stg/st) and inflection (e/lg/stg/st@Contlex) dictionaries can be observed in the etymon and audio pointers, as well as the derivation (e/lg/comp), translation (e/mg/tg) and example (e/mg/xg) groups. While the lemma, stem and continuation lexic data serve as vital information in transducer development, etymology, audio and compounding pointers provide for navigation between and within dictionaries. The etymon pointer



is used to access an online Sami language etymology dictionary, whereas an optional cognate sibling allows for pointing between languages in the MediaWiki infrastructure. Likewise, the inc-audio/audio pointers allow for accessing recordings in the Max Planck Institute archives at Nijmegen, and compounding group pointers offer access for navigation within the source language at the lemma and suffix levels.

Homonymy is addressed on a part-of-speech basis, with words bearing mutual etymological and inflectional data subordinated to single entries but feasibly different senses.

## 4 The Synchronized Dictionary System

The synchronized dictionary system we are proposing is meant to solve the problem of collaborative editing of XML dictionaries. Having multiple editors modifying the contents of pure XML dictionaries simultaneously is not an easy task to accomplish. It gets even more difficult if the editors have only a very limited technical background and from little to no understanding of the XML syntax. Large-scale tasks such as crowdsourcing of dictionary editing become next to impossible with plain XML files.

Another XML specific limitation our system is made to solve is breaking the tree structure of XML. Our dictionary system can build links in between different lexical entries even across multiple dictionaries to provide a more graph like structure of the dictionary data. This also makes it possible to conduct more complex queries to search for information stored in the dictionary system.

What makes our system synchronized is that we do not want to move entirely away from the XML standard, but rather build a system in which the same dictionary information can be edited in an easier crowdsourced fashion in a MediaWiki environment and also directly in the XMLs, so that edits at either end of the system will be made instantly available to all viewers of the dictionary system.

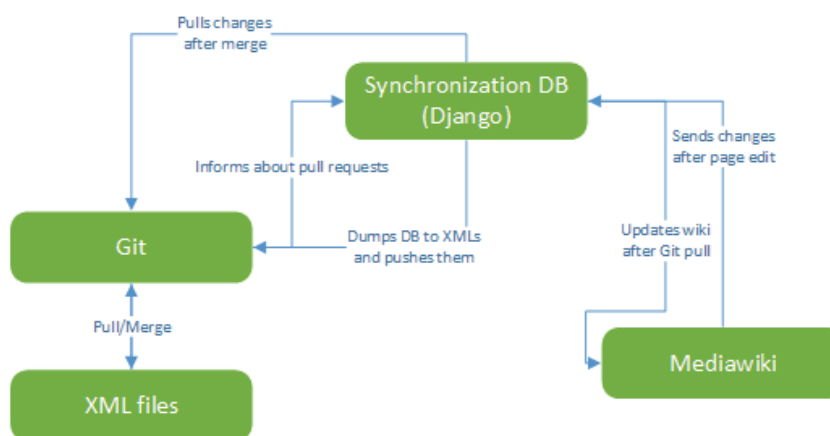


Figure 2: System architecture

The core of the dictionary system is the Django-based Synchronization DB seen in the middle of Figure 2. This database application provides APIs to the Git integration for making changes to the XML files and also APIs to the MediaWiki integration to communicate with the Wiki environment. The different parts of the system are discussed in more detail in the following subsections.

The non-functional requirements for the system are reliability and scalability. The dictionary system should not only serve researchers but also any language user outside of academia, which is a reason

why reliability of the platform is needed to guarantee a decent uptime. One of the design principles is that we should be able to include new dictionaries in the system, which means that it should scale well. The system should also be built fully on open source technologies in order to ensure its compatibility and maintainability in the future. Because the underlying MediaWiki platform and the XMLs will be used for other purposes than those required by our system, we also need to follow the idea of separation of concerns in order to fill a criterion of integratability.

#### 4.1 Synchronization Database

The role of the synchronization database is to keep the most up-to-date version of the data in all situations. This makes it possible to isolate the synchronization feature from the XMLs and the MediaWiki, making it possible to introduce new sources and views to the data in the future. These might be XMLs following a different structure or an entirely new system for collaborative editing. By embracing the notion of separation of concerns, we do not want to build the synchronization database to follow the structure of MediaWiki syntax or XML syntax, but rather we want it to have its own scalable structure.

When constructing the system, we want to keep the option open for introducing new data sources, in XML or in another format. This means that the contents of the data are not predictable, and thus defining an SQL database would make incorporating new kinds of data difficult. Storing data in plain XML format is not a viable option either, as using an XML database has a huge negative impact on the performance of the system (Nicola & John 2003).

We thus propose using MongoDB as a solution for storing the data in an effective fashion. MongoDB is a so-called NO-SQL database which does not require a predefined structure for the database. In performance terms, it can run faster than a traditional SQL database (Boicea, Radulescu & Agapin 2012), making it a good option for our purposes.

The database application is a Django-based web application. Communication from the Git side and MediaWiki side with the database is thus done by using HTTP requests to the web application API. The process of bringing XML files to the system is done over Git.

When XMLs are edited or brought for the first time to the system, the changes made in them are committed to a Git repository. If the XML dictionaries already exist in the system, the editor has to run a special command line script that will create a new branch and dump the data from the synchronization database into XML format in that branch. This leaves the conflict resolution to the editor of the XML files. He can compare his current working branch with the latest data in the synchronization database, resolve the possible conflicts and merge the branches to the master branch. When the repository is pushed, the synchronization database pulls the changes and updates its internal database after which it starts updating the MediaWiki side.

The XMLs are read into the internal JSON format of the system by language and/or XML structure specific modules. When the XMLs are requested from the system, a format and language specific Django template is used to produce the XML structure. This conversion process of the data is explained in more detail in the next section.

#### 4.2 Support for Multiple Languages

The system is built in a modular way to facilitate the inclusion of new languages or data sources. At the moment, all of the languages in the system follow the Giellatekno XML syntax, which means that the same modules are reused just with a different language flag. The system needs two language modules, one to handle the XML to JSON conversion for MongoDB and another to handle the JSON

to MediaWiki syntax conversion. We are dealing with languages whose orthographies contain special characters. This means, for multiple language support, that we have made sure the data is handled in UTF-8 format in all parts of the process.

Since the synchronization database itself is unaware of the contents of the data, how the XML gets transformed into JSON can be decided for each language module separately to better suit the needs of each dictionary type. The module currently developed for the Giellatekno XML does quite a direct transform of the XML data into JSON format. We do, however, handle homonyms differently in the JSON. In the Giellatekno format, homonyms are completely separate entries in the XML with a *hid* attribute to indicate the ID of the homonym. In the JSON format, it will be noted, we include all homonyms in a list under the main entry, which is identified by the lemma. The reason for this is simple: on the MediaWiki side all the homonyms are listed inside the same article which is identified by the lemma. Having all the different homonyms in the same entry in the synchronization database makes producing a MediaWiki page much simpler.

The other part of the language module is a script that can be run both in the synchronization system side and in the MediaWiki side to do a conversion between JSON and MediaWiki syntax. The important part is that the MediaWiki syntax is only used for the visualization of the dictionary data. For editing the dictionary entries in the MediaWiki side, a dump of the JSON data is included in the article in a hidden div element.

### 4.3 MediaWiki Integration

The MediaWiki integration is an extension which is isolated to work with a predefined set of namespaces. Our system creates a new MediaWiki namespace for each language. In practice, this means that each entry is prefixed by a three letter ISO language code, for example the Skolt Sami word *sokk* is stored inside of the MediaWiki article named *Sms:sokk*. The reason why it is important to limit the functionality with namespaces is not only that the namespace tells which language module should be used, but also that our dictionary system is a part of a shared MediaWiki dictionary of the Language Bank of Finland with multiple different data providers. This additional namespace restriction makes sure that our solution does not interfere with the MediaWiki entries other projects are building.

The MediaWiki extension of our system, in addition to communicating the changes to the synchronization database, provides the functionality for two MediaWiki article views: visualization and editing. A language specific module is used to construct a viewable version of a dictionary entry, or an article in the MediaWiki terminology. As described before, this viewable version stores the JSON structure as a hidden element for editing purposes.

The editing part of the MediaWiki extension solves the problem of XMLs requiring additional technical knowledge to be edited. The edit view of a MediaWiki article hides the MediaWiki syntax editor that would be shown by a MediaWiki based system by default. Instead, the editor constructs a form based on the language module and the hidden JSON element as seen in Figure 3. When we force users to edit the data through a form, we can make sure that the data is in a valid, parseable format. There is thus no possibility for the user to accidentally break the syntax of the data structure by, for example, forgetting a closing tag. Additionally, using a form for editing makes it possible for us to do form validation before saving the data in the system. At the moment, the validation means removing empty entries, such as a language entry without any translations.

Saving the edit form makes the system update the hidden JSON element and reconstruct the edit view based on the new JSON data using the exact same functionality as when a synchronization database pushes a JSON entry to the MediaWiki side. New changes are then immediately communicated to the synchronization database through the MediaWiki extension.

Figure 3: Form in MediaWiki

Since the MediaWiki stores each dictionary entry as a separate article, and the synchronization database does a similar separation, collaborative editing is made possible. Changes can be communicated between the two systems per entry basis without the need to parse an entire collection of lemmas, as in the case of XML. This structural separation of entries means that if different dictionary entries are edited simultaneously, there will not be any conflicts, but multiple edits can be synchronized in real time. The only case of simultaneous editing that is not supported is when the same MediaWiki article is edited at the same time by multiple users.

In addition to editing and visualizing the data, the MediaWiki integration has a search functionality for accessing the dictionaries. This is needed because the MediaWiki environment contains so many different dictionaries and word lists that using the default search box provided by MediaWiki makes it next to impossible to find the words in the system for an average user who is not familiar with the namespacing used in the system.

Figure 4: Easy search interface

The simplified search interface is depicted in Figure 4. It provides the functionality of picking the dictionary in which the words are searched, such as the Skolt Sami dictionary. Due to the highly inflectional nature of Uralic languages, a language learner might come across with a non-lemmatized form of a word. For this reason, our search interface incorporates morphological analyzers to lemmatize the user input word form. As seen in Figure 4, the search term used was *so'kke*, and the system found that it is an inflectional form of *suukkâd*, *sookkâd* and *sokk*. The inclusion of this feature is also motivated by previous research (Bergenholtz & Johnsen 2005) pointing out that online dictionary users use non-lemmatized word forms (the passive and imperative forms of a verb in their study) when consulting a dictionary.

It is also possible to use the same search to find words in the translations. This means that by inputting the English word *row*, the system will find the Skolt Sami entry *suukkâd*. The simplified search interface also provides a link to the full MediaWiki entry.

#### 4.4 Semantic MediaWiki

Semantic MediaWiki (Völkel et al. 2006) is an extension that has been used in the past in the Language Bank of Finland MediaWiki environment with good experiences in the context of online dictionaries (Laxström & Kanner 2015). The extension makes it possible to link MediaWiki articles together based on shared semantic characteristics. The aim of the extension is to make semantic knowledge in a MediaWiki environment machine readable.

We use Semantic MediaWiki to gain access to a more graph-like representation of the dictionary data. We use it to enhance the MediaWiki entries with property tags in an automated fashion. The property tags are added or updated to the MediaWiki articles automatically always when new edits are made.

**Semantic search** ? Help

**Query**

```
[[Lang::Sms]] [[tr_eng::no]] [[POS::V]]
```

**Additional data to display**  
(add one property name per line)

```
?Contlex  
?Assonance
```

Format as: Broad table (default) For a detailed description, please visit the [Broad table \(default\)](#) help page.

Search

Find results
Hide query
Show embed code

The query `[[Lang::Sms]] [[tr_eng::no]] [[POS::V]]` was answered by the `SMWSQLStore3` in 0.4906 seconds.

**Results 1 – 50** (Previous 50 | Next 50) (20 | 50 | 100 | 250) (JSON | CSV | RSS | RDF)

	Contlex	Assonance
-škue'tted	V SHKUEAQTTED	-CCue'CCeC
aaibšed	V TAARBSHED	AaiCCeC
aaibšeškue'tted	V SHKUEAQTTED	AaiCCeCCue'CCeC
aalgtoõllâd	V LAUKKOOLLYD	AaCCCõõCCâC
aassâd tâä'lv	V	AaCCâC Cää'CC

Figure 5: Semantic MediaWiki search

The property tags such as *tr\_eng* or *Contlex* make it possible to query the dictionary information more effectively through the Semantic MediaWiki query interface. In Figure 5, we see how we can get a list of all Skolt Sami (*Lang::Sms*) words that do not have an English translation (*tr\_eng::no*) and are



verbs (*POS::V*). We can also specify the property values we want to be visualized in the search results such as continuation lexicon (*Contlex*) and the assonance rhyme structure of each word (*Assonance*). These queries can be made within one dictionary or across multiple dictionaries stored in the system by altering the *Lang::* query parameter.

Furthermore, the extension allows us to access other entries of the same dictionary or entries of completely different dictionaries in the same system. This is achieved with the *pages that link here* functionality. This means that we can see, for each entry in the dictionary, if there is another entry possibly even in a different dictionary making a reference to a specific entry. Currently, these references might be translations, derivations or etymologies. In other words, just by having an etymological relation defined in the Skolt Sami dictionary, we can see the reference in the Erzya dictionary, for instance.

#### 4.5 The API

As the dictionary uses morphological tools for different tasks, such as producing inflection paradigms when viewing an article in MediaWiki or lemmatizing input words in the simplified search view, the dictionary system has in built functionality that can be of a general interest when doing NLP for Uralic languages. This is the reason why we have decided to serve the morphological tools over an API that is currently usable through a Python library called Uralic NLP<sup>5</sup> (Hämäläinen 2018).

The underlying functionality relies on finite-state transducers based on the HFST tool (Lindén et al. 2013). These are openly available in the Giellatekno infrastructure (Trosterud; Moshagen & Pirinen 2013) in a source code format. Our API provides easy access to precompiled versions of the FSTs for morphological analysis, generation and lemmatization. In addition to the FSTs, the API makes it possible to get full JSON entries for words in the dictionary.

Apart from our own extended API, the standard MediaWiki API and Semantic MediaWiki API are available for the users. These provide a standardized access to the data stored in the MediaWiki side of the system, such as using the Semantic MediaWiki query language.

## 5 Lexicographical Difference of the XMLs and MediaWiki

Each dictionary is tailored to a different audience or user group. Whereas the XML dictionaries have been set up to act as virtually stand-alone databases that can be used for deriving any variety of output sets, the MediaWiki dictionaries have been set up to provide a less cluttered experience. In fact, the visible code in the MediaWiki presentation is less than what can be found in the XMLs. This design decision was taken to better support the end user goals when using the dictionary. A typical dictionary user is more likely to be interested in definitions and translations than metadata or FST specific information needed to produce the morphological analyzers. Visualizing too much information that is irrelevant for the user goals makes it harder for the user to find the relevant pieces of information. This would cause higher cognitive load which would take up more working memory (Paas, Renkl & Sweller 2003), which is the very thing we want to avoid with our design choice. Previously it has also been reported that extremely extensive entries cause difficulties in using the dictionary (Selva & Verlinde 2002).

The MediaWiki dictionaries utilize three different types of links. Etymon and audio links provide access to sites of external institutions, such as the Sami-language etymological database Älgu at the Institute for the Languages of Finland in Helsinki, and the Max Planck Institute audio archives in

<sup>5</sup> Instructions and installation on <https://github.com/mikahama/uralicNLP>

Nijmegen. Cognate links facilitate navigation between languages in the namespace of our project on the CSC/Language Bank server, while compounding and derivation links enhance the navigation experience between compound words and their constituents in the same manner as derived words point to their derivational stems and morphemes. This interlinking provides a new alignment of semantic and morphological data not immediately accessible from the XML databases.

Not all homography is dealt with by means of Roman numeral identification. In fact, the development of XML dictionaries has led to the separation of homographs according to part-of-speech designation. When the MediaWiki dictionaries return all homographs to adjacent micro-entries within the macro-entries, micro-entries with the same part-of-speech designation are distinguished, as in the XML dictionaries, according to homograph enumeration, while other instances of homography are simply addressed with the help of part-of-speech marking.

Semantic tag values with synset distinctions are used in some language development at Giellatekno. In anticipation of shared meaning groups in source-to-multi-target-language dictionaries, this initial semantic tagging has been introduced in the XML dictionaries, where they reflect the same semantic tagging used in Constraint-Grammar disambiguation applied in the Giellatekno and Apertium infrastructures, and the ICALL infrastructure at Giellatekno. Initial outlines have also been drafted for editing semantic links that will enhance searches for various degrees of synonymy.

## 6 Discussion and Future Work

Our system is under continuous development, but it has reached a functional state. At the moment, we have several authors editing the Skolt Sami and Erzya dictionaries in the MediaWiki environment, while part of the dictionary editing is still ongoing in the XMLs. In this case of a handful of editors, the system has proved functional. The biggest limitation in the system, however, is the Semantic MediaWiki extension. Enabling the extension has a huge impact on the speed of the system when updating the entries in the MediaWiki side. We are currently finding ways to overcome this limitation.

The development has focused mainly on the technical side of the environment. Since the system is meant to be used by people with no linguistic or technical background, more research is needed in terms of usability and user experience of the system. This is especially needed and, in general, understudied in the context of editing the dictionary entries.

Giellatekno XMLs have the problem that they are not standardized by any means. This could be solved by remodeling the XML structure in a standardized TEI format. Since our system is built with multiple XML formalisms in mind, introducing a new TEI based format should not be too big of an issue. In fact, by writing a new template we can already start producing a TEI formatted version of the XML data.

The non-functional requirements of the system, reliability and scalability were solved by building the system on industry-scale open source technologies. These are MediaWiki, Django and MongoDB. Although these individual components are known to work reliably and scale well, there is a future problem of maintainability. This rises from the concern of the compatibility of our system with the future versions of MediaWiki and Django. Even during the two years we have been developing the system, a critical part of the MediaWiki API has already changed once. This required updates to our code in order to make our system work with the latest version of MediaWiki. This maintainability issue is solved by releasing the entire system as open source.

Currently, other users of the shared MediaWiki platform maintained by the Language Bank of Finland are showing interest in our system. Not only because it provides an already implemented way of

pushing dictionary data from another format to the MediaWiki system, but also because our system makes it possible to transfer the data edited in MediaWiki back to the original format.

## 7 Conclusion

In this paper, we have described our online dictionary system<sup>6</sup> with the aim of making XML based dictionaries editable by multiple users. We have described the advantages and limitations of Sematic MediaWiki in enhancing access to the dictionary data. Furthermore, the advantages of MediaWiki have been described. Our system is currently in use and has been proved to solve the problems we were set to solve with a small number of editors.

The dictionary system was originally developed for Skolt Sami, but we have successfully expanded it to cover 10 additional languages with minimal modifications. This has been possible due to the modular nature and ideology of separation of concerns embraced in the design process.

In addition to solving a dictionary editing problem, our efforts have made the XML formatted dictionaries available to a wider audience in an open MediaWiki format. The availability of these lexical resources online has a direct impact on the speakers and learners of these minority languages. The data has also been made available for research and technical purposes through the API of the system.

## References

- Älgu-tietokanta. (2002). Retrieved March 2018, from Kotimaisten kielten keskus: <http://kaino.kotus.fi/algu/>
- Antonsen, L., Johnson, R., Trosterud, T. & Uiho, H. (2014). Generating Modular Grammar Exercises with Finite-State Transducers. *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013*, (pp. 27-38).
- Bergenholtz, H. & Johnsen, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. *HERMES-Journal of Language and Communication in Business*, 34, 117-141.
- Boicea, A., Radulescu, F. & Agapin, L. I. (2012). MongoDB vs Oracle - database comparison. *Third International Conference on Emerging Intelligent Data and Web Technologies* (pp. 330-335). IEEE.
- Bothma, T. J., Gouws, R. H. & Prinsloo, D. J. (2017). The Role of E-lexicography in the Confirmation of Lexicography as an Independent and Multidisciplinary Field. *Proceedings of the XVII EURALEX International Congress*, (pp. 109-116).
- Hämäläinen, M. (2018, January). UralicNLP (Version v1.0). *Zenodo*. <http://doi.org/10.5281/zenodo.1143638>.
- Johannsson, E. T. & Battista, S. (2017). Editing and presenting complex source material in an online dictionary: the Case of ONP. *Proceedings of the XVII EURALEX International Congress*, 117-128.
- Laxström, N. & Kanner, A. (2015). Multilingual Semantic MediaWiki for Finno-Ugric dictionaries. *Septentrio Conference Series*, 2, pp. 75-86.
- Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T. & Silfverberg, M. (2013). HFST — A System for Creating NLP Tool. *International Workshop on Systems and Frameworks for Computational Morphology*, (pp. 53-71).
- Männamaa, K. & Iva, S. (2015). Võro-eesti-võro võrgosõnaraamat: synaq.org. In M. Velsker, & T. Iva, *Tartu Ülikooli Lõuna-Eesti keele- ja kultuuriuuringute keskuse aastraamat* (p. 147–150). Tartu: Tartu Ülikooli Kirjastus.
- Nicola, M. & John, J. (2003). XML Parsing: A Threat to Database Performance. *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 175-178). ACM.
- Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, 38(1), 1-4.

<sup>6</sup> The system has been released as open source in <https://bitbucket.org/mikahama/saame/>

- Rueter, J. & Hämäläinen, M. (2017). Synchronized Mediawiki Based Analyzer Dictionary Development. *The Third International Workshop on Computational Linguistics for Uralic Languages*, (pp. 1-7).
- Selva, T. & Verlinde, S. (2002). L'utilisation d'un dictionnaire électronique: une étude de cas. *Proceedings of the tenth EURALEX International Congress*, (pp. 773-781).
- Töpel, A. (2014). Review of research into the Use of Electronic Dictionaries. In C. Müller-Spitzer, *Using online dictionaries* (pp. 13-54). Berlin - New York: De Gruyter.
- Tender, T., Kallas, J., Laansalu, T., Nurk, T., Mihkla, M., Päll, P., Langemets, M., Soon, T. & Oro, K. (2017). *Eesti Keele Instituudi osakondade aruanded 2017*. Tallinn: Eesti Keele Instituut.
- Trosterud, T., Moshagen, S. & Pirinen, T. (2013). Building an open-source development infrastructure for language technology projects. *NEALT Proceedings Series*, 16, pp. 343-352.
- Völkel, M., Krötzsch, M., Vrandečić, D., Haller, H. & Studer, R. (2006). Semantic Wikipedia. *Proceedings of the 15th international conference on World Wide Web (WWW '06)* (pp. 585-594). ACM.
- Viikberg, J. (2008). Eesti keele kogud. In E. Parmasto, & J. Viikberg, *Eesti humanitaar- ja loodusteaduslikud kogud, seisund, kasutamine, andmebaasid* (pp. 95-112). Tartu: Tartu Ülikooli Kirjastus.

# Linking Corpus Data to an Excerpt-based Historical Dictionary

**Tarrin Wills, Ellert Þór Jóhannsson, Simonetta Battista**

*University of Copenhagen*

*E-mail: tarrin@hum.ku.dk, nk950@hum.ku.dk, sb@hum.ku*

## Abstract

*A Dictionary of Old Norse Prose (ONP)* is a digital dictionary that derives originally from an excerpt-based index of around 750,000 citations. This paper describes recent attempts to create two-way links between the growing body of digital texts encoded using TEI XML and the dictionary's word list, which forms the basis of the published dictionary. The process involves design challenges in bringing together very different digital structures, namely the text in an XML tree structure, and the dictionary in a relational database structure. Because of the very high levels of accuracy demanded by the end-users of the dictionary (particularly researchers in Old Norse studies), the linking process can only be automated for unambiguous cases, with remaining links entered manually. The application and interface that assists this process attempts to minimize the trade-off between automation and accuracy, and adds a range of tools to assist with the human lemmatizing process. We were able to achieve linking of lemmas in 90.4% of instances where the lemma was recorded in the TEI text, with very high levels of accuracy. Where no lemma was recorded, the application allowed an Old Norse scholar to link lemmas to previously unlemmatized words at an average rate of 4-7 seconds per word.

**Keywords:** online dictionary, corpus material, historical lexicography, Old Norse

## 1 Background

Texts written in Old Norse-Icelandic (henceforward Old Norse) form a major source for the study of literature, history and culture of Viking and Medieval Scandinavia. The material consists mainly of prose narratives (sagas) as well as legal, historical and learned texts, and charters. All of these are preserved in medieval and early modern manuscripts. A great emphasis is placed in the field of Old Norse studies on the material evidence for the texts as the foundation of the discipline, namely, the manuscripts from Norway and Iceland, particularly the latter.

The lexicography of Old Norse provides an important tool for understanding the history, literature and culture preserved in the texts. A reliable dictionary provides a means for locating lexical indicators of sociocultural phenomena and understanding the subtleties and variations of their use. The “gold standard” of lexicography in Old Norse should provide a link not only between the lexicon and the corpus but also to its material record. Researchers in the field using lexicographic resources expect very high levels of accuracy (at least 99.9%) and coverage (all instances of all low-frequency words, for example).

*A Dictionary of Old Norse Prose (ONP)* is a dictionary project hosted at the University of Copenhagen and part of the Arnarnagæan Institute of Old Norse Manuscript Studies. The dictionary has a long history, first as an unpublished archive of dictionary citations, but later as an incomplete print edition where four out of planned 12 volumes were published. Currently *ONP* is available as an online resource at *ONP.ku.dk*. The online *ONP* brings together unedited dictionary material, material from the printed volumes, as well as more recently edited dictionary entries. The work on the



dictionary continues with regular entries published online, as well as addition of new features to the online version (for a detailed overview cf. Johannsson & Battista 2014). The fundamental approach of the *ONP* is to elucidate the original medieval material while maintaining rigorous philological standards. The process involves strict textual principles and attention to orthographic details. This insistence on textual integrity has set *ONP* apart from earlier lexicographical works on Old Norse/Icelandic. *ONP*'s corpus derives from reliable diplomatic editions that are based in turn on direct readings of the manuscripts. Each citation is linked to a published edition as well as to the manuscript which represents its primary witness. The dictionary is edited and stored as a relational database, with linked tables representing the headwords, definitions, citations, editions and manuscripts (cf. Johannsson & Battista 2016).

In recent years the publication of digital scholarly editions of Old Norse texts has increased significantly. Many of these texts follow the standards set by the Medieval Nordic Text Archive (Menota) in its published handbook (Haugen 2008), which in turn is a minor extension of the TEI XML standard, particularly the element set designed for representing primary materials. Menota “aims to preserve and publish medieval texts in digital form and to adapt and develop encoding standards necessary for this work” (<http://www.menota.org>). Menota has made a large number of recent scholarly texts editions publicly available as encoded xml-files, amounting to a corpus of around 1.6 million words, most of which are within the scope of *ONP*'s coverage, and all of which are closely based on readings of the original manuscripts of the works, the “gold standard” for *ONP*'s corpus. Unlike *ONP*'s traditional excerpt-based corpus, these texts provide a potential direct link between the lexicon and the manuscript page, without an intermediate edition.

A third project, Lexicon Poeticum (LP), provides the structure and interface that allows the two others to come together. LP is effectively a poetic supplement to *ONP* which is based on the digital edition of the majority of the poetic corpus made by the Skaldic Project (<http://skaldic.abdn.ac.uk>; Clunies Ross et al. 2007-2017). This project uses a relational database for its edition, but the textual structure was developed from a TEI model and can therefore incorporate TEI-encoded texts (Wills 2013). In addition, the data on manuscripts and prose words in the Skaldic database is based originally on *ONP*'s data, and LP uses *ONP*'s wordlist as the basis for its own. The web application developed for the Skaldic Project and LP to enter and edit data is highly extensible and forms the framework for the application described below, where the Menota texts are first incorporated into the database structure and then linked to the *ONP* wordlist.

The desired outcome was that each project should benefit: Menota gains a means for automatically linking its lemmas to an authoritative external dictionary, as well as an application for assisting the manual linking process, all of which can be exported back as Menota-compatible XML; LP incorporates the remaining section of the corpus not covered by the Skaldic Project, namely the Codex Regius collection of poetry which has recently been edited according to the Menota guidelines; and *ONP* gains comprehensive coverage of selected manuscripts, greatly adding to its corpus, and in a format that can easily be added in the future to its own database.

Automatic lemmatizing systems of Old Norse tend to be directed towards full morphosyntactic analysis with lemma rather than as a lexicographic tool with unambiguous links between the corpus and dictionary. These automatic systems vary in accuracy, with recent published methods varying from 84% (Urban *et al.* 2014; 96% for word class) up to 92.7% accuracy for full morphosyntactic analysis (Rögnvaldsson & Helgadóttir 2011). These methods are unsuitable for a historical dictionary such as *ONP*: they use a highly normalized corpus compared with the manuscript-based text required by *ONP*; they do not provide a reliable way of linking accurately from the generated lemma to a curated wordlist; and the levels of accuracy, although gradually improving, are a very long way off what is required by the users of the dictionary, who demand close to 100% accuracy and coverage.

## 2 Method

The availability of an extensive number of digital texts that meet the textual standards of the *ONP* dictionary means that they can potentially be compared with the fragmentary digital corpus already in *ONP*. This requires aligning the corpora in a way that they can be linked and analyzed, which presents some challenges given the different nature of the data structures (relational data — linked tables; and XML — a tree structure). The two structures and points of connection between them are shown in Figure 1. *ONP*'s database encompasses a comprehensive index of all texts within its scope, and all manuscripts consulted in the editions of those texts. The Menota texts are based on transcriptions of individual manuscripts, divided into the texts which are recorded within them. TEI allows for a very complex textual structure with nested divisions, paragraphs, sentences or stanzas and lines, whereas such nesting is more difficult to encode in a relational data structure. The following process effectively “flattens” the XML structure into a series of words, but the two-way linking between the XML and database means that the details of the structure can be recovered at a later date if the database and application are updated to accommodate it.

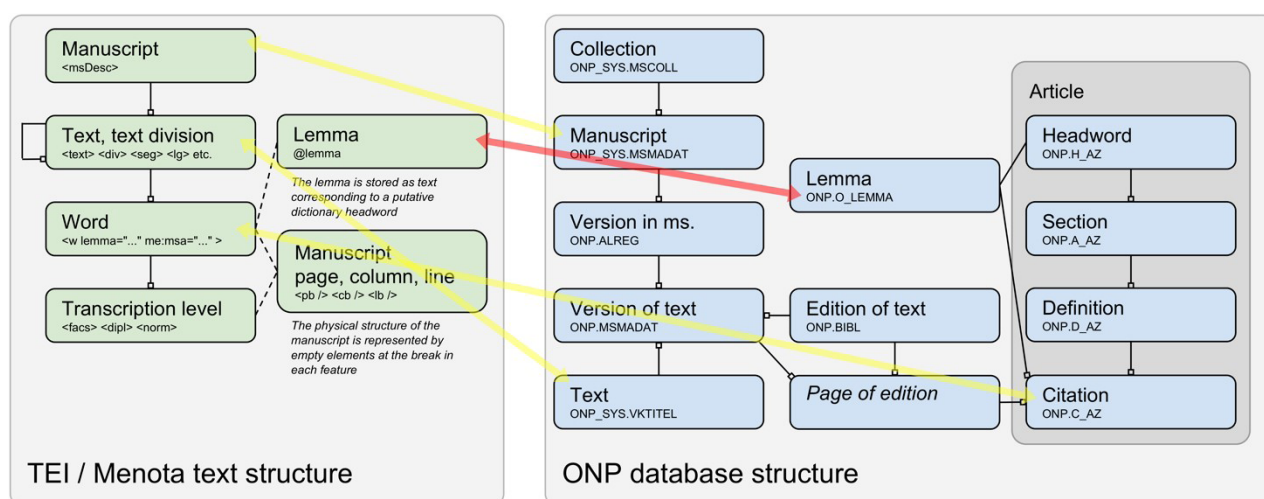


Figure 1: Menota XML structure and *ONP* database structure.

The Menota specification requires texts to be encoded with all words tagged with the `<w>` (word) element. This provides a functionally equivalent structure to the citation table in *ONP*, that is, an instance of a word in a text. Menota also advocates using the ‘lemma’ attribute of the `<w>` element to provide a dictionary headword for each word, providing a potential means for automating the linking process. Texts that use this attribute also encode further information in the ‘me:msa’ (Menota: morphosyntactic analysis) attribute, including word class, inflectional class and morphological categories.

### 2.1 Importing Menota XML

The Menota XML file is imported in two stages using the web application (not shown). In the first stage, the file is either uploaded or fetched via URL from the Menota catalogue and linked to both a manuscript and a text already in the database. The application reads through the file and detects all page, column and line breaks (marked up with standard TEI tags) and uses them to construct a unique identifier for all word (`<w>`) elements in the file, so that each word has a record of its location in the manuscript. The resulting data remains valid TEI XML and is stored on the server. In the second stage, the XML data is parsed and all word and punctuation elements (`<w>` and `<me:punc>` tags)

extracted using an XPath query. The three “levels” of textual representation defined by Menota are parsed, along with the lemma attribute and morphosyntactic tokens recorded in the me:msa attribute.

The result is that the database words table is populated with the following information for each word:

- A link to the text and to the manuscript in the corresponding tables in the database
- The xml:id value for linking back to xml file
- The raw me:msa attribute string (if present) and the raw lemma attribute string (if present)
- Up to three forms of the word (“facsimile”, “diplomatic” and “normalized” levels, if present)
- A number representing the position of the word in the text
- The manuscript page/folio and line number where the word begins (and column number, if relevant)
- Grammatical information from the parsed me:msa attribute: word class, strong/weak, gender, number, case, definiteness, degree of adjectives, person, tense, mood, voice, finiteness, suffix
- Punctuation for each form of the word, parsed from the following <me:punc> tag (if present)
- A link to the lemma table is undefined at this stage

This provides sufficient information for reconstructing the text, extracting the lexical context for each word, locating a word in the manuscript, and potentially automatically lemmatizing the word if sufficient information is present.

## 2.2 Automatic Lemmatizing

A large number of the texts in the Menota catalogue record a lemma string and morphosyntactic analysis as XML attribute values for each word. Together, the lemma value and word class provide a way of identifying and linking the word to an *ONP* dictionary headword (the ‘Lemma’ data types Figure 1). There is potential for ambiguity in instances where there are homographs with the same word class. Examples of this include a number of high-frequency words including the verbs *verða* (four headwords with 1,380 citations in *ONP*), *mæla* (three headwords with 948 citations) and *fá* (two headwords and 1,027 citations). In these examples one headword accounts for the overwhelming majority of instances of the word: between 96% (*mæla*) and 99.4% (*verða*) of citations, and a larger percentage of actual instances, as the citations for high-frequency words in *ONP* are not exhaustive. Researchers using the dictionary, however, require highly accurate coverage of low-frequency homographs. For example, the low-frequency homograph <sup>2</sup>*fá* (meaning to color, paint; 12 citations in *ONP*) is found in very early poetry and runic inscriptions and may be an indicator of an early date of a text. Finding all instances of this rare homograph may therefore be of interest to a historical linguist.

The following procedure is initiated by the database to avoid potentially incorrect linking of words to lemmas.

A temporary reference table is built using an SQL statement from the most recent *ONP* wordlist imported into the database, using the headword form, word class and noun gender to identify all unique homographs for each class/gender. This process means that all potentially ambiguous homographs are ignored, e.g. the verbs *fá*, *verða* and *mæla* will not be used, because there are multiple homographs with the same word class. Despite the fact that linking to the most frequent homograph would be at least 96% accurate in these cases, it is important that all such cases are checked manually in order to capture all instances of the low-frequency words.

The database attempts to link the reference lemma table to the word table, using the lemma, word class and gender values based on what was originally the lemma and me:msa attributes in the XML file. Where there is a match, a link is inserted to the lemma table for each matching word in the word

table. This process is initiated through the web interface and the whole process takes one or two minutes. An optional second pass attempts to match Old Norwegian variants in the lemma attributes of the remaining words and generally captures another 5% of words in Old Norwegian texts. Overall, the process captures around 90% of all words with high levels of accuracy (see Section 3 below).

The advantage of this method over one that captures more words but with lower accuracy is that the captured words do not need to be manually checked, at least in the initial stages. The remaining words comprise the ambiguous homographs mentioned above, lemmas which are not fully covered by *ONP*'s wordlist (particularly proper nouns and poetic words), or have been lemmatized in the XML according to differing practices from *ONP*'s process of determining headwords. These need to be lemmatized with human intervention, and the next stage involves using the application to assist in this process.

### 2.3 Assisted lemmatizing

At this stage the database contains a table of words in the manuscript text that can be used to reconstruct the text in various ways. The web application uses this table to create a web form with various features to assist in the lemmatizing of the words in the text which have not been automatically lemmatized, either because the Menota file lacked this information, or the lemmatizing process described in the previous section did not link the word to the lemma table.

The web form is shown in Figure 2 and consists of three columns, the first of which gives information about the word. In order to lemmatize the text the words must be understood in their syntactic context, and in practice it is fairly easy for the person using the form to follow the text as they scroll down the form. Where this is not possible, a pop-up can be opened which shows the word in the surrounding text, including any grammatical and lemma information that may have been previously entered.

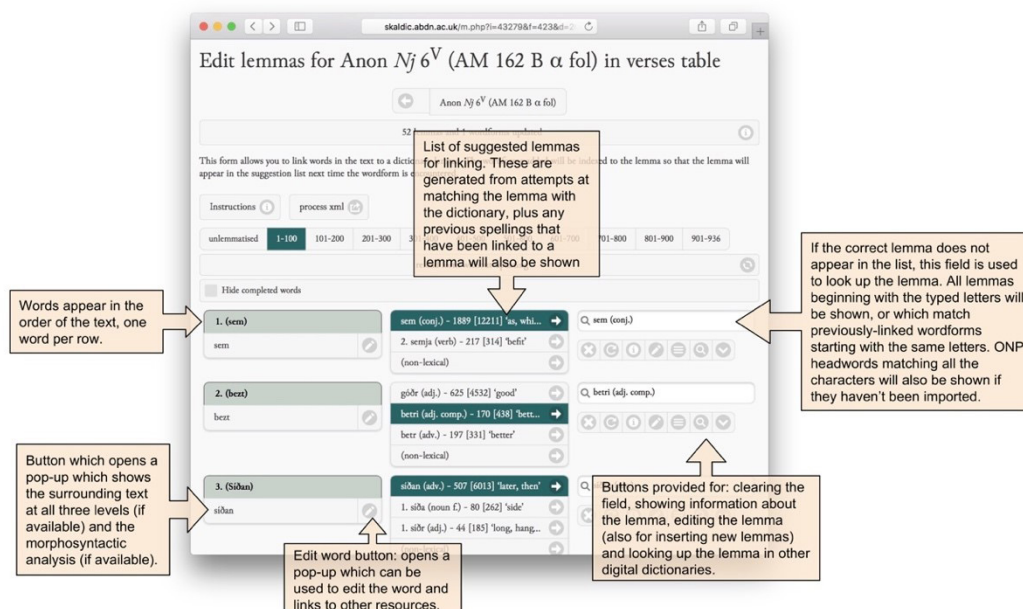


Figure 2: Assisted lemmatizer form, with explanations.

The second column is a list of potential lemmas for the word. This list is generated by searching the full corpus in the database for words which are already lemmatized and match the present form in the text. The database includes around 120,000 words from the Skaldic Project, 700,000 from *ONP*'s citation list, 1,500,000 from Málfrög's automatically-lemmatized corpus (cf. Rögnvaldsson



and Helgadóttir 2011), as well as the words already lemmatized using this process from the Menota catalogue (200,000 words at the time of writing). The resulting matching lemmas are listed according to overall frequency. In around 90% of cases this search of the 2,500,000-word corpus produces the correct lemma as one of the options in this list.

If this process does not find the correct lemma for the word, the word list can be manually searched in the third column. The results are listed in the same format as for the automatically-searched list, which includes word class, grammatical information and a gloss/definition, if available. To further assist in this process, a number of buttons provide pop-ups with further information once a potential lemma is selected, including a form for editing the lemma information itself and adding a new lemma, as well as looking up the lemma in various electronic dictionaries.

Because of the highly repetitive nature of this work there is a risk of repetitive strain injuries, particularly from using a computer mouse. The form is therefore designed to be used on a range of devices including desktop computers, tablets and phones. It is also being reviewed for optimizing for key-board input.

Up to 100 words at a time are shown in the form and are updated when the form is submitted.

## 2.4 Output

The two-way linking between the XML word elements and database word table allows the inputted information to be inserted back into the XML and exported. For each word (<w>) element in the TEI XML, a “me:ref” attribute is inserted with references to the external resources in URI format (the attribute is a Menota extension). The application inserts a reference to the database key for the word in the words table (e.g. ‘menota:word:ends:3593306’ for the word *laugberkf* on AM 162 B α fol, 1v/1) and the numeric identifier of the *ONP* headword (e.g. ‘menota:lemma:ONP:51275’ for *logberg*). These references can be resolved as URLs using the application API.

The screenshot shows a web application interface for the Menota project. The browser address bar displays `skaldic.abdn.ac.uk/m.php?p=menotalemma&i=49334`. The main heading is **langr (adj.)** [°*compar.* lengri, *superl.* lengstr] ‘long...’.

A note states: "Please note that the lexical concordance has not been reviewed and should not be referenced."

**Word forms:** langa, langar, langer, lango, langrar, langre, langri, langt, langu, laungv, lengra, *Lengra*, longu, longum, longum, longv, længra, længri, længst, længster, længster.

**ONP:** langr, "lagnt", [] *adj.*

**Concordance**

Filter items...

dyrin oc mælti. <i>Lengra</i> mundir þu rena i áve	Lbs fragm 82: 1va/15
Slag-brandar gorvir af <i>longum</i> room ok þungum	AM 1056 IX 4*: 1va/3
sua læiðar oc <i>langar</i> vesallder ef ek ma heðan	DG 4-7: 20va/26
sægi yðr Miok <i>longu</i> saker hæmsku minnar amællta ei	DG 4-7: 23b/32
en nu er <i>langt</i> liðit siðan ec for	DG 4-7: 25b/4
dom hafðe guð <i>longu</i> aðr dæmt oc upp sagt	DG 4-7: 30a/37
nu er miok <i>langt</i> siðan menn hava sua	DG 4-7: 28ra/34
æinn saman miok <i>langa</i> stund <i>konongrenn</i> sialfr læiddi	DG 4-7: 35b/17
fra þoum ækki <i>lengra</i> ævande. En þesttar ævðu lið	DG 4-7: 28ra/8

**Grammatical forms in corpus**

(Only includes words that have been morphosyntactically tagged)

		masc.	fem.	neut.
<b>strong</b>	<b>sg.</b>	<b>nom.</b>		langt (2)
		<b>acc.</b>	langa (6)	langt (9)
		<b>dat.</b>	langri (2)	longu (6)
			langre (1)	langu (2)
			lango (2)	longv (1)
	<b>gen.</b>		langrar (1)	
	<b>pl.</b>	<b>nom.</b>	langar (1)	
		<b>acc.</b>	langar (4)	
		<b>dat.</b>		
	<b>gen.</b>			
<b>weak</b>	<b>sg.</b>	<b>nom.</b>		
		<b>acc.</b>		
		<b>dat.</b>		

Figure 3: Concordance and morphological forms.



If a lemma has not already been inserted, the application will populate the “lemma” attribute with the headword form from *ONP*’s wordlist. Morphosyntactic analysis can also be inputted using the assisted lemmatizing form and will be exported in Menota’s format if entered. If no analysis is entered the “me:msa” will be populated with the information that can be extracted from the word list, that is, word class and gender in the case of nouns.

The application also inserts a revision description showing which changes were made by the application and how many changes were the result of particular users. The resulting file is Menota conformant and can be imported back to the Menota Catalogue with the additional data included.

The processed and inputted data remain in the database and can be used in various ways. For example, the text can be viewed with parallel transcription levels shown in columns. Each word is linked between the three forms (shown with highlighting) and clicking/tapping on a word shows a popup with information including a link to the lemma, the word in context and grammatical information if present. The lemma is linked to further information, and the wordlist can be searched for individual lemmas. Figure 3 shows the resulting information about each lemma: the grammatical form and gloss (deriving from *ONP*); a list of word forms from the corpus; a full concordance of the word in the corpus with surrounding text, including references to the individual manuscript; and where morphosyntactic analysis has been entered, a paradigm of the lemma and its morphological forms can be reconstructed.

Additional views show the full concordance for an individual text, the text on an individual manuscript page with parallel manuscript image, and further views for the text.

### 3 Results

Manual updates using the lemmatizing form are logged in a separate table and each word updated this way includes a link to the editor who performed the action. We can thus extract information from the log about how long each manual operation took and how many words were captured by the automatic lemmatization process.

Table 1 shows three longer Menota XML texts which were fully or partially lemmatized and processed using the automatic lemma linking procedure outlined in 2.1 above. The total words value only includes words that were lemmatized in the XML (the *Konungs skuggsjá* text has a total of 63,895 words).

Table 1: Capture of automatic lemmatizing.

Menota text	Linked lemmas	Total words	Percent
Strengleikar in DG 4-7	34788	38453	90.5%
<i>Konungs skuggsjá</i> in AM 243 b α fol.	37299	39537	94.3%
Barlaams saga ok Jósafats in Holm perg 6 fol.	67545	76411	88.4%
Total	139632	154401	90.4%

A random sample of 1,000 words with surrounding text was manually checked by the authors against the *ONP* wordlist for accuracy; all (100%) were found to be correctly linked to the *ONP* lemmas based on the lemma and word class information in the Menota XML files. Two were incorrectly lemmatized in the original XML, and a very small number, although technically correct, were linked to different headwords from *ONP* due to minor differences in word class classification between *ONP* and in the original XML file. The accuracy of this process is therefore solely dependent on the accuracy of the

lemmatization in the imported XML, and does not appear to introduce further errors. This means that no further systematic checking is required for the words linked by this method.

All updates using the web application are logged separately in the database and this log can be used later to analyze the processes initiated through the interface. The time taken for manual lemma linking can be calculated where there are multiple updates recorded in the database log within a limited period of time, disregarding what are clearly longer breaks between periods of lemmatizing (defined as longer than one hour). The time difference between each logged update, and the number of words altered in the relevant updates can together be used to calculate the average time per word taken to lemmatize the text. The results of this analysis are shown in Table 2.

Table 2: Time spent on manual lemmatizing.

Menota text	Total time (h:min)	Words lemmatized	Time per word
<i>Njáls saga</i> in AM 162 B α fol.	1:12	585	7s
<i>Njáls saga</i> in AM 162 B θ fol.	2:52	1515	7s
<i>Njáls saga</i> in AM 162 B κ fol.	0:33	457	4s
Strengleikar in DG 4-7	5:20	2227	9s
Total	9:57	4784	7s

The majority of words can be linked faster than the 7s average shown in the table, but a small minority require closer investigation and in some cases the addition of lemmas to the database. These cases slow down the overall rate of lemmatizing but are necessary for the high levels of accuracy which is the aim of the system. It should also be noted that the lemmatizing process for *Strengleikar* in DG 4-7 was slightly slower than for the other texts: this is due to the fact that this text was automatically lemmatized in the first instance and only the remaining 9.5% of words were manually linked. It is faster per word to lemmatize an entire text: the user mentally parses the full sentences rather than looking at isolated words. The overall speed, however, is much faster, when 90% can be linked automatically without need for further checking.

These results are based on a single user (Wills), and the results and accuracy for other users will be investigated further.

## 4 Conclusion

Existing Menota texts with lemmas recorded as XML attribute values can be automatically linked to *ONP*'s headwords with high levels of capture (around 90%) and very high levels of accuracy (> 99.9%), meaning that further checking is not required. This provides a basis for supplementing the dictionary's articles and citations with direct links to manuscript text and images (see <http://skaldic.abdn.ac.uk/m.php?p=ONP>: 'other resources' tab for lemmas).

The web application provides a very fast way for a human user to lemmatize words that have not been lemmatized in XML or not linked by the automatic system, with words in unlemmatized texts able to be processed by a user familiar with Old Norse in about 7 seconds each. This means that very large TEI texts can be processed in a comparatively short time, adding greatly to the texts linked to *ONP*. The process also provides a much quicker way than has previously been achieved for inserting lemma values into the TEI text for use in the Menota archive.

The resulting data can be used by both the lexicographic and editing projects, and the different formats (XML and database) remain linked by the use of unique identifiers for each word. The data from

this project, including the linked texts and concordance, are available at <http://skaldic.abdn.ac.uk/m.php?p=menota>.

## References

- Haugen, O.E. (2008). *The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources*. Version 2.0. Bergen: Medieval Nordic Text Archive. <[http://www.menota.org/HB2\\_index.xml](http://www.menota.org/HB2_index.xml)>
- Johannsson, Ellert Thor & Simonetta Battista (2014). "A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition", in Andrea Abel & al. (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15-19 July 2014, Bolzano/Bozen, 169-179.
- Johannsson, Ellert Thor & Simonetta Battista (2016). "Editing and Presenting Complex Source Material in an Online Dictionary: The Case of *ONP*", in Tinatin Margalitadze & Georg Meladze. (eds.): *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, 6-10 September 2016, Tbilisi, 117-128.
- ONP* = Degnbol, H., Jacobsen, B.C., Knirk, J.E., Rode, E., Sanders, C. & Helgadóttir, Þ. (eds.). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose*. *ONP* Registre (1989). *ONP* 1: a-bam (1994). *ONP* 2: ban-da (2000). *ONP* 3: de-em (2004). Copenhagen: Den Arnamagnæanske Kommission.
- ONP* Online. *Ordbog over det norrøne prosasprog* Online. Accessed at: <http://ONP.ku.dk> (20/03/2018)
- Urban K., Tangherlini T.R., Vijūnas A., Broadwell P.M. (2014). Semi-Supervised Morphosyntactic Classification of Old Icelandic. In *PLOS ONE*, 9(7), e102366. <<https://doi.org/10.1371/journal.pone.0102366>>
- Rögnvaldsson, E. and Helgadóttir, S.. 2011. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In C. Sporleder et al. (eds.) *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pp. 63-76. Berlin: Springer.
- Wills, T. (2015-). *Lexicon Poeticum*. Accessed at: <http://lexicONPoeticum.org> [20/03/2018]
- Wills, T. (2015). Relational data modelling of textual corpora: The Skaldic Project and its extensions. In *Literary and Linguistic Computing* 30(2), pp. 294–313. <<https://doi.org/10.1093/lc/fqt045>>



# Collocations Dictionary of Modern Slovene

**Iztok Kosem<sup>1,2</sup>, Simon Krek<sup>2</sup>, Polona Gantar<sup>1</sup>, Špela Arhar Holdt<sup>1</sup>, Jaka Čibej<sup>1,2,3</sup>, Cyprian Laskowski<sup>1</sup>**

<sup>1</sup>Faculty of Arts, University of Ljubljana, <sup>2</sup>Jožef Stefan Institute, <sup>3</sup>Faculty of Computer and Information Science, University of Ljubljana

E-mail: iztok.kosem@ff.uni-lj.si, simon.krek@guest.arnes.si, apolonija.gantar@guest.arnes.si, Spela.ArharHoldt@ff.uni-lj.si, jaka.cibej@ff.uni-lj.si, CyprianAdam.Laskowski@ff.uni-lj.si

## Abstract

The paper presents the compilation of the Collocations Dictionary of Modern Slovene, a new resource targeting the language production needs of Slovene speakers. An important aspect of the compilation of the dictionary is the immediate publication of all the entries, from automatic, postprocessed, finalized by lexicographers and so on, and indicating to the users their status, i.e. the stage in the compilation process. Furthermore, we discuss the introduction of crowdsourcing into the lexicographic workflow. The paper also focuses the development and presentation of the interface, which introduces new approaches to collocation presentation. The aim was to develop a collocation-driven interface that would allow different types of users a great deal of flexibility and customizability in exploring collocational information about words. In this way, the interface represents a hybrid between a more corpus-based presentation of collocations (e.g. in tools such as Word Sketch) and a traditional sense-driven presentation of collocations as found in existing collocations dictionaries.

**Keywords:** collocations, dictionary, database, interface

## 1 Background

In recent years, collocations have received a great deal of attention in Slovenian lexicography, initially mainly in relation to the conceptualization of a new monolingual dictionary of modern Slovene (Gorjanc et al. 2015, 2017). Relatedly, procedures for the automatic extraction of collocations and their examples have been developed and continuously improved (e.g. Kosem et al. 2013; Gantar et al. 2016). This means that lexicographers can now very quickly obtain large quantities of collocational information about words, which has facilitated the compilation of dictionaries, both general and terminological (e.g. Logar et al. 2013). However, despite these methodological advances, existing Slovene dictionaries, many of which are also outdated, offer users little help with language production tasks such as writing.

Interestingly, despite the advances in methods for collocation identification and the advantages offered by digital media, not many born-digital collocations dictionaries (i.e. dictionaries developed with a digital medium in mind) have been published. The examples the authors are familiar with include the Estonian Collocations Dictionary (Kallas et al. 2015; the dictionary will be published in 2018), the German Collocations Dictionary (Roth 2013; Häcki Buhofer et al. 2014),<sup>1</sup> and the Spanish Collocations Dictionary (DiCE; Vincze et al. 2011; Vincze & Alonso Ramos 2013). Furthermore, projects such as automatic collocation dictionaries (see Kilgarriff et al. 2013) and SkeLL have shown that even automatically extracted data can be useful for language users. All of the aforementioned dictionaries target L2 learners, as collocations specifically tend to pose significant problems for language learners (e.g. Granger & Meunier 2008; Schmitt 2004; Nation 2001); however, even L1 users

<sup>1</sup> <https://kollokationenwoerterbuch.ch/web/>. The dictionary was also published in paper format.



encounter challenges in language production and can benefit from having such resources at their disposal.

The lack of productively-oriented dictionaries for Slovene, regardless of the types of users, prompted the compilation of the Collocations Dictionary of Modern Slovene (CODICT, Gantar et al. 2015). A great deal of attention was paid to the design and customizability of the interface. One of the important decisions was not to publish only completed entries, but also include entries in various stages of completion, mainly in order to avoid causing user frustration due to the lack of collocational data for those headwords without complete entries in CODICT. As we wanted to inform design-related decisions with empirical data as much as possible, we first conducted a small test on a sample database to get feedback on the presentation of collocational information, and to investigate how Slovene users react to automatically created (not manually curated) content.

## 2 Initial Test with a Sample Collocations Database and Its Interface

In 2016, a sample of 2,500 automatically extracted collocational entries for Slovene (Krek et al. 2016) was extracted and published at <http://bkssj.cjvt.si/>. Each entry included collocations grouped by grammatical relation and their corpus examples (extracted using GDEX; Kilgariff et al. 2008). Some additional post-processing was conducted, e.g. putting the collocates in the required case, gender, removing duplicate examples, etc.

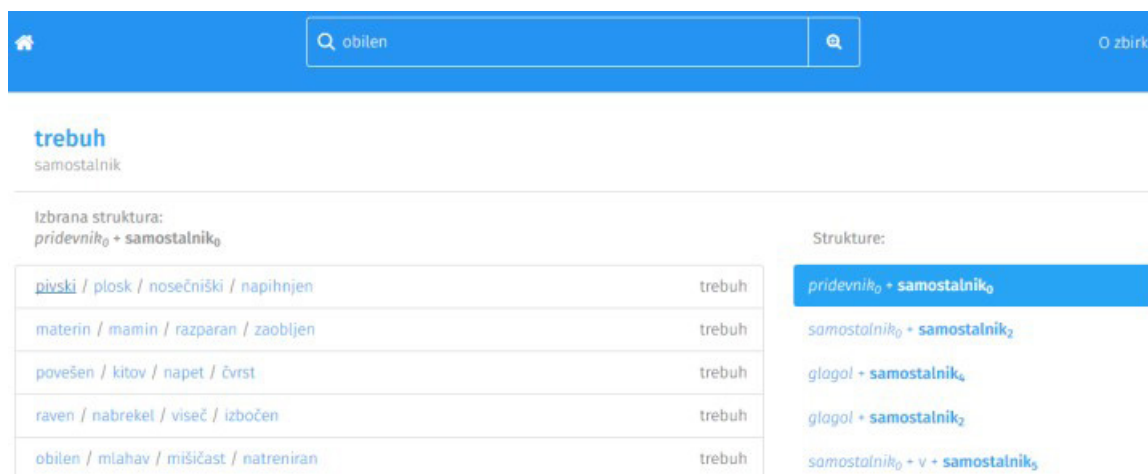


Figure 1: The interface of the sample collocations database.

The interface and the contents of this sample collocations database were then evaluated by a group of linguists and linguistics students. The data was found to be useful despite the fact that it contained a certain degree of noise. The users commented on several issues, which can be grouped into two categories. The first is related to content, such as large quantities of data, lack of data structure (e.g. missing sense information), small number of entries, distracting irrelevant or incorrect information (i.e. noise). The second category involves the presentation of lexical information, for example the lack of (statistical) information on collocations to help assess their relevance, the lack of options for sorting collocations, and so on. In sum, while the automatically extracted information was well-received among the users, they also noted a need for more structure being given to the data, and more options with regard to manipulating it.

### 3 Collocations Database and Collocations Dictionary of Modern Slovene

After concluding the initial test, we first extracted a much larger dataset, containing 35,989 entries with automatically extracted collocational data (nearly eight million collocations and nearly 37 million corpus examples), using the Sketch Engine API. The data was obtained from Gigafida, the 1.2-billion-word reference corpus of written Slovene (Logar Berginc et al. 2012). Compared to the sample database, several improvements have been introduced, both to the data extraction procedure and the post-processing of the extracted data. For example, good examples, five per collocation, were extracted using an updated GDEX configuration (e.g. penalties for sentences ending with an ellipsis and containing only upper-case letters have been introduced). Secondly, additional filtering of collocational data was used in the post-processing stage; this excluded all collocations containing the verb *biti* ('to be') and removed prepositional grammatical relations (preposition + noun in a specific case) that were not in accordance with the rules of the Slovene Orthography (Toporišič 2001).<sup>2</sup> This dataset presented a basis for CODICT.

Nowadays, dictionaries mainly use two approaches in the way they publish entries. One is to wait until the entire dictionary is compiled, and the other one, which has become the norm for online dictionaries, is to publish newly compiled entries at regular intervals (e.g. once a year) – this is what Klosa (2013) calls a dictionary under construction. For our purposes, none of these approaches seemed suitable, so we decided to use the approach proposed by Krek et al. (2013), where all the entries are made available to the users immediately, with a clear indication of their status in the lexicographic process. We introduced five stages: automatically extracted entry, postprocessed entry (semi-automatically cleaned and improved data based on lexicographic decisions), entry with validated collocations, entry with collocations distributed under senses, and final entries. There are several reasons for this, such as improvements in (semi-)automatic methods of processing of language data, especially collocations, the way that users can benefit from all this information immediately, the fact that in many cases clustered collocations already clearly indicate different senses, and so on. In addition, the dictionary will benefit from the results and findings of a research project called KOLOS (Collocations in Slovene), which is focused on collocations and aims to improve methods of collocation detection, collocation clustering, and the use of collocations in comparing synonyms.

The ultimate aim is to have all the entries manually edited, i.e. validated and cleaned data obtained with automatic extraction, with added information such as sense division, labels, collocate groupings (clusters), as well as providing collocations in their typical form (e.g. nouns in plural, adjectives in superlative, verbs in negative). For each collocation, at least two good examples should be provided. Rather than using definitions for senses, we decided to use short indicators, similar to signposts (used first by the Longman Dictionary of Contemporary English) and short definitions in menus (introduced by the Macmillan English Dictionary for Advanced Learners) (see also Kosem et al. 2017). The most important role of the indicators is thus not to explain the meaning, but to help the users clearly distinguish between different senses.

It is important to note that not all collocations that are validated by lexicographers are included in the final entries. This is because of the difference between statistical collocation, i.e. any combination of two or more words that is statistically relevant, and a collocation that is deemed relevant for inclusion in a collocations dictionary. Our inclusion criteria for collocations (and headwords) are less strict than the criteria used to make works such as the *Macmillan Collocations Dictionary*, where the authors excluded headwords such as *house*, *buy*, *good* on account of the fact they do not have any strong

2 Some of these excluded prepositional grammatical relations may contain valid information, but we will conduct a detailed analysis before including them in the database.

collocates.<sup>3</sup> In the digital age there is no longer any need to limit the number of headwords, but there remains a need to determine which collocations to include; for example, do we include numerous modifiers of the word *prestolnica* ('capital') related to countries, e.g. *Austrian*, *German*, etc., or do we include only the most salient ones, or none at all? In any case, even when statistically relevant collocations are excluded from CODICT, they are still kept in the database as they will be of use for the compilation of other resources, e.g. general language dictionaries, valency dictionaries, etc.

### 3.1 Implementing Crowdsourcing into Lexicographic Workflow

Considering the high number of collocations per headword and financial and time constraints related to the compilation of CODICT, we decided to introduce crowdsourcing methods into the lexicographic workflow. Previous tests (Kosem et al. 2013) have shown that tasks such as assigning collocations, via examples, to the relevant senses are not very demanding (even for non-linguists) and provide highly reliable results. We conducted such a task on 6,590 collocations (microtasks) for 88 sample headwords in CODICT, using four annotators (students of linguistics) and requiring three answers per microtask (see Figure 2). In addition to senses, the annotators were given the answer options "None of the above" (if the example indicated a sense of the headword not covered by the ones provided) and "I don't know". The results showed a high degree of annotator agreement: each pair of annotators agreed in 79-86% of the cases (83% on average, with an average Cohen's kappa of 0.83). A total of 4,258 collocation examples (65%) showed perfect agreement.

**obiskati bazar**

Ogledali si bomo mesto in *obiskali bazar*.

orientalska tržnica    prireditev

Nič od naštetega    Ne vem

---

Trenutno rešujete nalogo **12**.

Rešili ste naslednje število nalog: **0** od skupno **1**

Figure 2: The crowdsourcing task in Pybossa (an example of a microtask).

In addition to saving lexicographers' time and consequently speeding up the compilation of entries, the crowdsourcing method provides important feedback on sense division even during the process of dictionary compilation. As the experience from our crowdsourcing task has shown, the analysis of annotators' responses can reveal which indicators need to be improved, and potentially identify groups of collocates that might require their own sense, or senses that might have been overlooked. Furthermore, there were instances where the annotators' responses indicated that the sense division was too fine-grained.

3 <http://www.macmillandictionaries.com/features/how-dictionaries-are-written/macmillan-collocations-dictionary/>

### 3.2 Designing the Interface

A great deal of attention has been paid to the development of the interface,<sup>4</sup> which has to be easy to use and makes it clear to the users the status of the entry they are consulting. The status is indicated both directly, with the use of a pyramid icon with completed stages colored in red, and indirectly, for example the Sense filter is introduced only in the last two stages of entry compilation. The design has been greatly informed by user feedback on the interface of the aforementioned sample collocations database. An important decision made in the design process was that the interface would be collocation-driven rather than sense-driven; as a result, the user would be initially given a more general overview of the collocations of the word, and would then be able to explore collocational information further using sorting options and filters (by sense, grammatical relation, frequency, etc.). The interface was designed for different devices, i.e. a computer, tablet, and smartphone; some adjustments, e.g. exclusion of certain functions, had to be implemented for smaller screens such, as phones.

The initial view (i.e. general overview) provides a quick summary of most relevant collocations, divided by grammatical relations. We therefore attempt to maintain some grammatical diversity of the collocational overview; normally, there is one line per grammatical relation, although multiple lines can be allocated to a single relation if the number of (salient) collocations it contains is significantly higher than the number of collocations found in other relations. In this initial view, the users can select a specific grammatical relation to get a view of all the collocations in it, or they can click on a particular collocation to see its examples of use.

The screenshot displays the CODICT interface for the word 'jezik'. The top navigation bar is red with the 'cjvt kolokacije 1.0' logo, a search bar, and social media icons. Below the bar, the word 'jezik' is highlighted, and a date '2018-06-10' is shown. The main content area is divided into a left sidebar and a central table. The sidebar contains filters for 'Relevantnost', 'Gruče', and 'A-Ž', a frequency slider 'Pogostost', and a list of grammatical relations under 'Pomeni' (e.g., 'del telesa', 'jed', 'način izražanja') and 'Strukture' (e.g., 's samostalniki', 'z glagoli'). The central table lists collocations in four columns: 'materni jezik', 'lepljiv', 'goveji', 'dolg', 'odpor do jezika', 'odnos do', 'strast do', 'ljubezen do', 'govoriti jezik', 'stegniti', 'obvladati', 'iztegniti', 'jezik EU', 'narodnosti', 'države', 'manjšine', 'zakon o jeziku', 'znanost o', 'vedenje o', 'znanje o', 'čut za jezik', 'posluh za', 'talent za', and 'občutek za'. Each row has a vertical ellipsis icon on the right.

Figure 3: The interface of CODICT (first page of the entry *jezik*).

<sup>4</sup> <http://viri.cjvt.si/kolokacije/>.

The left-hand panel (located at the top in the mobile version) contains sorting and filtering options. Sorting is available only for the single relation view, and enables the user to sort collocates by relevance (default setting), semantic characteristics (clustering), and alphabetical order. There are four types of filters available:

- The frequency filter enables the users to filter collocates according to their frequency in the corpus. The filter can thus help the users to focus on a more frequent or rare collocates.
- *Pomeni* – the sense filter (available for entries in the final two stages) provides an overview of the senses in which the collocations can be found.
- *Struktura* – the grammatical relation filter provides the users with the option to limit the results to relations according to the word class of the collocate (e.g. noun, adjective, verb, adverb), and subcategories such as case or degree (e.g. superlative).
- *Predlogi* – the preposition filter applies to prepositional relations (trinary) in which collocations can be found, and is offered as a separate filter because it transcends different top-level categories in the grammatical relation filter.

As the idea is to give the users flexibility in limiting the amount of data displayed on the screen in order to facilitate finding the relevant information, multiple types of filters, as well as sorting, can be active at the same time. However, only one category within a certain type of filter, e.g. prepositions in the *Predlogi* filter, can be selected by the users at a time. There is a difference in behavior of the sense filter, where all the senses of a word always remain visible, even if some of the sense do not apply to the selection (those are then greyed out), and the grammatical relation and preposition filters, where only the selected or relevant relation (and its subcategory) is shown.

Even if the users are not using filters and are conducting their activities only in the right-hand of the interface (main panel), the filters remain dynamic and in that way informative, as they provide information on the current selection in the main panel. For example, if the user selects one of the grammatical relations in the main panel, they are taken to the collocations of that relation, and at the same time, the senses in which the collocations belonging to that particular relation are not found are greyed out, and grammatical relation and its subcategory in the grammatical relation filter is selected. This solution has been formed based on users' feedback to the sample collocations database interface, as they found listing all available grammatical relations on the right side overwhelming, and the names of the relations confusing (e.g. verb + noun<sub>4</sub> meant verb followed by a noun in the accusative case).

There are additional filters available in the right-hand panel (in the main view), which are based on the characteristics of the headword. For example, the collocations of the adjective headword in the grammatical relation adjective + noun can be filtered by gender. So, as shown in Figure 4, selecting *okusna* (the feminine form of the adjective *okusen*, 'tasty') shows only feminine nouns as collocates (*sladica* 'dessert', *hrana* 'food', etc.).

A very important feature of the interface is its search functionality, which offers searches by both headwords and collocations. It is important to note that if the user searches for a specific collocation, the result differs from the output obtained if opening the same collocation within a particular entry. This is due to the difference in user focus – if a specific collocation is selected within a certain headword, the user's focus comes from the headword, whereas when a specific collocation is searched for, the information on all its elements, and related collocations, is useful.

Finally, the interface also provides links to other resources. A page of each individual collocation, which contains corpus examples, contains a direct link to the corpus concordances for the collocation. Furthermore, the main features of the interface are shared by all the resources of the Centre for Language Resources and Technologies at the University of Ljubljana, so the user can, by clicking



on a specific button in the search row, obtain a quick view of links to all the available resources that contain entries related to his or her search.



Figure 4: An additional filter in the main collocations view (single relation).

## 4 Future Plans

Future plans for this project involve making several improvements to the methodology of compilation of collocational data. This involves improving the precision of collocation detection using approaches such as distributional semantics and extraction of collocations from parsed corpora. In addition, we plan to update CODICT with the information from a new version of the Gigafida corpus, which is due to be published at the end of 2018. We will also explore gamification forms of crowdsourcing to identify valid collocations.

Equally important as improving the methodology is testing the interface with different user groups. At the time of writing this paper we are already preparing a survey that will be conducted among the users, and will be complemented with interviews. The study will focus mainly on the content of the initial view, i.e. what different users would expect or want to be offered when opening the entry. Other plans include adding new filters based on the metadata of corpus texts, e.g. text type and year of publication.

## References

- Häcki Buhofer, A., Dräger, M., Meier, S. & Roth, T. (2014). *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.
- Gantar, P., Gorjanc, V., Kosem, I. & Krek, S. (2015). Going semi-automatic and crowdsourced: collocation dictionary of Slovene. In I. Kosem (ed.) *Electronic lexicography in the 21st century: linking lexical data in the digital*

- age. *eLex 2015, book of abstracts*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing, 2015, p. 37.
- Gantar, P., Kosem, I., & Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29 (2): 200–225.
- Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2015). *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.
- Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2017). *Dictionary of Modern Slovene: problems and solutions*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.
- Granger, S., & Meunier, F. (eds.) (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam: Benjamins.
- Kallas, J., Kilgariff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In I. Kosem, M. Jakubiček, J. Kallas & S. (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 1-20.
- Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds.) *Dictionaries. An international Encyclopedia of Lexicography*. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin in Boston: de Gruyter, pp. 517–524.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris. (eds) *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425–432.
- Kilgariff, A., Husak, M., Jakubiček, M. (2013). Automatic collocation dictionaries. *Presentation at eLex 2013 conference, Tallinn, Estonia*. Available at: <https://youtu.be/b3KyhPBeoLU>.
- Kosem, I., Gantar, P., Krek, S. (2013): Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, pp. 32-48.
- Kosem, I., Gantar, P., Krek, S. (2017). Sense menus in collocations dictionary of Slovene. *Electronic lexicography in the 21st century: lexicography from scratch*. Leiden: Dutch Language Institut; Brno: Lexical Computing; Ljubljana: Trojina Institute for Applied Slovene Studies, p. 43.
- Krek, S., Gantar, P., Kosem, I., Gorjanc, V. & Laskowski, C. (2016). Baza kolokacijskega slovarja slovenskega jezika. In T. Erjavec & D. Fišer (eds.) *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija = Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th - October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. Ljubljana: Znanstvena založba Filozofske fakultete: = Ljubljana University Press, Faculty of Arts, pp. 101-105.
- Krek, S., Kosem, I., Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika, version 1.1*. Accessed on 10 May 2018. [http://www.sssj.si/datoteke/Predlog\\_SSSJ\\_v1.1.pdf](http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf)
- Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Roth, T. (2013). Going Online with a German Collocations Dictionary. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 152-163.
- Rundell, M. (ed.) (2010). *Macmillan Collocations Dictionary*. Oxford, United Kingdom: Macmillan Education.
- Schmitt, N. (ed.) (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: Benjamins.
- Toporišič, J. (ed.) (2001). *Slovenski pravopis*. Ljubljana: Založba ZRC, ZRC SAZU.
- Vincze, O., Mosqueira, E., & Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In I. Boguslavsky & L. Wanner (eds.) *Proceedings of the 5th International Conference on Meaning-Text Theory*. Barcelona, pp. 275–286.

Vincze, O. & Alonso Ramos, M. (2013). Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Español. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, pp. 328-337.

## Acknowledgements

The paper was prepared as part of the two projects, *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (Collocations as a Basis for Language Description: Semantic and Temporal Perspectives, J6-8255) and *Nova slovnica sodobne standardne slovenščine: viri in metode* (New grammar of contemporary standard Slovene: sources and methods, J6-8256), which were financially supported by the Slovenian Research Agency. The paper was also supported by ELEXIS, a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 731015.

The authors would like to acknowledge the financial support from the Slovenian Research Agency's infrastructure programmes Centre for Applied Linguistics, Trojina Institute (I0-0051), and Centre for Language Resources and Technologies, the University of Ljubljana.

The interface was developed by Studio Kruh in collaboration with Leon Noe Jovan.



# Computerized Dynamic Assessment of Dictionary Use Ability

**Osamu Matsumoto**

*Waseda University*

*E-mail: matsumoto.kbx@gmail.com*

## Abstract

This paper demonstrates a new web-based dictionary training resource, named the Computerized Dynamic Assessment of Dictionary use Ability (C-DADA) intended for Japanese learners of English. C-DADA is a computerized format of dynamic assessment (DA), which originated in the works of Vygotsky, and is developed for lexicographic purposes. The fundamental aspect of DA is the integration of assessment and instruction. In addition, DA distinguishes two levels of ability: actual and potential. The former is the level at which the individual can perform by himself, while the latter is the level at which he can solve a problem with others' assistance. Likewise, C-DADA is designed to assess the learners' actual and potential levels of receptive dictionary use ability through providing feedback according to individual learners' responsivity, and so further promote their dictionary use ability. The paper first introduces the theoretical and methodological framework of DA, and then discusses the design and mechanics of C-DADA. Lastly, it gives a general overview of an on-going lexicographic project to examine the effectiveness of C-DADA.

**Keywords:** dictionary use, dictionary training, dynamic assessment, computerized dynamic assessment

## 1 Introduction

The need for dictionary training has long been advocated by many, such as lexicographers, applied linguists, curriculum designers and language teachers. To meet the demand, the author has developed a web-based dictionary training resource, named Computerized Dynamic Assessment of Dictionary use Ability (C-DADA). To better understand C-DADA, its theoretical and methodological framework will first be presented.

### 1.1 Dynamic Assessment (DA)

Dynamic assessment (DA) is an assessment method based on the sociocultural theory of mind, originating with Vygotsky, a Russian developmental psychologist. The fundamental idea of this theory is that social interaction is responsible for the development of higher mental functioning (Vygotsky 1978). DA is especially grounded in the concept of the zone of proximal development (ZPD). ZPD is "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky 1978, p. 86). It is also the place where learning will take place. The purpose of DA is to determine how much learning can take place in the ZPD during the task or test with or without others' assistance or mediation, while at the same time promoting this development as effectively as possible. Unlike traditional psychometric tests, in which there is no direct interaction between an assessor and learners during administration, in DA the assessor attempts to share the problems with the learners, to find out their actual level of performance on the task and to enhance their level of performance to overcome the problems they face (Reza & Barabadi 2012). In addition, it is possible for the assessor to track a learner's development during and across sessions. The assessor in DA is not a complete observer who maintains distance from the learner, but



rather a facilitator and mediator of learning. In this regard, the DA assessor and the learners are both active participants. The DA assessor interacts with the learners, helps them learn by giving feedback, clues, or mediational prompts as needed, and thus they collaboratively achieve the task. In this way, assessment and learning are seen inextricably mingled and not as separate processes (Reza & Barabadi 2012).

## **1.2 Computerized Dynamic Assessment**

DA first appeared in the field of developmental psychology to examine the learning potential of people or the developmental stages of children's cognition (e.g., Sternberg 2014; Sternberg & Grigorenko 2002). In the past two decades, it has been applied to educational contexts as a promising approach to build more effective learning environments. However, it has not been widely used in classroom settings. This is possibly because the administration of DA tends to be time consuming. In DA the assessor is required to pay careful attention to changes regarding the types of problems learners encounter, provide the appropriate assistance they require to overcome the problems with adjustment in their ZPD, and observe their responses to such assistance. In this respect, the generalizability of outcomes of DA is said to be very difficult. Other problems are related to a lack of adequate knowledge base and expertise in this field (Reza & Barabadi 2012), which may be more serious problems with regard to bringing DA into wider use. However, researchers have found a practical solution to address these matters by using state-of-the-art technology, by adapting DA into a computerized format, called Computerized-Dynamic Assessment (C-DA). Indeed, over the last decade, applications of C-DA have been increasing in foreign language education (e.g. Ebadi & Saeedian 2016; Poehner & Lantolf 2013).

## **2 Computerized Dynamic Assessment of Dictionary Use Ability (C-DADA)**

The Computerized Dynamic Assessment of Dictionary use Ability (C-DADA) is designed to blend assessment, instruction, and training of receptive dictionary use in one holistic activity. To do this, C-DADA gives the task to the learners with providing feedback to help solve a problem that they face while they are working on. The type, level and frequency of feedback are further used for numerically evaluating their dictionary use ability.

### **2.1 Task and Feedback on C-DADA**

To the best of the author's knowledge, DA has not been discussed in the lexicographic literature to date. Accordingly, C-DA, a derivative of DA, has not been developed and studied either. However, C-DA is technically applicable and very beneficial, especially in pedagogical lexicography, in which L2 learners' lack of dictionary use skills and the necessity of dictionary instruction have long been noted (e.g., Chan 2012; Nesi & Hail 2002; Wingate 2004). If C-DA for dictionary use is available, it may enable teachers to manage to instruct and evaluate learners' dictionary use skills without needing much time for both. In addition, like other e-learning resources, this type of C-DA use may establish an environment where learners study dictionary use by themselves outside the classrooms. It may further enable teachers to easily track and record their achievement and progress in the digital format. However, the most significant aspect is that the application of C-DA may maximize the learning potential of dictionary use rather than its practicality, as DA does. Based on these assumptions, the author presents a newly developed C-DA as a lexicographic resource, named the Computerized Dynamic Assessment of Dictionary use Ability (C-DADA).

Following the principles of DA, C-DADA is designed to unify assessing and instructing dictionary use skills in a single task. More specifically, C-DADA assesses how learners' actual and potential

levels of receptive dictionary use ability while learning could be promoted as much as possible. It should be noted, however, that because the target user of C-DADA is a learner of English at a beginning or lower intermediate level, the task on C-DADA focuses on some core skills necessary to search the contextual meaning of a word in the dictionary rather than comprehensive skills for receptive use of a dictionary. The next is the steps that a learner takes in C-DADA.

First, the learner is shown a short sentence which includes a bold, italicized word as the targeted word, the meaning of which is to be searched. Then the three blanks to be filled appear in the following order:

- (1) Form
- (2) Syntactic category
- (3) Meaning

In the form section, the learner needs to identify whether the target word is inflected. If so, he has to change it to the appropriate word form as a headword. Next the learner is required to decide on the syntactic category in context and select the most appropriate syntactic category from the dropdown list, including eight syntactic categories. As for verb, the distinction between transitive and intransitive is necessary because conventional English-Japanese dictionaries usually treat each as an independent subentry. Finally, after syntactic categorization, the learner fixes the contextual meaning. The learner needs to select the most suitable meaning from the dropdown list with six to ten meanings.

There are four levels of feedback given to the learner according to the correctness of the answer and the level of the preceding feedback for each section. The amount of feedback C-DADA provides per section is from one to four items. When a correct answer is provided by the learner, C-DADA gives feedback which verifies its correctness and allows him to go to the next step. In contrast, for a wrong answer, different types of feedback are given, which are structured and graduated from implicit to explicit. This way of feedback is an instance of locating and targeting learners' ZPD in order to be able to provide maximally beneficial assistance (e.g. Poehner, 2009; Rassaei, 2014). The first feedback is the most implicit that simply gives the second chance to answer. Likewise, the second feedback aims to direct the learner's attention to the part(s) relevant to solve the problem given in each section. The third feedback presents more explicit or direct clues or hints to help the learner to answer by adding more information to the second one. The fourth feedback is the last and most explicit, which presents the answer with an explanation. This feedback is given when the learner fails to answer correctly after receiving the third feedback. Except for the items, language used in C-DADA is basically the learner's L1 (i.e. Japanese), due to the reality of their dictionary use. This is because the target users of C-DADA are at a beginning or lower-intermediate level of L2 English, so that it is very likely that they may not understand the language of feedback in L2. It is also common knowledge that bilingual dictionaries have always enjoyed greater popularity with foreign language learners, even those at an advanced level (Dick-Bursztyn 2014).

Table 1 summarizes the types and levels of feedback given for each section. Note that comments in quotations in the table are translated into English for the convenience of the reader of this paper.

To understand the mechanics of C-DADA more clearly, showing the following example should be helpful. Suppose there are two students (student A and student B) who are individually working on the following item on C-DADA.

He *succeeded* to his father's business.

First student A is expected to identify whether the italicized target word '*succeeded*' is a base form (lemma) working as a headword in the dictionary. When he recognizes it as a base form and types '*succeeded*' into the space provided for the form section, C-DADA will give the first feedback

Table 1: The types and levels of feedback.

Feedback	Preceding Answer	Function	Section		
			Form	Syntactic Category	Meaning
1st	Wrong	To give the second chance to answer	To present “That’s wrong! Try again!”	To present “That’s wrong! Try again!”	To present “That’s wrong! Try again!”
	Correct	To lead the learner to the next step	To present “That’s correct!” and lead the learner to the syntactic category section	To present “That’s correct!” and lead the learner to the meaning section	To present “That’s correct!” and lead the learner to the next item
2nd	Wrong	To direct the learner’s attention to the surrounding context of the target word	To point out some part(s) of the target word or some word(s) relevant to identify the base form	To point out some word(s) relevant to identify the syntactic category of the target word	To point out some word(s) relevant to identify the meaning of the target word
	Correct	To lead the learner to the next step	To present “That’s correct!” and lead the learner to the syntactic category section	To present “That’s correct!” and lead the learner to the meaning section	To present “That’s correct!” and lead the learner to the next item
3rd	Wrong	To give more direct clues / hints than 2nd feedback	To explain morpho-syntactic context of the target word more directly than 2nd FB	To exhibit syntactic relationship between the word(s) pointed out in the 2nd FB and the target word	To give L1 translation of the surrounding words or partial translation of the sentence except the target word
	Correct	To lead the learner to the next step	To present “That’s correct!” and lead the learner to the syntactic category section	To present “That’s correct!” and lead the learner to the meaning section	To present “That’s correct!” and lead the learner to the next item
4th	Wrong	To present the answer with explanation	To present the correct base form of the target word with explanation	To present the correct syntactic category of the target word with explanation	To present the correct meaning of the target word with explanation
		To lead the learner to the next step	To lead the learner to the syntactic category section	To lead the learner to the meaning section	To lead the learner to the next item

Note: FB = feedback

“That’s wrong! Try again” (in Japanese). Then he gets rid of the ‘-ed’ inflected morpheme, and types ‘succeed’ into the same place, the second feedback is the given to verify this answer’s correctness by presenting “That’s correct!”, and displays the direction to go to the next section. Simultaneously the

dropdown list for the syntactic category appears. The student is then required to select one syntactic category from the list. It should be noted here that transitive and intransitive verbs are listed as an independent syntactic category. This means that if he first selects transitive verb for the above case C-DADA will give a comment “That’s wrong! Try again.” as the first feedback. Then if he continuously fails to select the correct syntactic category, the second feedback is presented. In the above case, because ‘succeeded’ is followed by a prepositional phrase, C-DADA attempts to direct student A’s attention to it by commenting “There is ‘to his father’s business’ after ‘succeeded’”. However, if he still cannot select the proper syntactic category, the third feedback is given. This time C-DADA gives a more explicit clue by presenting “There is ‘to his father’s business’ as a prepositional phrase and no noun phrase working as the object of ‘succeed’”. Then, when he selects intransitive verb from the list, C-DADA verifies it as the correct answer and gives the direction to go to the meaning section. Immediately, the dropdown list for the meaning comes out. For a wrong selection in the meaning section the first feedback gives the student another chance to answer, just as in the other two sections. In the second feedback, C-DADA points out some surrounding words of the target word which are expected to be helpful to guess its contextual meaning. If the student fails three successive times, C-DADA gives L1 translation of the words which are presented in the second feedback. However, if student A still not select the correct meaning in context from the list, C-DADA displays the answer with an explanation and allows him to move to the next item. Next the case of student B is shown.

Student B is working on the same item as student A. Like student A, student B’s first trial in the form section results in failure. After receiving the first feedback, she types ‘succeed’. Soon after verifying it as the correct answer, C-DADA allows student B to go forward to the syntactic category section. Again, her first trial results in failure. When the first feedback given, she successfully chooses the answer intransitive verb from the list and is led to the meaning section. However, here she has more troubles than the previous two sections. She repeatedly fails in finding the appropriate meaning in context. She needs the third feedback before her fourth answer is verified as the correct one.

This example thus illustrates how C-DADA works for the two students’ different answer patterns. Figure 1 shows a sample image of C-DADA.

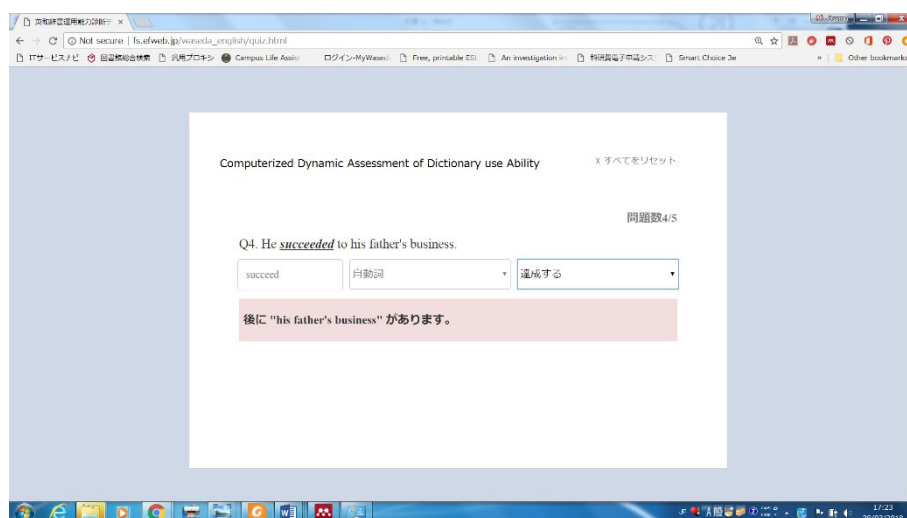


Figure 1: Sample image of C-DADA.

As in Figure 1, C-DADA presents items to be solved one by one. The learner must reach the meaning section and successfully answer regardless of the level of feedback or receive the fourth item of feedback due to making four successive mistakes before being able to see the next item. The learner

cannot go back, either. After the learner finishes the 50th item, the score report with task completion time will be automatically generated.

## 2.2 Scoring

In C-DADA the scoring system is based on Poehner and Lantolf (2013). There are two different scores per item: actual and mediated. The actual item score is a non-assisted score, the result of the first trial for each section. The point per section will be either 0 or 1 point, so that the score per item is the sum of the points of the three sections, which can range from 0 to 3. The mediated item score, in contrast, is an assisted score, the result of the learner's answer with feedback. If the learner can correctly answer for any section without feedback, she will receive the highest score (i.e. 1 point). However, with the increase in the number of items of feedback, the score per section will fall from 1.0 to 0 by 0.25 points. The points for the three sections are summed up so that the mediated score per item will range 0 to 3 by 0.25 points. Each item score is then summed up so as to produce two different types of scores for the entire task performance: actual and mediated. Both scores will range from 0 to 150 points, since C-DADA contains 50 items and first set can get 1 point each and the second 0.25. Furthermore, the points for the three sections (form, syntactic category and meaning) are individually summed up, which gives the three section scores. Like the item and task scores, each section score distinguishes the actual score (non-assisted) and the mediated (assisted) score (see Appendix 1 for the flow of the task and scoring procedure per item).

To understand the scoring process more clearly, the previous two students' cases (A and B) are illustrated. Student A solves a problem in the second trial in the form section, in the fourth trial in the syntactic category section, but cannot do so in the fourth trial in the meaning section. The points that he receives for the three sections will thus be 0.75, 0.25, and 0, respectively. Then these are summed up, resulting in 1.0 points as the mediated item score. Simultaneously C-DADA will present 0 points as the actual item score, because student A does not successfully answer in any section in the first trial. In contrast, student B succeeds in the second trial in the form section, in the second trial in the syntactic category section and in the fourth trial in the meaning section. C-DADA will thus give her 0.75, 0.75, and 0.25 for the respective sections, which makes the mediated item score 1.75 points. At the same time C-DADA will presents 0 points as the actual item score, because student B does not solve a problem in the first trial in any sections, just like student A. In sum, student A and B's actual item scores are the same (0) while their mediated item scores are different: 1.0 and 1.75, respectively. From this, student B receives more gains from feedback than student A. This may suggest that student B is a potentially better dictionary user than student A, while they are superficially at the same level for this item in that both of them receive 0 points as the actual item score. In this way, C-DADA aims to numerically distinguish the learner's actual and potential level of dictionary use ability.

## 2.3 Target Word

There are 50 items presented in C-DADA. Each item consists of a short sentence with a bold, italicized target word. They are in either uninflected or inflected forms. All of the items are constructed based on the following criteria: First, the target items are polysemous words that have more than six meanings in the entry or subentry, and the majority of which have multiple syntactic categories. This may prevent the learners from selecting the meaning in context with ease. Second, non-target words, words used in the sentences, are carefully chosen to be possibly familiar to students. In fact, most of them are usually taught in junior high school. While some non-target words may be unknown to students, their meanings are supposed to be easy to guess from context. Third, parts of idioms or phrasal expressions are not set as target words. Although the ability to find them in the dictionary is undoubtedly important, it is outside the focus of this study. Fourth, the sentence structure of each



item is designed to be simple to support student understanding. For example, the items do not include relative clauses, participle clauses, or subjunctive mood, which are reported to be difficult grammatical items or structures for Japanese learners of English to understand (Chujo, Yokota, Hasegawa & Nishigaki 2012).

### 3 Future Research

While developing C-DADA is itself a lexicographic project, the future research goes one step further. The effectiveness of C-DADA will be examined in two different ways.

The first study will take the format of pretest-instruction-posttest: C-DADA is introduced as the instruction phase. The pre- and posttests have the same type of the task on C-DADA: to search the contextual meaning of the target word in a dictionary. After taking the pretest, the learners work on C-DADA. Then they take the posttest (retest). The results of the pretest and posttest will be compared to statistically examine the effects of C-DADA on their development of dictionary use ability.

In the second study, the learners' on-going development while working on C-DADA will be investigated. As already mentioned, assessing and learning are unifiable in DA; DA constructs an environment where learning occurs with the help of the assessor or others. Accordingly, it is expected that learners may show on-going development of dictionary use ability by receiving feedback from C-DADA. To find this, the individuals' task performances on C-DADA will be examined in detail. For example, if a learner solves a problem with less feedback at each section per item as he proceeds to the last item (i.e., the 50th item), he is assumed to have developed his dictionary use ability while working on C-DADA. Furthermore, if there are learners with the same actual task scores, their differences in potential dictionary use ability will be revealed by examining their mediated task scores and the amount and type of feedback they received to solve the problems.

### 4 Conclusion

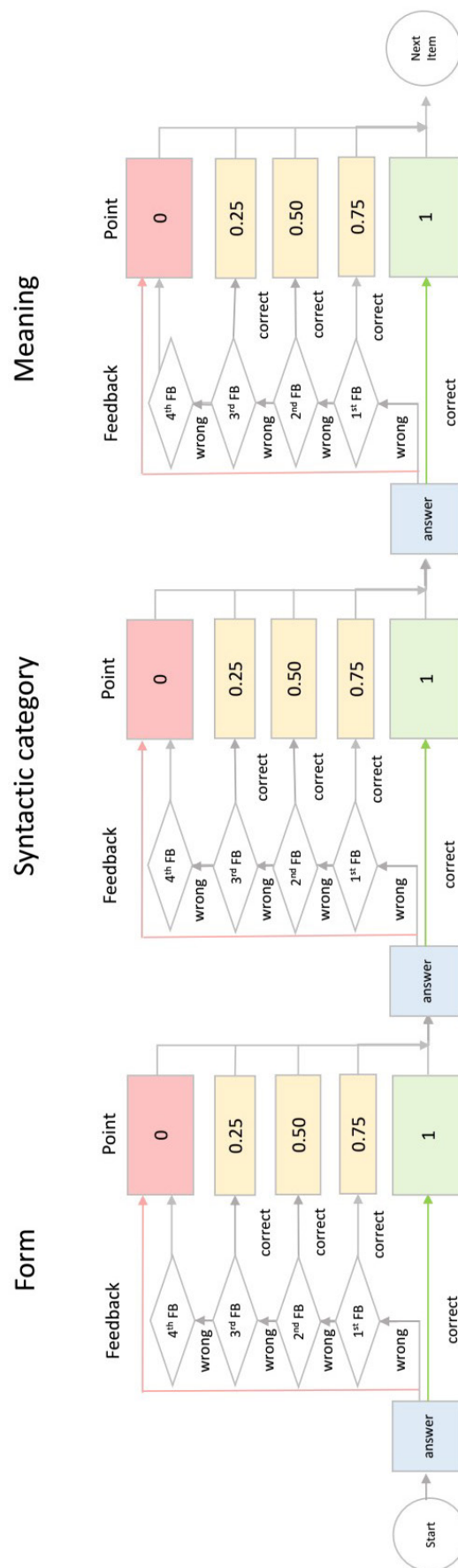
This paper has presented C-DADA, a newly developed web-based learning resource. As seen so far, C-DADA has some unique features. First, C-DADA is an e-learning resource for lexicographic purposes. Second, C-DADA is a theory-driven approach, applying the principles of dynamic assessment which originated in Vygotsky's sociocultural theory of mind. Unlike the traditional assessment, C-DADA unifies instruction and assessment as a single unit. C-DADA assesses individuals' actual and potential levels of dictionary use ability while attempting to promote their learning. This will provide very valuable information on learners' ability of dictionary use to assessors. Furthermore, if the on-going project shows the effectiveness of C-DADA, it will be applicable to other language learning contexts as well. It is expected that C-DADA will make a contribution to language learning and the progress of lexicographic research.

### References

- Chan, A. Y. W. (2012). The use of a monolingual dictionary for meaning determination by advanced Cantonese ESL learners in Hong Kong. in *Applied Linguistics*, 33(2), pp. 115–140
- Chujo, K., Yokota, K., Hasegawa, S., and Nishigaki, C. (2012). Identifying the general English proficiency and distinct grammar proficiency of remedial learners. In *Journal of the College of Industrial Technology, Nihon University (Nihon Daigaku Seisan Kougakubu Kenkyuu Houkoku)*, 45, pp. 43–54.

- Dick-Bursztyn, M. (2014). Facing grammar problems with the aid of lexicographic tools. In *Journal of Language and Cultural Education*, 2(2), pp. 281–290.
- Ebadi, S., & Saeedian, A. (2016). Planning future instructional programs through computerized L2 dynamic assessment. In *Teaching English with Technology*, 16(4), pp. 12–32.
- Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. In *International Journal of Lexicography*, 15(4), pp. 277–305.
- Poehner, M. E. (2009). Group dynamic assessment: mediation for the L2 classroom. In *TESOL Quarterly*, 43(3), pp. 471–491.
- Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: capturing L2 development during computerized dynamic assessment (C-DA). In *Language Teaching Research*, 17(3), pp. 323–342.
- Rassaei, E. (2014). Scaffolded feedback, recasts, and L2 development: a sociocultural perspective. In *Modern Language Journal*, 98(1), pp. 417–431.
- Reza, P., & Barabadi, E. (2012). Constructing and validating computerized dynamic assessment of L2 reading comprehension. In *Iranian Journal of Applied Linguistics*, 15(1), pp. 73–95.
- Sternberg, R. J. (2014). The development of adaptive competence: why cultural psychology is necessary and not just nice. In *Developmental Review*, 34(3), pp. 208–224.
- Sternberg, R. J., & Grigorenko, E. L. (2002). Difference scores in the identification of children with learning disabilities it's time to use a different method. In *Journal of School Psychology*, 40(1), pp. 65–83.
- Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.
- Wingate, U. (2004). Dictionary use - the need to teach strategies. In *Language Learning Journal*, 29(Summer), pp. 5–11.

## Appendix



Appendix 1: Flow of the task and scoring procedure per item in C-DADA.



# Creating a List of Headwords for a Lexical Resource of Spoken German

*Meike Meliss, Christine Möhrs, Dolores Batinić, Rainer Perkuhn*

*Institut für Deutsche Sprache, Mannheim*

*E-mail: meliss@ids-mannheim.de, moehrs@ids-mannheim.de, batinic@ids-mannheim.de, perkuhn@ids-mannheim.de*

## Abstract

Except for some recent advances in spoken language lexicography (cf. Verdonik & Sepesy Maučec 2017, Hansen & Hansen 2012, Siepmann 2015), traditional lexicographic work is mainly oriented towards the written language. In this paper, we describe a method we used to identify relevant headword candidates for a lexicographic resource for spoken language that is currently being developed at the Institute for the German Language (IDS, Mannheim). We describe the challenges of the headword selection for a dictionary of spoken language, and having made considerations regarding our headword concept, we present the corpus-based procedures that we used in order to facilitate the headword selection. After presenting the results regarding the selection of one-word lemmas, we discuss the opportunities and limitations of our approach.

**Keywords:** list of headwords, spoken German, corpus-based methods

## 1 Introduction

In the project “Lexik des gesprochenen Deutsch” (=LeGeDe)<sup>1</sup> a corpus-based lexical resource for standard spoken German in interaction is being developed. In this resource, lexical and interaction-specific features are to be gathered and presented in a multimedia manner (cf. Meliss & Möhrs 2017; Möhrs, Meliss & Batinić 2017). Identifying and defining the characteristics of spoken German are important tasks for creating this new type of lexicographic resource. These tasks are directly related to the selection of headwords and to the methodological procedures for creating the resource (see Section 4).

The results of two surveys on the expectations for and requirements of a resource for the standard lexicon of spoken German, (cf. Meliss, Möhrs & Ribeiro Silveira 2018) confirm that the lexicographic codification of spoken language and its interactional characteristics are not satisfactorily addressed in current dictionaries (cf. Meliss 2016: 195; Eichinger 2017: 283), despite the call made almost 15 years ago by Trap-Jensen that “we should pay more attention to spoken language when making our dictionaries” (2004: 311). Apart from particular recent advances in spoken language lexicography (cf. Verdonik & Sepesy Maučec 2017; Hansen & Hansen 2012; Siepmann 2015), both consideration of and experience with spoken language in lexicography are still rare. Hence, the LeGeDe resource can barely rely on existing models that could provide guidance in the creation of the list of headwords.

For developing this novel type of a spoken language lexical resource, the definition of “headword” from the research tradition of dictionaries based on written language has to be supplemented with new perspectives to address the peculiarities of the spoken form (cf. Deppermann, Proske & Zeschel

<sup>1</sup> The project LeGeDe (<http://www.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html>; [16/05/2018]) is a third-party funded project of the Leibniz Association. It is being implemented as a cooperation project of the Department of Pragmatics and Lexical Studies at the Institute for the German Language in Mannheim and has a duration of three years (2016-2019); for a detailed project description, see Meliss & Möhrs 2017.



2017). The headword concept has to include considerations that take into account the one-word lemmas that have specific meanings and uses in spoken language in comparison to written language. In addition to one-word lemmas, multi-word expressions and constructions that have specific functions in interaction (e.g. *ich weiß nicht* or *keine Ahnung*; cf. Bergmann 2017) are also of interest as headword candidates (see Section 5).

The aim of our contribution is to present the corpus-based and interpretative method we used for detecting salient terms of typical spoken lexicon. In this study we focus on one-word lemmas (see Sections 3 and 4).

## 2 The Project LeGeDe

The aim of the LeGeDe project is to develop a corpus-based electronic resource that addresses lexical peculiarities of spoken German in interaction. The specifics of spoken language lexis were rather neglected in previous lexicographic codifications and research (Meliss & Möhrs 2017: 47). Within the framework of the (present) project work, however, they are to be identified, analyzed and described via different corpus-based and interpretative methods.

The subject matter covered by the LeGeDe project is the lexicon of spoken German characterized by the feature “standard”<sup>2</sup>, which allows a differentiation to other medial language varieties. Of particular interest are the distinctive features of the spoken lexicon in comparison to the lexicon of the written standard language. In order to find these, we work with the largest corpus of spoken German in interactional settings (FOLK: Research and Teaching Corpus of Spoken German, Schmidt 2014; 1,96 Million tokens). One of the key research and methodology issues addressed by the LeGeDe project is the detection of differences of the lexicon of written and spoken language, and thus the selection and description of headwords that represent the most typical and distinctive phenomena of spoken lexis.

## 3 Corpus-based procedures for assisting the creation of a list of headwords

In order to assist the selection of headword candidates for the LeGeDe resource, we performed a lemma comparison between FOLK and the German reference corpus of written German (DeReKo 2017 I, cf. Kupietz/Keibel 2009; 30 Billion tokens). Since we used DeReKo as a representation of current written language, we excluded the data containing conceptual spoken language represented in Wikipedia discussions, as well as the sub-corpus “Sprachliche Umbrüche”, dating from 1945 to 1968. FOLK is lemmatized automatically with TreeTagger (Schmid 1994) by using a parameter file trained on a manually annotated gold standard (Westpfahl & Schmidt 2016). For DeReKo several different annotation layers are provided (with many different suggestions about lemmatization, lemma forms, and, also, part-of-speech information, cf. Stadler 2014). For ease of comparison of spoken and written data, we selected only the highest ranked suggestion of the TreeTagger lemmatization. The TreeTagger was applied to the written data with an especially customized parameter file that was prepared by the author of the tool for DeReKo.

Firstly, we simplified the part-of-speech tags in FOLK (cf. Westpfahl 2014) to more universal categories (V for verbs, N for nouns etc.) to facilitate the headword selection. Then, we aggregated the part-of-speech tags with which a lemma has been tagged in FOLK (V/N if a word was tagged as verb and as a noun). In addition, we marked the lemmas that we did not want to consider in further examinations as outliers. These were: lemmas tagged as proper names, numerals, lemmas affected by orthographic reforms, idiosyncrasies and lemmas that occurred only in one transcript.

<sup>2</sup> This means that the project will not consider dialects (such as Bavarian), sociolects (such as adolescent language) or idiolects.

We calculated the difference in lemma distribution in the corpora by using different effect size measures (odds ratio, %diff, relative risk, binary log of relative risk, and frequency classes) and measures of statistical significance (log likelihood ratio and chi square). Then, we integrated the table containing the lemma comparison into a tool we developed for facilitating, filtering and sorting the data in a fast and user-friendly way. With the help of this tool, the headword candidates can be assessed, performed and explored dynamically, and the parameters can be adjusted to the needs of lexicographers. Moreover, for each lemma a direct link to the corpus examples has been provided.

After examining the output of different measures of frequency comparison, we chose to work with the difference of “frequency classes” (‘Häufigkeitsklassen’; cf. Keibel 2008), a measure inspired by the word distribution in allusion to Zipf’s law, which is relatively intuitive to understand and commonly used in German lexicography (cf. Klosa 2013a). The most common word in a corpus is in frequency class 0, the word(s) being approximately half as frequent as the most frequent word are in class 1, the words being approximately half as frequent as those in class 1 are in class 2, etc. We defined the shift in frequency classes of two corpora as the “difference of frequency classes” ( $fc\_diff = fc(dereko) - fc(folk)$ ). Since the lemma *kriegen* (en: *to get*) in FOLK is in class 5 and in DeReKo in class 12, the shift between the classes amounts to 7; hence, the difference in frequency classes ( $fc\_diff$ ) between the two corpora is 7 (see Table 1).

As shown in Figure 1, lemmas which have the highest  $fc\_diff$  are in higher  $fc$  and are hence rare in both corpora, which is in part due to the fact that DeReKo, being much larger, contains a higher number of frequency classes (31) than FOLK (16).

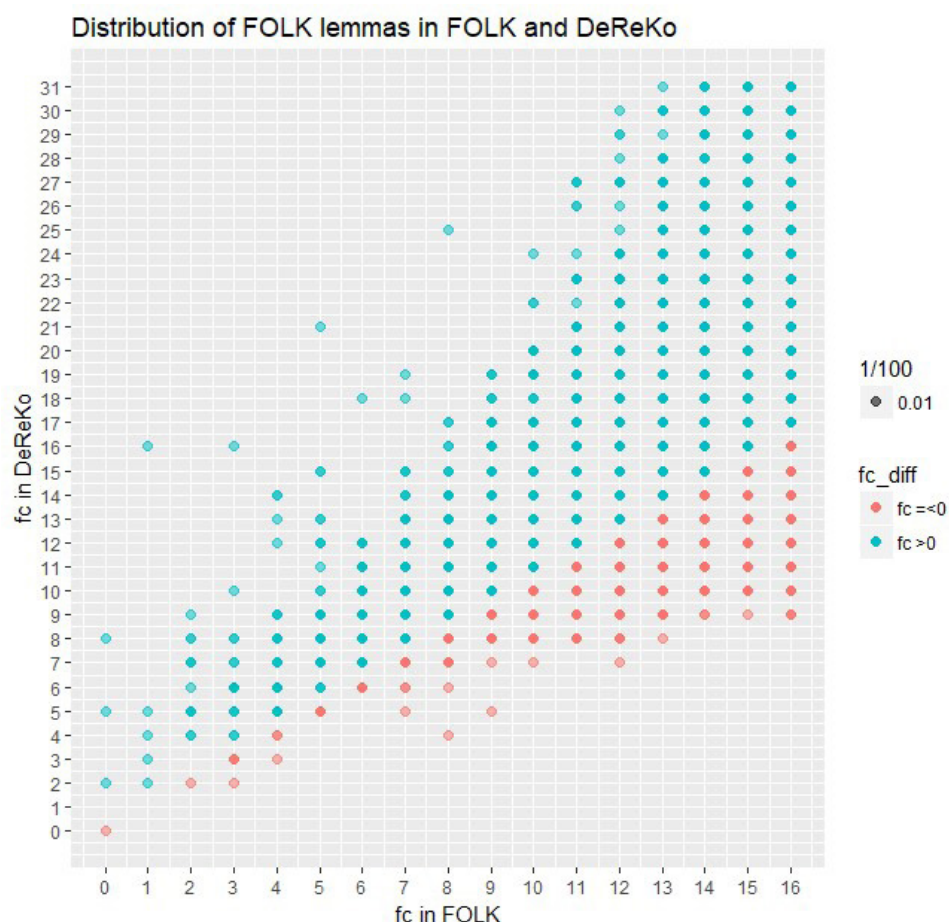


Figure 1: The distribution of the FOLK lemmas according to their frequency classes in FOLK and DeReKo. The positive  $fc\_diff$  (in blue) shows the lemmas being more frequent in FOLK than in DeReKo, the negative/equal  $fc\_diff$  (in red) shows the lemmas being less (or equally) frequent in FOLK than in DeReKo.

Since we wanted to concentrate particularly on lemmas being frequent in FOLK, we assumed that the relevant candidates for our headword list would be found by selecting the lemmas having the FOLK *fc* lower than 9, and DeReKo *fc* lower than 15. By setting these parameters, we were able to filter out those lemmas that have a high *fc\_diff* but are rare in both corpora, such as, for instance, *Spinosaurus*, having a *fc\_diff* of 7 and *fc* 15 in FOLK and *fc* 22 in DeReKo (see Table 1).

Table 1: Examples of FOLK lemmas having the *fc\_diff* of 7 with respect to DeReKo.

Lemma	FOLK <i>fc</i>	DeReKo <i>fc</i>	<i>fc_diff</i>
nein	3	10	7
kriegen	5	12	7
irgendetwas	7	14	7
Schornsteinfeger	9	16	7
Transistor	11	18	7
Proseminar	13	19	7
Spinosaurus	15	22	7

After sorting the lemma list according to the descending *fc\_diff*, we manually examined 322 one-word lemmas whose *fc\_diff* amounted to at least 2 (see Figure 2), with the aim to investigate whether they are suitable headword candidates for our resource. The corpus-specific lexemes that occur in very specific contexts or transcripts in FOLK, such as *Zentimeter* (maptask) and *Dialekt* (language biographical interviews), were sorted in the manual analysis.

## 4 Selection of headword candidates

With the method described above (see Section 3) we found lemmas that cover different subject areas of interest, such as particles, interjections, routine formulae, expressions of vagueness, deictic expressions, and other peculiarities in relation to style and register.<sup>3</sup> These subject areas have been mentioned before in the research literature regarding spoken German (see for example Schwitalla <sup>4</sup>2012; Fiehler 2016). The evidence for the pervasiveness of these phenomena in spoken language can be detected already by observing the top 25 lemmas with the highest *fc\_diff* (see Figure 2). For example, the modal particles, such as *halt* and *mal* as well as interjection particles such as *ah*, *ach*, *oh*, *eh* are very common in spoken German, but insufficiently covered in German monolingual dictionaries (see Section 1). The lemmas *irgendetwas*, *irgendwie* and *sozusagen*, that also occur in the top 25 seem to testify to the prevalence of vagueness in everyday speech.

Other examples of lemmas that we identified as potential headwords among the candidates of our headword list are “passepartout” words, such as *Ding*, *Sache*, *machen*, *tun*. In addition, we detected several lexical alternatives to the expressions of standard written German, such as *kriegen* to *bekommen*. When inspecting the adverbs in our sample of 322 headword candidates, we found different types of deictic expressions, such as temporal (*nachher*, *jetzt*) and local adverbs (*da*, *hier*). Some adjectives like *gut*, *klar*, *cool*, *toll*, *super*, *fertig*, *geil*, *krass*, *ehrlich*, that were also detected as having high *fc\_diff*, are also a focus of our interest because in spoken language they may also serve as discourse particles, among other functions. We have identified several verbal lemmas that can be assigned to different semantic subgroups of interest, such as perception (visual, auditory): *gucken*, *schauen*, *sehen*, *hören*; cognition (mental): *wissen*, *glauben*, *denken*, *meinen*, *überlegen*, *verstehen*, *kennen*; locomotion: *gehen*, communication: *reden*, *sagen*, *fragen*, emotion: *mögen*, *gefallen*, and

<sup>3</sup> A detailed overview of the subject areas can be found in Meliss & Möhrs 2017: 43.

■ FOLK vs. DeReKo

Column visibility CSV Show 25 entries

Lemma	FOLK HK	DeReKo HK	HK Diff	Filter	PoS
okay	4	14	10	1	NG
ah	4	14	10	1	NG
ach	4	13	9	1	NG
ja	0	8	8	1	PTK/NG
oh	5	13	8	1	NG
gucken	5	13	8	1	V
halt	4	12	8	1	PTK/NG
du	2	9	7	1	P
nachher	7	14	7	1	ADV
danke	7	14	7	1	NG
irgendetwas	7	14	7	1	P
na	5	12	7	1	NG
irgendwie	5	12	7	1	ADV
kriegen	5	12	7	1	V
nein	3	10	7	1	NG
mal	2	8	6	1	ADV/PTK
eh	7	13	6	1	ADV
Mama	7	13	6	1	N
cool	7	13	6	1	NG/ADJ
drin	6	12	6	1	ADV/PTK
drauf	6	12	6	1	ADV/PTK
dran	6	12	6	1	ADV/PTK
raus	6	12	6	1	PTK/ADV
sozusagen	6	12	6	1	ADV/NG
dein	5	11	6	1	P

Search Lemma <9 <15 >1 1 Search PoS

Showing 1 to 25 of 322 entries (filtered from 52,966 total entries)

Previous 1 2 3 4 5 ... 13 Next

Figure 2: Top 25 lemmas with the highest *fc\_diff* between FOLK and DeReKo, having FOLK *fc* (FOLK HK) lower than 9, DeReKo *fc* (DeReKo HK) lower than 15, and *fc\_diff* of at least 2 classes. *HK Diff* stands for *fc\_diff*; *Filter: 1* stands for the lemmas we consider for the headword list (no numerals, no proper names, etc.); *PoS* stands for the aggregation of parts-of-speech with which a lemma occurs in FOLK.

modal verbs: *können*, *müssen*. Some qualitative studies demonstrate that the verbs of these semantic subgroups have an important potential in interactional contexts to realize several special meanings, functions and formal particularities according to the different subject areas described above (cf. Dep-permann et al. 2017). Additionally, they can be the lexical basis of several multi-word-lemmas, which must be considered in further work.

After defining the headword candidates based on the corpora comparison, we analyze the corpus examples and focus on the peculiarities in meaning, use and function in talk-in-interaction by following the approaches of interactional linguistics and lexicology.



## 5 Discussion

The selection of headwords is related to the scope of the dictionary and its static vs. dynamic “dictionary under construction” conception (on these aspects, cf. Atkins & Rundell 2008: 160 ff.; Klosa 2013: 518; Schnörch 2005: 71; Wiegand 1983). Since our lexicographic resource is intended to be a digital one, in terms of quantity, our list of headwords is per definition dynamic and extensible. As shown in Section 4, for the first step of dictionary creation we used the most common one-word lemmas with a prominent frequency difference between FOLK and DeReKo. The combination of automated procedures and manual analysis of the data has proven to be an effective and sustainable way of approaching the task of headword selection. With the help of the tool for automatic sorting and filtering of the headword candidates according to their parts of speech, frequency and other features, the selection of more headwords can be carried out in a transparent and scalable way in further steps of our project, and can be enhanced in significantly.

The measure of frequency class difference proved to be a suitable measure for detecting the difference in the lemma distribution in the two corpora. Since we integrated the other measures that were commonly used for lemma frequency comparison as well, we can further investigate the effects of other measures and compare them with each other in further studies.

We did not encounter any great discrepancies that emerged from different lemmatization conventions in the two corpora, presumably because we worked with frequent lemmas and closely comparable operationalizations. However, the lemmatization process and the varieties of conventions in the German language corpora are certainly an issue that has to be taken into consideration when comparing the sets of lemmas in two corpora in future works.

The detection of one-word lemmas is only the first step towards a set of headwords needed to represent the lexicon of spoken German in interaction, which is also built upon multi-word expressions to a significant degree. In further steps of our project, we will work on the lexicographic implementation and the detection and integration of such expressions into our headword concept.

## References

- Atkins, B. T. S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bergmann, P. (2017). Gebrauchsprofile von *weiß nicht* und *keine Ahnung* im Gespräch - Ein Blick auf nicht-responsive Vorkommen. In H. Blühdorn, A. Deppermann, H. Helmer, T. Spranz-Fogasy (eds.) *Diskusmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung, pp. 157-182.
- Deppermann, A., Proske, N., Zeschel, A. (eds.) (2017). *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr (= Studien zur deutschen Sprache, Band 74).
- Eichinger, L. M. (2017). Gesprochene Alltagssprache. In *Deutsche Akademie für Sprache und Dichtung / Union der deutschen Akademien der Wissenschaften* (eds.) *Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache*. Tübingen: Stauffenburg, pp. 283-331.
- Fiehler, R. (2016). Gesprochene Sprache. In: A. Wöllstein (ed.) (2016): *Duden – Die Grammatik. Unentbehrlich für richtiges Deutsch*. Berlin: Dudenverlag, pp. 1181-1260.
- FOLK. *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (Release 2.8 vom 06.04.2017). Accessed at: <http://agd.ids-mannheim.de/folk.shtml>. [16/05/2018].
- Hansen, C., Hansen, M. H. (2012). A Dictionary of Spoken Danish. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012*. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929-935.
- Institut für Deutsche Sprache (2017). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release 08.03.2017). Mannheim: Institut für Deutsche Sprache. Accessed at: <http://www.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html>. [16/05/2018].



- Keibel, H. (2008). *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung*. Mannheim: Institut für Deutsche Sprache. Elektronische Ressource.
- Klosa, A. (2013a). Aktuelle Tendenzen in der deutschen Lexikographie der Gegenwart. In G. Stickel, T. Váradi (eds.) *Lexical Challenges in a Multilingual Europe. Contributions to the Annual Conference 2012 of EFNIL in Budapest*. Frankfurt am Main: Lang, pp. 75-93. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 99).
- Klosa, A. (2013b). The lexicographical process (with special focus on online dictionaries). Berlin/New York: de Gruyter. In R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds.) *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, pp. 517-524 (= Handbücher zur Sprach- und Kommunikationswissenschaft. Bd. 5.4).
- Kupietz, M., Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi, Y. Kawaguchi (eds.) *Working Papers in Corpus-based Linguistics and Language Education*, no. 3. Tokyo: Tokyo University of Foreign Studies (TUFS), pp. 53-59. Accessed at: [http://cblle.tufts.ac.jp/assets/files/publications/working\\_papers\\_03/section/053-059.pdf](http://cblle.tufts.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf). [16/05/2018].
- Meliss, M. (2016). Gesprochene Sprache in DaF-Lernerwörterbüchern. In B. Handwerker, R. Bäuerle, B. Sieberg (eds.) *Gesprochene Fremdsprache Deutsch*. Baltmannsweiler: Schneider, pp. 179-199. (= Perspektiven Deutsch als Fremdsprache, Band 32).
- Meliss, M., Möhrs C. (2017). Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe. In *Sprachreport* 4/2017, pp. 42-52. Accessed at: <http://pub.ids-mannheim.de/laufend/sprachreport/pdf/sr17-4.pdf>. [16/05/2018].
- Meliss, M., Möhrs, C., Ribeiro Silveira, M. (2018). Erwartungen an eine korpusbasierte lexikografische Ressource zur Lexik des gesprochenen Deutsch in der Interaktion: Ergebnisse aus zwei empirischen Studien. In *Zeitschrift für Angewandte Linguistik*. 68/1, pp.103-138.
- Möhrs, C., Meliss, M., Batinić, D. (2017). LeGeDe - towards a corpus-based lexical resource of spoken German. In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Leiden, the Netherlands, 19.-21. September 2017. Brno: Lexical Computing CZ s.r.o., 2017, pp. 281-298.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. Accessed at: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf>. [16/05/2018].
- Schmidt, T. (2014). The research and teaching corpus of spoken German – FOLK. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.) *Proceedings of the 9th Conference on International Language Resources and Evaluation (LREC'14)*. 26-31 May 2014. Reykjavik, Iceland, pp. 383-387. Iceland: ELRA. Accessed at: <http://www.lrec-conf.org/proceedings/lrec2014/index.html>. [16/05/2018].
- Schnörch, U. (2005). Die elexiko-Stichwortliste. In U. Haß (ed.) *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin/New York: de Gruyter, pp. 71-90. (= Schriften des Instituts für Deutsche Sprache 12).
- Schwitalla, J. (2012): *Gesprochenes Deutsch. Eine Einführung*. Berlin: Schmidt. (= Grundlagen der Germanistik, Band 33).
- Siepmann, D. (2015). Dictionaries and spoken language: A corpus-based review of French dictionaries. In *International Journal of Lexicography*. 28 (2), pp. 139-168.
- Stadler, H. (2014). Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora. (= *OPAL - Online publizierte Arbeiten zur Linguistik* 5/2014). Mannheim: Institut für Deutsche Sprache. Accessed at: <http://pub.ids-mannheim.de/laufend/opal/opal14-5.html>. [16/05/2018].
- Trap-Jensen, L. (2004). Spoken Language in Dictionaries: Does it Really Matter? In G. Williams, S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress. 6-10 July 2004*. Lorient, France: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, pp. 311-318.
- Verdonik, D., Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. In *International Journal of Lexicography*. 30 (2), pp. 143-166.
- Westpfahl, S. (2014). STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In L. Levin, M. Stede (eds.) *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 1-10.

- Westpfahl, S., Schmidt, T. (2016). FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.) *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 1493-1499.
- Wiegand, H.-E. (1983). Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. In H.-E. Wiegand (ed.) *Studien zur neuhochdeutschen Lexikographie III*. Hildesheim et al.: Olms, pp. 401-474. (= Germanistische Linguistik, 1-4/82).

# fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data

**Peter Meyer, Mirjam Eppinger**

*Institut für Deutsche Sprache, Mannheim*

*E-mail: meyer@ids-mannheim.de, eppinger@ids-mannheim.de*

## Abstract

We present the conceptual foundations and basic features of fLexiCoGraph, a generic software package for creating and presenting curated human-oriented lexicographical resources that are roughly modeled according to Měchura's (2016) idea of graph-augmented trees. The system is currently under development and will be made accessible as open source software. As a sample use case we discuss an existing online database of loanwords borrowed from German into other languages which is based on a growing number of language-specific loanword dictionaries (*Lehnwortportal Deutsch*). The paper outlines the conceptual foundations of fLexiCoGraph's hybrid graph/XML data model. To establish a database, XML-based resources may be imported or even input manually. An additional graph database layer is then constructed from these XML source documents in a freely configurable, but automated way; subsequently, the resulting graph can be manipulated and enlarged through a visual user interface in such a way that keeps the relationship to the source document information explicit at all times. We sketch the tooling support for different kinds of graph-level editing processes, including mechanisms for dealing with updated XML source documents and coping with duplicate or inconsistent information, and briefly discuss the browser interface for end users.

**Keywords:** graph-based dictionaries, editorial process, data modeling, linked data, historical lexicography

## 1 Introduction

With the rise of Linked Data approaches to lexicography (cf. Bosque-Gil, Gracia & Gómez-Pérez 2016), graph-based data modeling for dictionaries has gained considerable momentum. For the time being, however, there are hardly any ready-made tools that would assist a genuine graph-based lexicographical editing and online publishing process. Much current research has instead concentrated on interlinking existing resources and converting them to RDF triples. Proposals on 'Linked Data-native' models of lexicographical editing have appeared on the horizon only recently (cf. Gracia, Kernerman & Bosque-Gil 2017), and are typically focused on creating standards-compliant graph resources that are well-suited for NLP applications. Older work on graphs in lexicography and lexicology (e.g. Polguère 2012) has largely remained experimental. The software package fLexiCoGraph presented in this paper is intended as a practical tool that enables working lexicographers to create, manually edit, validate and publish human-oriented lexicographical resources, using a graph-based data modeling layer only to the extent necessary and desired, with the help of an easy-to-use administration interface. Later conversion to a Semantic Web-ready format is possible and will be supported by the software.

In order to illustrate our general ideas on data modeling and editing, we will consider, as a running sample use case for the software, an existing online database of German loanwords in other languages (*Lehnwortportal Deutsch*) which is currently in the early stages of being reimplemented with fLexiCoGraph as its new backend. In this use case, graphs are an indispensable tool to represent complex relationships between loanwords and etyma, such as borrowing histories of words (possibly spanning

multiple languages); formation of derivatives and compounds from loanwords in the recipient language; etc. (Meyer 2014; Bowers & Romary 2017). Our choice for the sample use case has mainly expository reasons, however, and is not meant to indicate a preferred area of applicability. We believe that the flexible architecture of the software and data model makes it suitable for a wide range of applications in lexicography; cf. Section 5.

## 2 Data Modeling

With a clear focus on human-oriented lexicography (instead of computational lexical resources), our approach shares many important features with Měchura's (2016) idea of *graph-augmented trees* in that it does not consist in representing all possible relations between the entities, attributes and so on as graphs, but in using conventional XML documents (henceforth, *resource documents*) as a starting point for the lexicographical process and to superimpose a graph-based data structure (henceforth, *graph component*) only where useful and appropriate for the lexicographical task at hand.

In our sample use case, the resource documents of the *Lehnwortportal Deutsch* represent entries on German loanwords in other languages, taken from different loanword and etymological dictionaries (resources) and encoded in XML. The XML schema typically varies from resource to resource. Resource documents may contain references to still other source files in various digital formats. The graph component<sup>1</sup> is supposed to represent the network of relationships between the words treated in these entries in a cross-resource, unified way. We posit two distinct node types, one for words (more generally, lexical units) and one for word senses, and an array of inter-word relations ('is borrowed from', 'is derived from', 'is a diasystemic variant of', ...) as edge types. Given the possible structural and conceptual heterogeneity of the resources, fLexiCoGraph does not impose any restrictions on data modeling. No assumptions regarding the resource documents' XML schemas are made. No particular ontology is presupposed for the graph component, such that the node and edge types needed for a specific application can be configured with respect to their attributes as one wishes.

A central organizational principle of fLexiCoGraph is the separate treatment of two interrelated parts (non-overlapping but interconnected subgraphs) in the graph component, viz. the *source layer* of information as provided by the individual underlying resource documents and the cross-resource *curated layer* representing edited, corrected and annotated lexicographical data that would typically be presented to the end user. In our use case, different loanword dictionaries may provide complementary or even contradictory data on etyma, loanwords, and their relations to each other (source layer); these data must be homogenized and interconnected in manual lexicographical work for the online presentation (curated layer). The hybrid graph/XML data model of fLexiCoGraph thus generalizes Měchura's proposal of "a data structure that allows fragments of entries to be 'shareable', able to appear in multiple entries" (Měchura 2016:98). While the source layer subgraph simply reproduces data already contained in the resource documents, the vertices and edges added on the curated layer enriches the source layer information with arbitrarily complex lexicographical identifications, specifications, abstractions, corrections and generalizations.

The process of graph creation in the source layer is, of course, driven by the resource documents. These files ultimately define conventional "units of presentation" (which might or might not correspond to traditional dictionary entries) that either contain or reference the data to be processed and

<sup>1</sup> The implementation currently used for the *Lehnwortportal Deutsch* internally represents relationships between words through a directed acyclic graph modeled in a relational database. This representation is highly limited in scope and generalizability, however, and must be recreated from scratch each time the underlying resource data (set of dictionaries or their entries) changes. See the online documentation and Meyer (2014) for more details.

presented. The mapping process of data and structural information in the XML files onto graph constellations of the source layer is freely configurable for individual resources through either XSLT or a special domain-specific language created for fLexiCoGraph. Each source layer vertex with all of its properties must correspond to an XML element, however large or small, in some resource document; similarly, each source layer edge represents, in a pre-defined way, some kind of structural configuration between the two XML elements corresponding to the vertices connected by the edge. The resulting source layer subgraph *cannot* be edited manually, since it is supposed to be a faithful portrait of information represented in the original resource. In our use case, source layer vertices with their attributes represent XML elements that model words with all their relevant properties as to grammar, part of speech, language/dialect and so on, or that model word senses (and are therefore typically descendant elements of the word elements).

In a similar fashion, a corresponding subgraph on the curated layer is bootstrapped on first import of the resource in a freely configurable and scriptable way. In our sample use case, the curated layer subgraph automatically created from a resource document is often just a replica of the source layer subgraph, where corresponding vertices are interconnected by a dedicated ‘source-to-curated’ edge type; many automated graph reconfiguration processes take place during bootstrapping, however. Amongst other things, we map homographic German etyma as they appear in different loanword dictionaries (and thus produce multiple etymon graph nodes for homographic words in the source layer) onto only one shared etymon node in the curated layer that is connected to each corresponding source layer node by a ‘source-to-curated’ edge, thereby formalizing the fact that the different source dictionaries probably refer to the same lexeme. This is a typical way of creating ‘links’ between different, originally independent resources on the curated layer. In such cases, the nodes or edges in question can be marked automatically by certain searchable attributes (flags) for later lexicographical review; after all, homographic words might still belong to different lexemes. Other flags may mark whether information available at the different source layer nodes linked to one and the same curated layer node is contradictory. The curated layer can be edited manually by the lexicographer and corresponds to the final graph-based lexicographical information the end user will be presented with. In all cases the connection of curated layer data to the original resource data is formally reconstructable in the graph component by following paths from the curated to the source layer.

### 3 User Interfaces for Data Management and End Users

fLexiCoGraph offers a large number of browser-based administration and editing tools for the working lexicographer. In this paper, only a cursory overview of some of the more important functions and features will be given.

- There are two different, but interrelated presentation and navigation/search modes, both for end users and lexicographers: A template-based one based on resource documents (more or less corresponding to the activity of “browsing through entries”), and one based on the graph component (allowing users to navigate through the graph). Both layers of the latter are navigable, searchable and (for the curated layer) modifiable in an intuitive and interactive visual graph editor. Editing includes deleting and adding edges and vertices, changing their property sets, and merging nodes. Many of the relevant ‘data fusion’ tasks and problems that arise with graph-editing are well-known from other processes of combining several resources into a homogenized product (cf. Bleiholder & Naumann 2009).
- There are, of course, importing and editing options for resource documents. In the case of changed source data or when individual altered resource documents are imported again, alterations in the source layer of the graph component automatically percolate up to the curated layer, again



triggering flags on nodes and edges where subsequent editorial decisions are needed. This percolation and revision process can be configured by the lexicographer.

- fLexiCoGraph offers a dedicated *graph/XML editor* component that may be used to create and edit *graph-like* XML resource documents. Graph-like XML resource documents mostly or exclusively represent source-layer subgraphs in that they mainly contain (i) elements representing nodes of a graph and (ii) elements explicitly specifying edges between these nodes, using something like an REFID mechanism to refer to the nodes connected by the edge; cf. (Bowers & Romary 2016) for a survey of the state of the art for the sample case of coding etymological relations in XML. Graph-like XML documents can trivially be converted into source-layer subgraphs. In our use case, dictionary entries in loanword and etymological dictionaries must, in certain cases, be excerpted manually into graph-like XML resource documents that specify which words – etyma, loanwords, derivatives of loanwords, etyma of etyma, etc. – in the original dictionary entries (to be modeled as nodes) stand in which relations (borrowing; language-internal variation or diachronic development, etc.; to be modeled as edges) to one another. The original entries may contain information on remarkably complicated borrowing histories spanning multiple languages and leading to graph constellations with long and ramified paths. Manually editing XML documents representing such constellations in a conventional XML editor would be difficult and error-prone. In the graph/XML editor, lexicographers simply construct the ‘borrowing graph’ for the entry to be excerpted in a visual way by ‘painting’ nodes and edges between them on a browser canvas. Upon creating a new resource document or selecting an existing one from a list of entries, the GUI displays all related information and allows the creation of new nodes (=words) and edges for the graph of the chosen entry as well as the deletion or edition of existing nodes and edges. Figure 1 shows an example screenshot for our use case. The upper pane is the graph editor

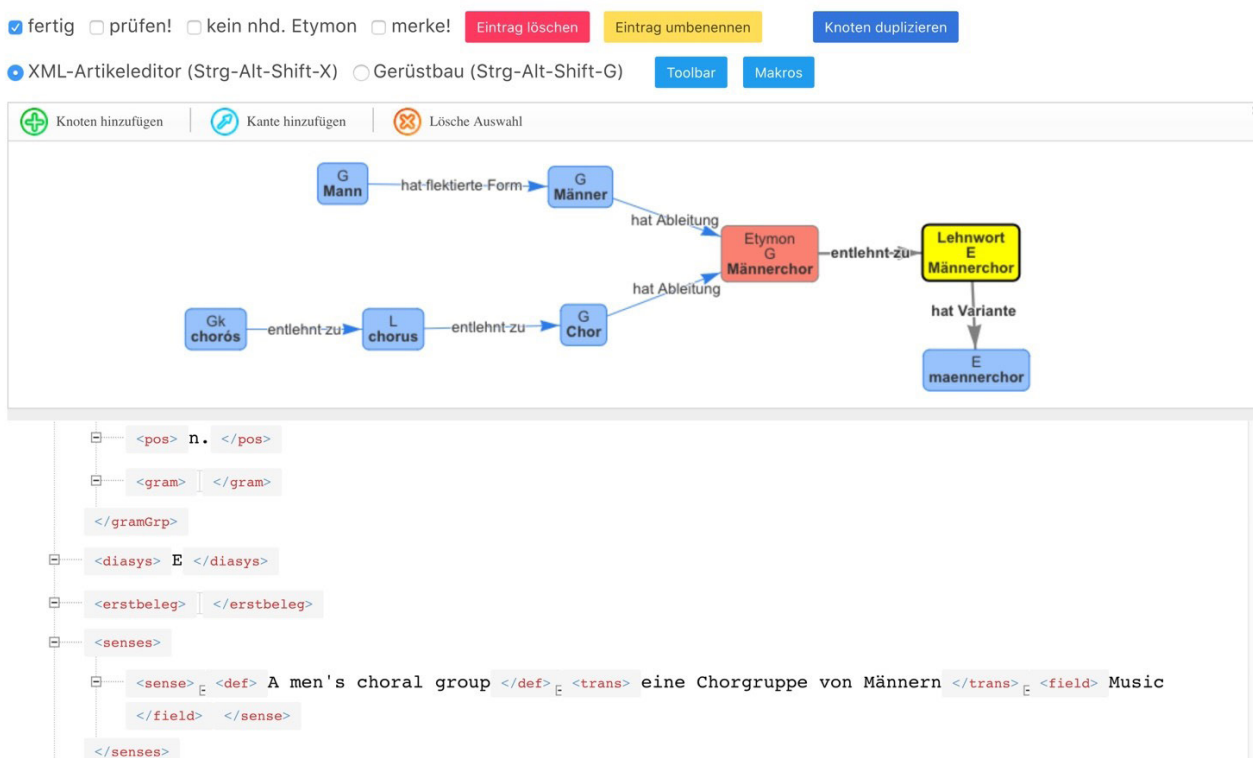


Figure 1: Partial screenshot of a customized graph/XML editor instance; the XML code pertaining to the user-selected yellow node (representing the English loanword *männerchor*) can be edited in the lower pane.

component and shows a manually scalable diagram of the entry-related subgraph. The lower pane, the XML editor component, allows the lexicographer to edit the XML fragment related to the node selected in the upper panel, taking advantage of dropdown lists with pre-defined content or attribute options for certain XML tags (depending on the configuration defined by the user). It is possible to duplicate a node if another shares a lot of XML detail information with it, e.g. in the case of variants that differ only in their word form. There is a separate scaffolding mode that makes fast creation of a first version of a subgraph possible by pasting the original entry into a separate pane (not shown here) and simply selecting words that should be represented as new nodes in the graph.

- Other functionalities include fine-grained user, resource, and editorial rights management, versioning, editorial logging, backup/restore options, and the possibility of assigning user or editorial comments to (parts of) resource documents as well as nodes and edges.
- Complex graph administration tasks can be scripted in the Gremlin graph traversal language.

The browser interface provided by fLexiCoGraph for end users is essentially a stripped-down version of the presentation and search tools at the lexicographer's disposition, without the possibility to alter data and (typically) without direct access to the source layer of the graph component. Note that the online presentation as a whole need not be tied to a network/graph metaphor. Instead, even in the graph-based navigation mode mentioned above the included template-based system allows for defining customized HTML presentations attached to only certain node types (e.g. 'headwords'), and possibly including data from relevant sections of the corresponding resource documents.

## 4 Software Architecture

fLexiCoGraph is shipped as a cross-platform server application for the Java Virtual Machine with an integrated graph database management system, editing components as well as a web server. In its default configuration, the program runs as a self-contained web application that provides the management interfaces for managing graph-based lexicographical resources and a template-based, freely configurable presentation layer for end users. The software has a pluggable architecture which can also be used

- with a Tinkerpop3-compliant third-party graph OLTP database such as Neo4J;
- with an external XML editor provided it can be customized to exchange XML data with fLexiCoGraph (as is the case at least with most commercial products);
- with a third-party web server provided it implements the Java Servlet specification.<sup>2</sup>

The software is geared towards small to medium size projects. Software development is currently in the prototyping stage; an alpha version with most of the essential features will be available for demonstration in mid-2018. The final product will be made available on the website of the authors' affiliation as open source software – Java source code and binaries for server-side installation – at a later time.

## 5 Further Areas of Application

We hope that the software package briefly presented in this paper will be useful for a variety of lexicographical tasks, such as those discussed in Měchura (2016) – complex multilingual resources and

<sup>2</sup> As an alternative option, fLexiCoGraph may be used only as editing tool; graph data can be exported in a JSON-based representation and published elsewhere.

the treatment of multi-word units – that are best solved with a graph-based approach. A particularly interesting example would be the way resources specifically dedicated to multi-word expressions (MWEs) should be organized. Sets of different entities such as partial hierarchies of ever more specific MWE construction patterns with their slots, fillers and fixed lexical elements as well as distributional contexts exhibit a complex network of interrelations (cf. Steyer and Brunner (2014) for a discussion and Steyer, Brunner and Zimmermann (2013) for a graph-like interactive online presentation). In a graph-augmented model, the vertices representing these entities would be linked to resource documents with full lexicographic descriptions, including corpus examples.

The lexicographical products created with fLexiCoGraph should easily be convertible to a Linked Data format. Tasks such as converting graph data to an RDF format, translating SPARQL queries to database-native Gremlin etc. are indeed conceptually straightforward in fLexiCoGraph. The exact extent to which default tools will be offered in the software presented here remains to be determined.

## References

- Bleiholder, J. & Naumann, F. (2009). Data fusion. In *ACM Computing Surveys (CSUR)*, 41(1), pp. 1-41.
- Bowers, J. & Romary, L. (2016). Deep Encoding of Etymological Information in TEI. In *Journal of the Text Encoding Initiative*, 10. Accessed at: <http://journals.openedition.org/jtei/1643> [31/03/2018].
- Bosque-Gil, J., Gracia, J. & Gómez-Pérez, A. (2016). Linked Data in Lexicography. In *Kernerman Dictionary News* 24, pp. 19-24. Accessed at: <http://kdictionaries.com/kdn/kdn24.pdf> [31/03/2018].
- Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward Linked Data-Native Dictionaries. In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19 – 21 September 2017*. Brno: Lexical Computing CZ s.r.o., pp. 550-559. Accessed at: <https://elex.link/elex2017/proceedings-download/> [31/03/2018].
- Lehnwortportal Deutsch*, ed. by Institut für Deutsche Sprache, Mannheim. Accessed at: [lwp.ids-mannheim.de](http://lwp.ids-mannheim.de) [31/03/2018].
- Měchura, M. (2016). Data structures in lexicography: from trees to graphs. In A. Horák, P. Rychlý, A. Rambousek (eds.) *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016*, pp. 97-104. Accessed at: <https://nlp.fi.muni.cz/raslan/raslan16.pdf> [31/03/2018].
- Meyer, P. (2014). Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries. In A. Abel, Ch. Vettori, N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen*. Bolzano/Bozen: EURAC research, pp. 1135-1144. Accessed at: [http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014\\_gesamt.pdf](http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf) [31/03/2018].
- Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. In *International Journal of Lexicography*, 27(4), pp. 396-418.
- Steyer, K., Brunner, A. & Zimmermann, Ch. (2013). *Wortverbindungsfelder Version 3: Grund*. Accessed at: <http://wvonline.ids-mannheim.de/wvfelder-v3/> [31/03/2018].
- Steyer, K. & Brunner, A. (2014). Contexts, Patterns, Interrelations - New Ways of Presenting Multi-word Expressions. In V. Kordoni, M. Egg, A. Savary, E. Wehrli, S. Evert (eds.) *Proceedings of the 10th Workshop on Multiword Expressions (MWE), Gothenburg, Sweden, 26-27 April 2014*. Association for Computational Linguistics, pp. 82-88. Accessed at: <http://www.aclweb.org/anthology/W/W14/W14-0814.pdf> [31/03/2018].

# Wordnet Consistency Checking via Crowdsourcing

**Aleš Horák, Adam Rambousek**

*Natural Language Processing Centre, Faculty of Informatics, Masaryk University*

*E-mail: hales@fi.muni.cz, rambousek@fi.muni.cz*

## Abstract

Large ontologies and semantic networks represent complex multilevel structures, which are incredibly resistant to standard proof checking procedures. Automatic consistency checks can discover system errors such as missing intralingual links, but to find a missing word sense is a difficult task. Standard solutions rely on successive consultations of multiple information sources in a multi-level review process. In this paper, we present a new approach of supplementing such multi-level reviews with engaging the dictionary users in WordNet error corrections and enhancement proposals via systematic crowdsourcing. This approach defines an early release phase with the full dataset published to the target audience followed by a continuous workflow consisting of structured adjustment suggestions obtained from the public users and of the complete editing process by expert reviewers. The review team members are handling prestructured review tasks organized in aggregated forms with correction proposals, the revision management and the appropriate editing of proposed changes. Both the users and reviewers have access to the complete revision history, which allows them to handle repeated proposals responsibly.

**Keywords:** WordNet, semantic network, ontology, consistency checking

## 1 Introduction

Long-term development and management of a large ontology or semantic network is a tedious and time-consuming process taking a lot of manual work. Even though automatic ontology consistency checks have been developed since the creation of the first digital semantic databases (Alvez et al., 2008; Tufis & Cristea, 2002; Rath, 1999), full control of the database content is always obtained by manual inspection of the data. A small team of experts cannot completely finish such a process, and new errors are always discovered in the published versions by a broader audience.

A recent example may be the discussion at the WordNet users mailing list<sup>1</sup> initiated by John McCrae in August 2017. He pointed out that the term “*church mouse*” is monosemous in the English WordNet (Fellbaum, 1998):

**WordNet Search – 3.1<sup>2</sup>**

**Noun**

- {02454543} <noun.animal>[05]

S: (n) **church mouse#1**

**(church mouse%1:05:00::)** (a mouse created by Lewis Carroll)

<sup>1</sup> <https://wordnet.princeton.edu/wordnet/contact/>

<sup>2</sup> <http://wordnetweb.princeton.edu/perl/webwn>

The further talk by several wordnet experts (including Christiane Fellbaum, the current principal English WordNet coordinator) revealed that there are at least two idiomatic senses of the term missing in the database: a “*poor creature*” (“poor as a church mouse”) and a “*quiet creature*” (“quiet as a church mouse”). The following comments showed that many WordNet developers keep their own list of discovered discrepancies in the published WordNet, and stressed the need for a standard way of reporting them and possibly incorporate the suggestions in the core WordNet.

In the following text, we summarize the current state of WordNet consistency checking approaches and issues, and present a new interface for crowdsourcing, standardizing and speeding up the process of correction of (usually small) errors discovered by the wider public. The discussed tool is developed within the DEB (Dictionary Editor and Browser (Rambousek & Horák, 2016; Horák et al., 2008)) framework used for developing a number of national WordNets.

## 2 WordNet Development and Issues

With the WordNet concept being the best known and most widespread language ontology approach, now introduced for nearly a hundred languages, some mistakes or questionable content that can appear in the released data are generally unavoidable. Issues in WordNet may be divided into two main categories:

- surface errors – problems with synset description, e.g. spelling errors in literals or definitions,
- structural errors – issues with semantic relations, appropriate literal selection, varying subtrees depth, and granularity, or orphaned synsets.

Two general methodologies defined during the EuroWordNet project (Vossen, 1998) are general used to build new WordNets:

- Expand model – with this approach, Princeton WordNet (or its part) is translated into a new language, keeping the semantic relations mostly intact. Some projects translated the synsets semi-automatically, which may introduce surface errors if the results are not verified thoroughly.
- Merge model – new WordNet is created either from scratch or based on an existing dictionary, which does not contain semantic relations and entries are not grouped to synsets. WordNets utilizing this method tend to contain more structural errors.

Many of the errors may be prevented during the WordNet development phase. The important part is to design and follow detailed guidelines (Pociello et al., 2011; Tufis & Cristea, 2002). Software tools may help significantly. WordNet editing software should check for a range of errors, from spell-checking to semantic relations completeness (Horák et al., 2006). Some projects also use periodical heuristic testing to check recently added or updated synsets (Čapek, 2012).

## 3 Crowdsourcing in Linguistics

In linguistics and NLP research, crowdsourcing is generally used to manually annotate large datasets with semantic or syntactic information (Grác, 2013), word sense disambiguation (Rumshisky, 2011), or to evaluate the results of automatic tools (Nevěřilová, 2014), but may even help to detect the outbreak of epidemics (Munro et al., 2012).

The results of crowdsourcing experiments in NLP research have been evaluated multiple times, with the results showing that combining annotations by several “unskilled” annotators may result



in cheaper and faster annotation. A study by Snow et al. (2008) found that, on average, voting on four non-expert annotations achieved the equivalent precision as a single expert annotation. Another experiment (Callison-Burch, 2009) evaluated machine translation using crowdsourcing, and concluded that a combination of many non-expert evaluations provides comparable quality to that obtained with experts.

In the field of lexicography, Wiktionary<sup>3</sup>, a sister project of Wikipedia, is one of the most prominent crowdsourced resources. The goal of Wiktionary is to create a freely available “dictionary of all words in all languages” (Wikipedia, 2017) edited by volunteers. Several analyses (Hanks, 2012; Meyer & Gurevych, 2012; Fuertes-Olivera, 2009) found Wiktionary to be a useful linguistic resource, although the entry quality varies from well-crafted to unreliable.

We have previously applied crowdsourcing principles in various annotation projects, see for example Grác (2013), Nevěřilová (2014), or Kovář (2016). Based on the experience with annotation results, we have decided to develop a tool allowing users to participate in new WordNet updates.

## 4 Crowdsourcing Tool and Review Process

Czech WordNet (CzWN) was first published as a part of the EuroWordNet and Balkanet projects (Vossen, 1998; Christodoulakis, 2004) and since then CzWN was mostly just maintained. However, there are several versions with various amount of edits, as well as a version semi-automatically extended using a large English-Czech translation dictionary (Blahuš & Pala, 2012). The NLP Centre (the CzWN developer) is currently running a project to integrate all updates to Czech WordNet and publish a new Open Czech WordNet linked to the Collaborative Interlingual Index (Bond et al., 2016).

The Czech WordNet was developed using the Expand model, translating the English WordNet synsets. The most notable example of errors caused by this approach are the synsets containing words that are not exact synonyms, or that are rare in the Czech language, but present in the Czech WordNet because of the translation from English. For example, the English synset *cabriolet:1, cab:2* has the equivalent Czech synset *kabriolet:2, dvoukolový jednospřežní povoz:1, koňská drožka:1* (*cabriolet, two-wheeled one-horse cart, horse-drawn carriage*). Although the translation is correct, this sense of *kabriolet* in Czech is very archaic, and in the current spoken language the only sense used is *the convertible car*. Another problem is the inclusion of multiword expressions in the synset, which may be justified in some cases, but which are not fixed lexical units in the Czech language. However, during the integration we will not have enough resources and lexicographers to check all the synsets and relations in the Czech WordNet. We are receiving reports and emails about issues in the Czech Wordnet, but not always in the exact form, and it is time-consuming for editors to find the right synset and fix the error. A standardized way to report errors would make the whole process much faster and more comfortable.

Based on the feedback from the DEBVisDic users, both viewers and editors, and developers of WordNet-based applications, we have developed a new software tool to enable anyone to report issues in the WordNet data. The list of features was drafted with potential future users in mind, mostly editors of the Czech WordNet and users who use DEBVisDic for WordNet browsing. Although we are testing the tool on the Czech WordNet, it is language-independent and available for all WordNets developed using the DEBVisDic editor.

3 <http://www.wiktionary.org>

The development started in summer 2017, and the first version of the application was released in October. Currently, the prototype is in testing with the Czech WordNet data. We will evaluate the testing phase and user feedback in August 2018.

The application is developed using the client-server model, programmed in Python. The server part is responsible for the suggestion storage in the database and the connection to the DEBVisDic server. The client part is a user web interface, written in JavaScript. All parts of the application are published as open-source and available for download<sup>4</sup>.

The tool is not directly integrated into the DEBVisDic editor, but it uses the DEBVisDic server API to access the WordNet data. It is possible to add new modules for integration with other wordnet editors if they provide API for the synset data update. On the other hand, all available synset representations (the editor, the simplified browser, the API calls) will enable users to move to the error reporting application efficiently.

The users are presented with a data from of the synset they were browsing and they may update any item – change an existing value, add a new one if some part of the synset is missing, or remove an unwanted item. See Figure 1 for an example of the user feedback form. The updates are stored in a separate database as suggestions. Each value (e.g. a gloss or a relation) is stored as a single suggestion.

**Road, route [n]**

Definition: [an open way \(generally public\) for travel or transportation](#)

Domain: [town\\_planning](#)

Sumo: [StationaryArtifact](#)

Sumo type: +

---

**Usages** [Add](#)

---

**Synonyms** [Add](#)

[road 1](#)

[route 2](#)

---

**Relations** [Add](#)

[ENG20-01895340-v:route:2](#)

[ENG20-01897936-v:route:1](#)

[Cancel](#) [Save](#)

Figure 1: Reporting an error in wordnet synset

<sup>4</sup> Source code repository available at <https://github.com/jirkle/DEBVisDic-Report>

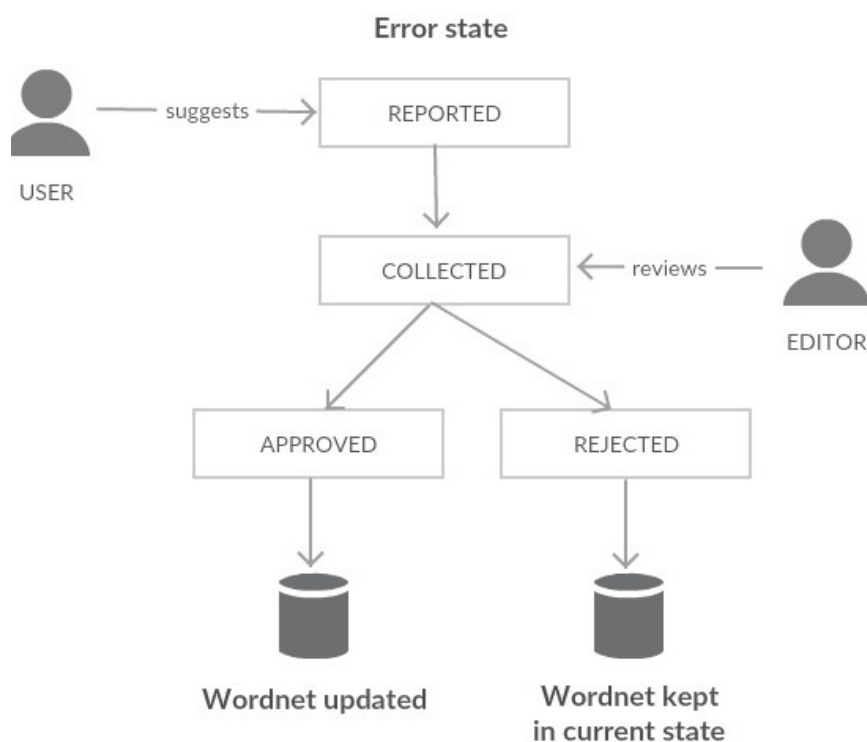


Figure 2: Review process schema

House [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Add"/> Synonym → home		<input checked="" type="checkbox"/> <input type="checkbox"/>
<u>Car, auto, automobile, machine, motorcar</u> [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Remove"/> Synonym motorcar →		<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Edit"/> Synonym machine → motorcar		<input checked="" type="checkbox"/> <input type="checkbox"/>
Love, passion [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Add"/> Usage → love is in the air		<input checked="" type="checkbox"/> <input type="checkbox"/>
Cat, true cat [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Edit"/> Domain zoology → catology		<input checked="" type="checkbox"/> <input type="checkbox"/>
Cat, true cat [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Add"/> Relation → ENG20-02352256-n:[n] paw:1		<input checked="" type="checkbox"/> <input type="checkbox"/>

Figure 3: Administrator's review of submitted suggestions

Any member of the editing team with access permissions to the given wordnet may browse all user suggestions (or filter them by the reporting user, the information type, or the review status). The editor may approve or reject any single proposal or approve/reject all suggestions for any synset at once. Of course, it is also possible to accept/reject all proposals based on the selected filter. Before deciding, the editor may compare the user feedback with previously approved or dismissed updates for the chosen synset. See Figure 2 for the schema of the review process and the suggestion life cycle, and Figure 3 for an example of the administrator interface for the review of any suggestions.

All the approved suggestions are immediately transferred to the development version of the WordNet database and presented to the users. When a reviewer rejects a user's feedback, the information is kept in the database and future users trying to suggest the same update are notified about the previous refusal.

In future versions, the reporting tool will support more detailed management of user roles with the possibility to provide reliable public users with tools to moderate suggestions and also enable discussion and voting about ambiguous synsets. Based on the prototype evaluation, we will consider extensions to the data presentation, e.g. enable users and developers to use data with suggestions and marking "synset reliability."

## 5 Conclusion

We have presented a new infrastructure for wordnet consistency checking and error reporting via crowdsourcing. The process covers all the necessary phases of the database enhancement workflow, starting with a structured proposal for an error fix in the WordNet data by a public user, followed by aggregated semi-automatic checks reviewed by a WordNet editor and projecting the correction in the development as well as the stable version of the covered WordNet database.

In the future, we will carry out a thorough public testing of this infrastructure with the Czech WordNet, and finally propagate the interface to all WordNets developed within the DEB (Dictionary Editor and Browser) framework.

Once the tool is thoroughly tested on WordNet data, it will be extended for use with any dictionary in general. The inclusion of the public enhancement proposal capability in the DEB framework will then allow to further unify and generalize the process of aggregating user suggestions to the dictionary content, and offer a straightforward application of the crowdsourced data editing to other dictionary writing applications.

## References

- Alvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., and Rigau, G. (2008). Complete and consistent annotation of wordnet using the top concept ontology. In *Proceedings of LREC 2008*.
- Blahuš, M. and Pala, K. (2012). Extending Czech WordNet using a bilingual dictionary. In Christiane Fellbaum et al., editors, *6th International Global Wordnet Conference Proceedings*, pages 50–55, Matsue, Japan. Toyohashi University of Technology.
- Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In Verginica Barbu Mititelu, et al., editors, *Proceedings of the Eighth Global WordNet Conference*, pages 50–57. Romanian Academy.
- Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

- Christodoulakis, D. (2004). *Balkanet Final Report*. University of Patras, DBLAB. No. IST-2000-29388.
- Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press Cambridge.
- Fuertes-Olivera, P. A. (2009). The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary. In Henning Bergenholtz, et al., editors, *Lexicography at a Crossroads*, pages 103–120. Peter Lang, Bern.
- Grác, M. (2013). *Rapid Development of Language Resources*. Ph.D. thesis, Faculty of Informatics, Masaryk University.
- Hanks, P. (2012). Corpus evidence and electronic lexicography. In Sylviane Granger et al., editors, *Electronic Lexicography*, pages 57–82. Oxford University Press, Oxford.
- Horák, A., Pala, K., Rambousek, A., and Povolný, M. (2006). DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In *Proceedings of the Third International WordNet Conference - GWC 2006*, pages 325–328, Jeju, South Korea. Masaryk University, Brno.
- Horák, A., Vossen, P., and Rambousek, A. (2008). A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In *Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing*, pages 1–15, Haifa, Israel. Springer-Verlag.
- Kovář, V. (2016). Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pages 127–134. Tribun EU, Brno.
- Meyer, C. M. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger et al., editors, *Electronic Lexicography*, pages 259–291. Oxford University Press, Oxford.
- Munro, R., Gunasekara, L., Nevins, S., Polepeddi, L., and Rosen, E. (2012). Tracking Epidemics with Natural Language Processing and Crowdsourcing. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS12-06. AAAI.
- Nevřilová, Z. (2014). Annotation Game for Textual Entailment Evaluation. In *15th International Conference, CICLing 2014, Part I*, pages 340–350. Springer, Heidelberg.
- Nevřilová, Z. (2014). *Paraphrase and Textual Entailment Generation in Czech*. Ph.D. thesis, Faculty of Informatics, Masaryk University.
- Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language Resources and Evaluation*, 45(2):121–142, May.
- Rambousek, A. and Horák, A. (2016). DEBVisDic: Instant WordNet Building. In *Proceedings of the Eighth Global WordNet Conference, GWC 2016*, pages 25–29.
- Rath, H. H. (1999). Technical issues on topic maps. In *Proceedings of Metastructures 99 Conference*. GCA.
- Rumshisky, A. (2011). Crowdsourcing word sense definition. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 74–81. Association for Computational Linguistics.
- Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating nonexpert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.
- Tufis, D. and Cristea, D. (2002). Methodological issues in building the romanian wordnet and consistency checks in balkanet. In *Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation*, pages 35–41.
- Čapek, T. (2012). SENEQA - System for Quality Testing of Wordnet Data. In Christiane Fellbaum et al., editors, *6th International Global Wordnet Conference Proceedings*, pages 400–404, Matsue, Japan. Toyohashi University of Technology.
- Piek Vossen, editor. (1998). *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.
- Wikipedia. (2017). Wiktionary — Wikipedia, The Free Encyclopedia. [www.wiktionary.org](http://www.wiktionary.org) [Online; accessed 28-May-2018].

## Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and by the Grant Agency of CR within the project 18-23891S.



