Univerza v Ljubljani
FILOZOFSKA
FAKULTETA

Edited by Vojko Gorjanc, Polona Gantar, Iztok Kosem and Simon Krek

DICTIONARY OF MODERN SLOVENE: PROBLEMS AND SOLUTIONS

Edited by Vojko Gorjanc, Polona Gantar,
Iztok Kosem and Simon Krek

DICTIONARY OF MODERN SLOVENE: PROBLEMS AND SOLUTIONS

Book series Prevodoslovje
in uporabno jezikoslovje

Ljubljana 2017

DICTIONARY OF MODERN SLOVENE: PROBLEMS AND SOLUTIONS
BOOK SERIES PREVODOSLOVJE IN UPORABNO JEZIKOSLOVJE
ISSN 2335-335X

Contents

Introduction

In the autumn of 2015 we published a monograph titled *Slovar sodobne slovenščine: problemi in rešitve* (Znanstvena založba Filozofske fakultete UL, 660 pages), which presents the results of studies focussed on some of the key questions in the conceptualisation of a state-of-the-art dictionary of Slovene; a dictionary that would address the challenges of modern lexicography, and would promote Slovenian lexicographic theory and practice internationally. The monograph contained 32 chapters, co-authored by as many researchers.

Our point of departure were two main aims of the dictionary: to inform Slovene native speakers and other users about lexical and grammatical characteristics of the Slovene language using state-of-the-art lexicographic practices, and to provide a lexical and grammatical resource for the development of language technologies. In order for the dictionary to be useful for language technology applications, it should be conceptualised as a machine-readable database, available under open access. This will also enable the compilation of dictionaries for other target users, as we are aware there is a need for language resources that meet the needs of different types of users, from pupils and students to language professionals such as translators and editors, from native speakers to non-native speakers of Slovene.

Although the monograph focussed on the Slovene language, the presentations of the results abroad have shown that the studies are also of great interest to colleagues in other countries. This led to the decision to make a selection of relevant chapters, and translate them or adapt them for international audience. Several contributions have been in the meantime published in international journals or conference proceedings, so this monograph contains only those that have not yet been published in English.

The monograph, containing 13 chapters, presents the compilation of a dictionary that utilizes different technologies available, and is conceptualised around language technologies, i.e. it uses state-of-the-art methods of language analysis, data extraction and data storage, and visualisation. The technical aspects of the dictionary such as designing the dictionary database are presented and discussed by Klemenc et al. Then, Dobrovoljc, and Dobrovoljc et al. discuss the role of morphological information in dictionaries of Slovene, and the role of the Sloleks morphological lexicon in future dictionary projects and planned developments of the resource, respectively. A dictionary project needs to pay a great deal of attention to its users, and Rozman et al. and Mikolič provide some insights into the needs of Slovene users, with particular focus on language learners (both young native speakers and non-native speakers) and creative writers. Language is constantly changing, and at the same time, language technologies are being improved; consequently, corpora as a basis of modern language description need to be updated regularly. Thus, chapters by Logar, Erjavec et al., and Erjavec et al. address the role of the reference corpus for Slovene in dictionary compilation,

and discuss improvements needed. One of the most important aspects of the proposed dictionary is a new approach to lexicographic analysis, which includes automatic extraction of lexical data from the corpus, presented in detail by Gantar et al. In the modern society, specialised lexis plays an increasingly important role, and the chapter by Vintar discusses to what extent such a lexis should feature in the dictionary, and the methods of its selection for inclusion. The monograph is concluded by chapters by Fišer and Čibej on the potential of crowdsourcing in lexicograpy, and a few suggestions on its implementation in different stages of dictionary compilation.

This research undertaking, started at the Centre of Language Resources and Technologies, University of Ljubljana, has established a wide research network and prompted collaborations between researchers from different disciplines, extending its influence and relevance far beyond lexicography. One of the aims of this monograph is to extend our collaborations internationally, to start or strengthen links with researchers and lexicographers working on similar research topics.

Editors

Ljubljana, March 2017

Technological Design of a State-of-the-art Digital Dictionary

Bojan Klemenc, Marko Robnik-Šikonja, Luka Fürst, Ciril Bohak and Simon Krek

Abstract

An important building block of a state-of-the-art digital Slovene language dictionary is its technological framework, which is briefly presented in this paper. We view the dictionary as a multi-tier architecture with a presentation tier, a middle application tier (a back-end application system with a component for semi-automatic data extraction), and a data tier. In their natural form, the language data are multidimensional. In a printed dictionary, there is just the presentation tier, and thus many relations contained in the underlying data are difficult to access or are even lost. By contrast, in electronic dictionaries there are no such restrictions. We can preserve the data in all its complexity and present it in various ways, since there is a distinction between the data and their presentation. This separation is the key factor in integrating various data sources (different corpora and external databases) into a unified database. Various users or programs can query different parts of the database based on their interests, and the presentation tier displays or returns the data at different levels of granularity. For each tier, we present the structure and review some of the technological considerations, which guarantee good extensibility, reliability, and adaptability of the final solution.

Keywords: digital dictionary, multi-tier software architecture, presentation layer, relational database, data extraction

1   INTRODUCTION

To create a modern digital dictionary of Slovene, technological considerations are no less important than lexicographical ones. This paper thus focuses on the technological aspects of such a dictionary. In particular, we first describe the core components of a modern digital dictionary, and then outline some ideas for its implementation. When designing a digital dictionary it is now crucial to consider the issues of sustainability, scalability, adaptability, and reliability.

Early implementations of digital (or rather digitized) dictionaries were, from a data-modelling perspective, a more or less direct mapping of the existing paper-based dictionaries to the digital form (cf. Urdang 1984; Boguraev and Briscoe 1989; Hajnšek-Holz 1993; Krek 2014b). Specifically, dictionary entries, together with their hierarchical organization and tags, were stored in formats such as XML (*eXtensible Markup Language*) files or, in case of web dictionaries, HTML (*Hyper-Text Markup Language*) files. In the latter case, the logical structure of a dictionary entry is intertwined with its presentation (appearance). By contrast, an XML dictionary entry specifies only the structure of the entry, whereas its presentation is generated using template-based transformations. Such templates may be defined by, for example, the CSS (*Cascading Style Sheets*) markup language. In the case of XML, we thus have a basic separation between the data and their presentation. The text of a dictionary entry may also contain references to other dictionary entries or their components.

The search queries supported by digitized versions of paper-based dictionaries are typically limited to headwords, a restricted set of elements (usually those specified in XML), and general text search. Search results are always presented in the same way: a dictionary entry (or perhaps several entries) that match(es) the query, possibly with highlighted portions of the matching text. Unfortunately, it is impossible to obtain a query-specific presentation of search results, since the organization of the dictionary data supports only a fixed number of predefined search result views. Such an organization of the dictionary data (and entries) is natural when dealing with a medium such as paper, where the data have to be organized and stored in their final, permanent form. However, dictionaries designed for digital media do not suffer from this physical limitation. Therefore, in designing a digital dictionary, we have to think beyond paper limitations and beyond static data structures, as the data have to be stored in their natural multidimensional form. Based on the desired queries, the data then have to be suitably filtered, rearranged, and presented.

It is therefore crucial to separate the presentation of the data from the data themselves when producing a digital dictionary. In this manner, the data can be stored

in their entire complexity and presented from different viewpoints and at different levels of granularity. Technologically, it is thus important to separate the implementation of a digital dictionary into the *presentation tier* (or *front end*) and the *data tier* (or *back end*). The user does not have direct access to the data tier; he or she interacts with the data only through the presentation tier. The presentation tier presents the dictionary data to the user, intercepts the user's queries in the broad sense of the word (mouse clicks, search queries, etc.), and visualizes the results of the queries. The third component is the so-called *application tier* (or *intermediate tier*), whose role is to connect the data and presentation tiers. In particular, the application tier converts queries at the presentation tier into a form that can be used to retrieve the corresponding data from the data tier. The application tier then filters and reorganizes the retrieved data and forwards them to the presentation tier.
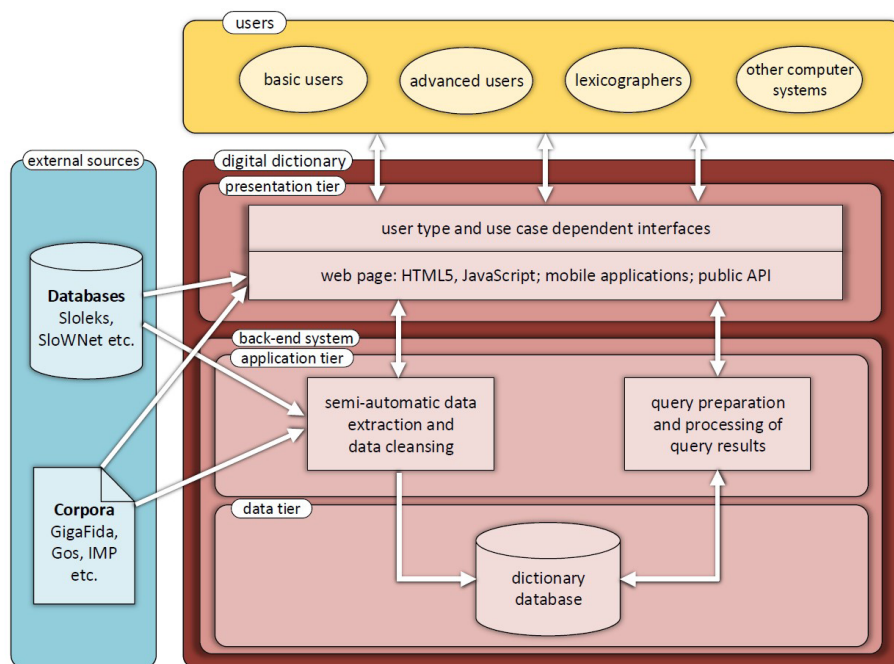
Figure 1: Architecturally, a digital dictionary is divided into three tiers: the presentation tier, the intermediate application tier, and the data tier. The user interacts solely with the presentation tier (through the webpage or mobile applications), which presents suitably selected and processed data from the data tier. The role of the intermediate tier is to connect the presentation tier and the data tier and to make it possible to fill the dictionary database from external sources.

We thus obtain a three-tier architecture (Figure 1), in which the user is only able to interact with the upper (presentation) tier, whereas the intermediate (application)

tier and lower (data) tiers are invisible. Incidentally, the application and data tiers may be collectively called the *back end*. The fact that the architecture of the system is divided into multiple tiers makes it possible for individual parts to be relatively independent of each other, and higher tiers interact with lower ones through pre-defined programming interfaces. Consequently, a given tier can be replaced with another without having any negative effect on the other tiers. In addition, the separation of the presentation tier from the database makes it possible to integrate various dictionaries and sources. The idea is to have a single unified database and multiple "views" at the presentation tier, which can visualize different subsets of the database, e.g., written language, spoken language, modern language, archaic language, regional varieties, different combinations of criteria, and so on. At the presentation tier, we might also present different user interfaces to different types of users. For example, a high school student who uses the dictionary to write an essay might want to interact with a completely different interface (with different data and a different hierarchical structure of the data) than a linguist or a lexicographer. Although all users access the same database, there may thus be substantial differences in the level of granularity of the presented data and in the possibility of reading, writing, or modifying them. For instance, a lexicographer is allowed to modify the dictionary data, while other users are not.

The dictionary database may be updated both by the manual work of a lexicographer or by crowdsourcing (cf. Kosem et al. 2013a; 2013b). In addition, the system enables automatic extraction of the dictionary data from external sources, such as corpora. Data extraction is a repetitive rather than one-time process, since the language and hence the corpora constantly change. Therefore, in addition to serving as a connection between the presentation tier and data tier, the intermediate application tier also has to connect to external sources and make the initial data extraction process possible.

Technologically, the dictionary can be divided into four main components, which we briefly describe below:

1. **The database**, being the most important component of the data tier, is implemented as a unified relational database. Its role is to store the language data and the information extracted from the corpora.

2. **The back-end application system** (the intermediate application tier) integrates the entire solution and contains programming interfaces for interacting with the presentation modules (the web application and mobile applications) and programming code for interacting with the database.

3. **The automatic data extraction component** is, in fact, part of the intermediate application tier. However, because of its complexity, we will deal with it separately. Its role is to fill and update the database with

the data extracted from external corpora and databases. As part of the lexicographical process, the automatic extraction of data is presented in Gantar et al. (2015a).

4. **The presentation tier**, both in the form of a web portal and in that of applications for different mobile platforms (e.g., Android, Apple iOS, and Windows Phone), presents the lexicographical data to different types of users, makes it possible to search and browse the data, and facilitates data corrections and updates as part of the lexicographical process (ibid.). The presentation tier is not used only by people, and thus it also includes a programming interface through which other computer systems can interact with the dictionary.

It makes sense for the implementation of the dictionary to be based on open-source solutions to the greatest extent possible. This is because such solutions are now sufficiently powerful to support advanced operations and a high number of users. The division of the system into tiers enables us to select the most appropriate technology for each and then replace individual tiers if the need arises. The same principle holds for individual components. For example, the component for the automatic extraction of data is separated from other components at the application tier; if necessary, it communicates with them via programming interfaces.

The communication between individual tiers is based on the client-server paradigm. The client sends a request to the server, and the server replies with the appropriate response. This approach makes it possible for clients within the dictionary system to have comparatively modest demands for memory and processing power, since the data are mostly stored and processed on the server, whereas the client (at the presentation tier) merely displays the results of the user's query. Lower computational demands imply a lower energy consumption, which in turn enables the use of the dictionary on less powerful mobile devices, provided that they have a data connection to the server. Since the data in the database are regularly updated, the users always have access to the most up-to-date version. Such an architectural solution does not imply that the clients and servers have to be strictly separated; however, if they are installed on the same physical device, the database or a part thereof is replicated, and so we have to ensure that the individual copies of the database are synchronized (typically with one of the canonical copies of the database). To illustrate the usefulness of such a solution, let us note that (even on mobile devices) the dictionary can be used without an Internet connection.

The multi-tier and modular structure enables us to build, evaluate, and test individual components of the dictionary in parallel. However, the necessary prerequisite for such an approach is that the connections between the individual tiers, such as programming interfaces, are well defined in advance.

2   THE DATA MODEL AND THE DATABASE

A unified database and a separate presentation tier make it possible to integrate dictionaries and sources that were previously isolated. To build a suitable unified database, we first have to define an appropriate data model that will be able to store integrated data from various existing and newly-formed databases. Besides this, the data model has to support a broader set of queries, and has to cover those that were being executed on the existing databases, and enable additional queries on the integrated data. We also have to pay attention to the fact that the integration increases the quantity of the stored data that (still) has to be quickly accessible.

Table 1 shows the data sources, their inclusion into the unified database, and the existing format of individual lexicographical data that will be displayed in the user interface. The data can either be included directly in the database (YES in Table 1) or be accessible via a link to some external source, such as corpora (NO in Table 1). For a more detailed discussion on integrated dictionary sources and corpora, see Krek et al. (2013).

Table 1: Types of displayed data, their sources, their inclusion into the database, and their current format. The labels of formats are as follows: TEI (Text Encoding Initiative), LMF (Lexical Markup Framework), and LBS (Leksikalna baza za slovenščino – Slovene Lexical Database).

Displayed data	Source of the data	Inclusion into the database	Current format
phrases	extracted data	YES, as the lexicon	XML LBS
collocations - concordances	Gigafida (Slovene language corpus)	NO, a reference to the concordancer	-
parts of speech	Sloleks (Slovene morphological lexicon)	YES, as the lexicon	XML LMF
synonyms and translations into selected foreign languages	SloWNet (Slovene semantic lexicon)	YES, as the lexicon	XML DEBDIC
history, words	IMP (Corpus of the older Slovene language)	YES, as the lexicon	XML TEI
history - concordances	IMP (Corpus of the older Slovene language)	NO, a reference to the concordancer	-

Displayed data	Source of the data	Inclusion into the database	Current format
speech, words	Gos (Corpus of the spoken Slovene language)	YES, as the lexicon	(XML TEI - implementation in the project)
speech - concordances	Gos (Corpus of the spoken Slovene language)	NO, a reference to the concordancer	-
visualization of relationships	extracted data	YES	XML LBS
multimedia	WikiMedia, ...	YES, also as external sources	different multimedia formats
lexicographical statistics	Gigafida (Slovene language corpus)	YES	-

Sources in the textual form are usually stored in XML or plain text files. In addition to the contents, the XML files also store the structural data. Since different types of data have different structures (this is in part due to the type of contents they represent), it does not make sense to keep the XML structure in the database. (However, there are some exceptions where it is reasonable to keep smaller XML parts, such as emphases in descriptions.) Being a hierarchical form of data storage by its nature, XML is not very suitable for storing non-hierarchical data, such as dictionary data. However, owing to its hierarchical layout, XML is appropriate for serialization, and an XML file itself contains the data about the structure of the underlying data. For these two reasons, XML can be used for data interchange. (In the case of the dictionary, the data is interchanged with external sources and with external applications that interact with the dictionary through programming interfaces.)

It is important to consider relationships between individual records when organizing the data in the dictionary database. These relationships can be modelled by graph or relational databases. In terms of performance (Vicknair et al. 2010), both types of databases are able to handle large quantities of data that are typically associated with a dictionary. Several query languages have been defined for both graph and relational databases. For example, there are SPARQL (SPARQL Query Language for RDF) and several non-standard solutions (Wood 2012; Haase et al. 2004) for graph databases, and SQL and SQL/PSM for relational ones. Graph databases are highly flexible, since they do not have an explicitly defined structure, and are thus suitable for data with a variable structure. On the other hand, relational databases have an explicitly defined structure, which compels us to define the data model in advance. Besides that, we also have to consider which

database queries are possible and which are not. Nevertheless, even the relational data model can be adapted in such a way that part of the structure is stored as data (Newman 2007).

Multimedia sources are stored as references in the database. To facilitate search queries, they are appropriately tagged.

Owing to the maturity of the corresponding technological solutions, the dictionary database is designed as a relational database. A simplified conceptual model of the database core is shown in Figure 2.
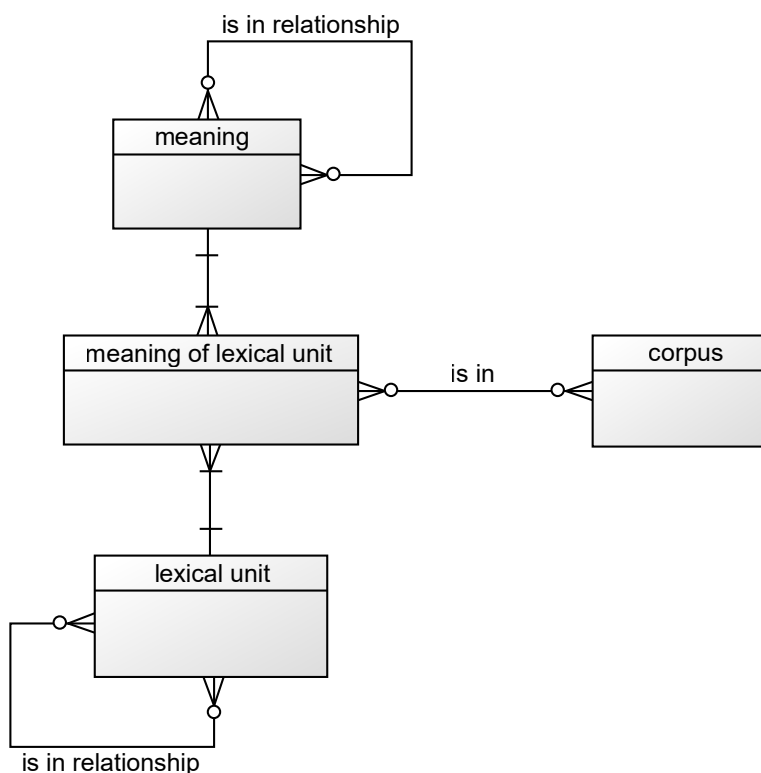
Figure 2: A simplified conceptual model of the database core, displayed in Martin's notation. This model serves as the starting point for the design of the entire database.

A lexical unit conveys a single meaning or several meanings, which can be in different relationships with one another. Lexical units can take the form of lexemes, phrases, phrasemes, or even parts of words, and can also be in different relationships with each other. For lexical units with a certain meaning, we store (aggregated) data about the sources in which they have been found.

The data model is designed in a sufficiently general way to enable the set of stored data to be extended to multiple language varieties and to treat these varieties equally. Moreover, when designing the data model we have to pay attention to the level of granularity of the data (a lower level of granularity means that we store a greater amount of aggregated data or lower-precision data, which in turn implies that it will not be possible to answer certain queries). Granularity is important both for data extraction and filling the database, since it determines what data have to be extracted and what extra amount of work will have to be carried out, e.g., in crowdsourcing or in the final processing performed by a lexicographer. For example, if, in the process of extracting data for lexical units, we do not record the time span during which individual lexical units occur, it will not be possible to restrict search queries to the lexicon from a given time span.

Several database management systems are available, and since relational databases are now well-established, there are a number of open-source solutions, although not all of these have the necessary functionalities. For our purposes, the database management system also has to support so-called recursive queries and SQL/PSM (procedures stored in the database). An example of such a system is PostgreSQL.[1]

3 THE BACK-END APPLICATION SYSTEM

The back-end application system serves as a link between the data and presentation tiers. Automatic data extraction is also part of the application system; however, owing to its complexity, we deal with it in a separate section. The role of the application system is to (re)format data requests received from the presentation level and to forward the requests to the database or external sources, such as corpora or external databases. Subsequently, the application system processes and filters the responses from the database or external sources and sends them back to the presentation level.

It is important to distinguish between the data themselves and additional restrictions and rules defined over the data, as these restrictions and rules can also change over time. For instance, collocations associated with individual lexical units can be recorded for a long time span (e.g., several centuries), but we might want to impose a rule to display only collocations occurring within, say, the last ten years. In this case, not only the data that match the rule but also the rule itself changes over time. The application tier has to make it possible to define such rules, and it has to formulate database queries based on the imposed rules and restrictions. This implies that we have to be restricted to time spans defined by

---

1   www.postgresql.org

the imposed rules and through the user interface, and we should be able to define the desired time span explicitly.

The application system provides its services in the form of a programming interface. An advantage of having separated tiers is that the source code of the application tier may be changed (completed, corrected, or improved) without affecting the programming interface, which means that the clients at the presentation level (the web and mobile applications) can still make use of the services without any modifications being needed. In addition to the clients at the presentation tier, the access to the programming interface has to be provided to other computer systems that would like to retrieve the data. We also have to enable connectivity in the sense of a semantic web (i.e., linked data).

Since the presentation and application tiers communicate according to the client-server paradigm, another important task of the application tier is to prepare the data in such a way that the clients receive only those data that they truly need, without any unnecessary data transfers.

4   AUTOMATIC DATA EXTRACTION

As shown in Table 1, the data in the dictionary database are extracted from different external sources. There are two main problems associated with data extraction: first, how to cope with the sheer quantity of the data in the external sources (for instance, the Gigafida corpus currently contains approximately 1.2 billion words), and second, how to ensure the quality of the extracted data. In addition, data extraction is not completed when the dictionary is published; rather, it is an ongoing process, since the language changes over time.

In the first stage, data are extracted automatically, and the results are then validated. Reliable data are written directly into the database, while those with a lower degree of reliability undergo a further filtering and manual processing stage.

To implement the automatic data extraction stage, we build upon the data extraction approaches developed for the purpose of creating the *Slovene Lexical Database* (*Leksikalna baza za slovenščino* in Slovene) within the project *Communication in Slovene* (*Sporazumevanje v slovenskem jeziku* in Slovene) (Gantar 2009; Gantar and Krek 2011), augmenting these approaches with more recent findings and technologically improved tools. For the entire lexicon that will be visualized, the following data can be automatically extracted: the headword in the base form (lemma), its part of speech, its frequency in the corpus, its grammatical relationships (which, in the database, are transformed into patterns),

and the corresponding collocations together with their examples. As an important step in the process of automation, the so-called word sketch grammar within the Sketch Engine[2] tool has already been created. With the help of a designated software script that contains the descriptions of all relevant grammatical relationships for extracting collocations, we can retrieve a set of good candidates for usage examples of individual headwords within a realistic textual environment (Kosem et al. 2011). The software script makes use of the so-called GDEX (abbreviation for *good dictionary examples*) configuration, which defines the properties of such examples.

In the second stage of the data extraction process, the data are manually inspected before being included in the dictionary database. This work is carried out with the help of crowdsourcing, in the context of which the users label possible anomalies or errors in the data. Eventually, the data are formatted and confirmed by a lexicographer. The errors that have been confirmed to originate from the automatic extraction process are labelled and fed back to the data extraction system, which in turn learns from the errors using machine learning techniques, and thereby improves its performance.

Automatic data extraction belongs to the back-end system. Both the partially and completely processed data are written into the dictionary database. In the database, the data that have not yet been completely processed are appropriately tagged, which means that they may be either displayed or not displayed at the presentation tier. For example, both a lexicographer and general user access the same database, but the lexicographer will, besides interacting with a different user interface, also see the data that have not yet been completely processed and will be able to process them. The users participating in crowdsourcing have their own view of the data too. For the purpose of crowdsourcing, we can use existing platforms such as PyBossa,[3] which simplify creation of crowdsourcing applications (cf. Fišer et al. 2015).

5 THE PRESENTATION TIER: THE WEB PORTAL AND MOBILE APPLICATIONS

When designing the presentation tier, and consequently also the user interfaces for different applications, we have to focus primarily on user experience. The unified visual design of the applications is no less important. One of the goals of the presentation tier is to display the data on the web pages and popular mobile platforms in a consistent way.

---

2   http://www.sketchengine.co.uk/
3   http://pybossa.com/

When developing mobile applications it is advisable to take the so-called hybrid approach, which is the best way to port applications between different mobile platforms while ensuring the maximum reusability of individual components. A reasonable option to develop the basic functionality is to use the HTML5 and JavaScript technologies. The application core developed in this way can then be embedded into the application frameworks of the individual mobile platforms that have to be supported. Such a development is supported by numerous open-source tools, e.g., PhoneGap[4], which is based on the Apache Cordova[5] platform. The hybrid approach facilitates and accelerates the development of applications for all supported platforms. In addition, it ensures a unified presentation tier on all platforms and facilitates the upgrading of the applications. The core of a mobile application created in such a way may serve as a basis for developing a web portal.

For the purposes of achieving recognisability and a consistent user experience, it is advisable to design a unified visual identity for the entire user interface. It is important to follow the WCAG 2.0 (Web Content Accessibility Guidelines 2.0) standard and thereby ensure that the applications are also suitable for users with special needs.

6   CONCLUSION

In the technological design of a modern digital dictionary Slovene, a key concept is the separation of the presentation of the data from the data themselves. By following this route, the data can be stored in their entire complexity and presented from different viewpoints and at different levels of granularity. The dictionary is designed as a three-tier architecture, consisting of a presentation tier, intermediate application tier, and data tier. The task of the presentation tier is to retrieve the requested data from the data tier and display them to the user. Between the presentation and data tiers there is the intermediate application tier, which converts the user queries from the presentation tier into a form suitable for a direct execution in the data tier (on the database), and transforms the data retrieved from the data tier into the form required by the presentation tier. Another role of the intermediate application tier is the automated extraction of data from various corpora and external data sources. Since the language is constantly evolving, automated data extraction is an ongoing process that also involves lexicographers, who access the data through the suitable views at the presentation tier.

The separation between the data and their representation plays a key role in the integration of different sources (corpora and external data sources) into a unified

---

4   http://phonegap.com/
5   https://cordova.apache.org/

database. Different users, as well as external computer systems, may retrieve the desired data from the database using queries forwarded from the presentation tier. The presentation tier then also displays the retrieved data.

The main advantage of the multi-tier architecture is the independence of individual tiers, as long as the programming interfaces through which higher tiers interact with lower ones are appropriately defined. At each tier, we can therefore choose the most suitable implementation technologies, and a change at one tier does not affect others, as long as the programming interface remains intact.

We have followed the above-mentioned principles in our proposed implementation of a modern dictionary Slovene. In particular, we have divided its technological design into four components: a database (the data tier), a back-end application system with a component for partially automated data extraction (both belong to the intermediate application tier, but the data extraction component is treated separately because of its complexity and importance for the entire system), and a presentation component with the web portal and mobile applications (the presentation tier).

The technological design of the dictionary that we have described in this paper ensures that the solution to be built upon will serve as a central web-based language portal involving all levels of the Slovene language vocabulary. The key components, which enable the sustainable development of both the web portal and mobile applications, will be made available for further improvement under a free software license.

Morphological Information in Modern Slovene Dictionaries

Kaja Dobrovoljc

Abstract

Although morphology in lexicography is generally considered to be a solved problem which mostly deals with user-oriented evaluations of its comprehensibility, online dictionaries bring new possibilities for both dictionary users and makers alike. In the context of planning a future dictionary of modern Slovene, this paper explores the language users' need for morphological information, and the different aspects of its inclusion in a born-digital online dictionary. Preliminary analysis of inflection dictionary log files confirms that there is a great need for the inclusion of inflectional information, and that users tend to search for both regular and irregular inflectional paradigms. However, this need is not sufficiently met within the recently issued edition of the reference The Dictionary of Slovene Literary Language, as decoding inflectional and other morphological information requires substantial cognitive effort and metalinguistic knowledge that cannot be expected from most users. Given that Slovenian is a morphologically rich language with extensive inflectional information, we take into account the idea of a separate machine-readable morphological database intended for use in language guides and various NLP applications. This database brings many advantages for dictionary users, such as the display of full inflectional, pronunciation and derivational paradigms, normative information, hyperlinking, improved searching, corpus linking, speech synthesis and voice search recognition. At the same time, it demands careful consideration of the content-related, visual and technical issues that arise when interlinking two distinct databases, in particular morphology-dependent polysemy and variant spelling synonymy.

Keywords: morphology, inflection, morphological lexicon, dictionary database

1   INTRODUCTION

In addition to information on the semantic properties of lexical items, dictionaries usually also include information on their formal properties, such as pronunciation, inflection, orthography and so on. Contrary to the reception-based semantic description, such information advises users on how to use lexical items in the process of actual production. This has also been standard practice in Slovenian lexicography, as ever since the first edition of *The Dictionary of Slovene Literary Language* (DSLL) most subsequent dictionaries have considered information on pronunciation, inflection and other morphological features as an indispensable part of a dictionary description, regardless of the dictionary type, i.e. general, specialised, terminological, historical, dialectical or any other type of monolingual dictionary.

Despite the fundamental role of morphological information in lexical descriptions of a language, however, there has been relatively little research on questions related to this particular aspect of lexicographic work, in Slovenian and general lexicography alike. While research related to (paper-based) dictionaries for morphologically less complex languages mostly discusses how much morphological information should even be included in dictionaries, in addition to irregular morphological phenomena, and to what extent can regular morphological patterns be predicted by non-native dictionary users (Jackson 2002: 105−107; Honselaar 2003: 355−356; Caluwe and Taeldeman 2003: 73−77), research related to morphologically rich languages mainly focuses on the micro-structural issues of the optimal presentation of inflectional information, such as ways of abbreviating inflected forms or cross-referencing paradigmatic patterns, and the level of comprehensibility with regard to dictionary users (Vikør 2009: 140; Kola 2012). On the other hand, rather than ways of encoding morphological information, Slovenian linguistics has mainly been concerned with the question of its suitability from the viewpoint of (literary) language standardisation (see Toporišič 1971a and 1971b; Rigler 1971 and 1972).

Given the many possibilities that the online dictionaries bring to dictionary users and makers alike, the present paper aims to explore the prospects of describing and presenting morphological information in a born-digital dictionary of modern Slovene. We first perform an empirical analysis of the user needs for morphological information in Slovenian dictionaries (section 2), and investigate how these are met within the recently issued reference *The Dictionary of Slovene Literary Language, second edition* (section 3).[1] Given the general consensus of storing morphological information in the form of a separate machine-readable morphological database (morphological lexicon), we discuss the possible advantages of this approach for dictionary users (section 4.1), and also emphasize the need

---

1   Although this method is applicable to morphology in general, the remainder of this paper mostly focuses on inflectional information.

for a clear distinction between information in a morphological database and its presentation in a dictionary (section 4.2), as well as the distinction between a lexicon entry and a dictionary entry (section 4.3).

2 USER NEEDS

As a starting point for evaluation of user needs with regard to including morphological information in online language resources, this section presents an initial analysis of query log files of the Amebis inflection dictionary,[2] developed as one of the modules of the Besana grammar checking application (Holozan 2012). The demo version of this module is designed as an online dictionary portal that provides information on inflected forms (both standard and non-standard), derived forms and grammatical features of words or multi-word expressions entered by the user (Figure 1). The inflection dictionary is based on the ASES lexical database (Arhar and Holozan 2009), which is continuously developed and currently contains approximately 244,000 lexical entries.

Figure 1: An example of the Amebis Besana dictionary entry for the noun *gospa* **'lady/Mrs' (Slovenian interface only).**

---

2   http://besana.amebis.si/pregibanje/

Our analysis is based on an extensive log file for a six-year period from January 2009 to January 2015, which has been compiled as a two-column list of distinct query strings (words or multi-word expressions entered by the user) and the number of such queries. As can be seen in Table 1, 2,350,778 queries of 787,751 distinct query strings were recorded in this period. Thus, on average, more than 1,000 queries were recorded daily,[3] which confirms the significant need for this type of linguistic information by users, especially given that Besana is only one of several freely available online inflection dictionaries for Slovenian.[4]

Table 1: Number of queries within the Amebis Besana dictionary in the period 2009–2015.

Type of query string	Number of queries	Distinct query strings
Word	2,250,705	723,608
Multi-word expression	100,073	64,143
TOTAL	**2,350,778**	**787,751**

To gain a better understanding of which lexical items users investigate most frequently and in what way, we limited the subsequent qualitative analysis to query strings occurring in 300 or more queries. Even though these include only 571 distinct strings, they represent more than 25% of all queries (619,117 queries in total), which signals that speakers of Slovenian find the inflection of some lexical units significantly more problematic than others.

The results given in Table 2 show that these mostly include common nouns, such as *hiše*[5] (English 'houses'; 26,115 queries)*, otrok* ('child'; 22,164), *dan* ('day'; 15,488), *hči* ('daughter'; 14,046), *mati* ('mother'; 10,824), *gospa* ('lady'; 10,756), *človek* ('man'; 6,941), *tla* ('floor'; 6,006), *otroci* ('children'; 4,838), *vodja* ('leader'; 4,782), *pljuča* ('lungs'; 4,501), *vrata* ('door'; 4,408), *drva* ('wood'; 4,199), *oko* ('eye'; 4,034), *hiša* ('house'; 3,957), *dno* ('bottom'; 3,333), *pes* ('dog'; 3,296), *breskev* ('peach'; 3,032), *okno* ('window'; 2,991), and *leto* ('year'; 2,967). These are followed by verbs, such as *zvedeti* ('to find out'; 4,426)*, dati* ('to give'; 3,401)*, biti* ('to be'; 3,394)*, iti* ('to go'; 3,201)*, jesti* ('to eat'; 2,259)*, imeti* ('to

---

3   As a point of comparison, the online portal for the reference Slovenian orthography guide (http://bos.zrc-sazu.si/sp2001. html) recorded an average of 400 queries daily in the period from March 2010 to June 2015. In their overview of the frequency of usage for different online dictionaries, Bergenholtz and Johnsen (2005: 122–126) report on a range from a few hundred to a few thousand queries per day, for languages or language combinations with a considerably higher number of speakers than the two million seen for Slovenian.

4   A similar type of full paradigm querying is offered by the Sloleks morphological lexicon interface (available as part of the http://eng.slovenscina.eu/sloleks and http://www.termania.net portals), while abbreviated inflectional information is also included in most of the dictionaries produced by the Fran Ramovš Institute of the Slovenian Language (available as part of the www.fran.si, http://bos.zrc-sazu.si/ and http://termania.net portals).

5   The list of most frequent queries presented in this paper does not exclude queries suggested as demo queries or those used in system testing, such as *hiše* 'houses', *hiša* 'a house', or *Oselica* (name of a village).

have'; 1,736)*, vedeti* ('to know'; 1,474)*, moči* ('to be able'; 1,100)*, delati* ('to work'; 1,090)*, poslati* ('to send'; 1,076); pronouns, such as *on* ('he'; 4,325), *nič* ('nothing'; 3,518), *jaz* ('I'; 3,334), *ta* ('this'; 3,292), *ona* ('she'; 3,059), *kaj* ('what'; 2,473), *kar* ('which'; 2,205), *kateri* ('which'; 2,079), *ti* ('you'; 1,901), *moj* ('my'; 1,892); and proper nouns, such as *Oselica* (20,731), *Miha* (5,271), *Luka* (5,120), *Marko* (3,115), *Jaka* (2,310), *Žiga* (2,144), *Mitja* (2,046), *Grosuplje* (1,985), *Sašo* (1,722), *Klemen* (1,598). There is significantly less recorded queries for adjectives, e.g. *lep* ('beautiful'; 1,183), *nov* ('new'; 690), *dober* ('good'; 686), numerals, e.g. *dva* ('two-masculine'; 2,530), *tri* ('three'; 1,500), *dve* ('two-feminine'; 921), and adverbs, e.g. *lahko* ('easy'; 685), *dobro* ('well'; 593), *rad* ('gladly'; 562), which suggests users find these less problematic due to their regular inflectional patterns. The analysed list does not include any multi-word expressions, as even the most frequently queried multi-word unit (*dve leti* 'two years') does not reach the threshold, with only 241 queries in total.

Table 2: The list of most frequent queries per part-of-speech category in the Amebis Besana inflection dictionary.

	distinct queries	all queries
common nouns	336	398,658
verbs	71	53,633
pronouns	64	72,247
proper nouns	63	66,681
adjectives	14	6,878
numerals	10	9,003
adverbs	7	3,344
other	6	8,673
TOTAL	**571**	**619,117**

As expected, the most frequently queried strings include well-known words with irregular conjugation or declension patterns, which are also frequently discussed in language-related online forums (Dobrovoljc and Krek 2011; Bizjak Končar et al. 2011) and amongst the most common mistakes in student essays (Kosem et al. 2012a). On the other hand, our query log analysis reveals a surprisingly high number of queries related to seemingly unambiguous words, which inflect by regular patterns and have thus not been given any special consideration in existing language manuals so far, such as *avto* ('car'; 2,034), *mama* ('mother'; 1,578), *miza* ('table'; 1,565), *stol* ('chair'; 1,319), *fant* ('boy'; 1,070), *ura* ('clock'; 922), *knjiga* ('book'; 918); *delati* ('to work'; 1,090), *videti* ('to see'; 776), *hoditi* ('to walk'; 744), *govoriti* ('to talk'; 612), *dobiti* ('to get'; 494); *lep* ('beautiful'; 1,183), *nov* ('new'; 690), *prvi* ('the first'; 447), *star* ('old'; 368), and *zanimiv* ('interesting'; 309).

Even though the original log files lack other potentially relevant metadata on individual search queries, such as user ID, user demographics or look-up duration, which could give better insights into the user profile and the relevance of the obtained results (see for example the Wiktionary log files used in Müller-Spitzer et al. 2015), the results of this elementary query log analysis nevertheless illustrate there is a significant need to include inflectional and other morphological information in future lexical descriptions of Slovenian, and at the same time indicate this need is not limited to a closed set of well-known exceptions, but also includes lexical items with regular inflection.

3   MORPHOLOGICAL INFORMATION IN DSLL2

In the introduction section, the authors of the second, revised and partially updated edition of the *Dictionary of Slovene Literary Language* (DSLL2), the reference dictionary of standard Slovenian, describe the dictionary as a source of information on both semantic and formal properties of Slovenian lexica, since "*for each word, the dictionary explains how it is written and pronounced, what are its dynamic and pitch accents, how it inflects, what it means and what are the relations between individual meanings*" (Gliha Komac et al. 2014: 25, translated by K. D.). In both printed and online versions, DSLL2 continues the tradition of the first edition (DSLL, issued in 1970–1991), in which the information on inflection is presented as a combination of abbreviations in the dictionary entry, with instructions on how to interpret these in the dictionary's introduction. In order to access information on inflection of a lexeme, the dictionary users therefore first need to know how this information is encoded and then familiarize themselves with specific decoding instructions in the introduction section for their appropriate interpretation. In general, this can be described as a four-stage process consisting of (i) identification of the headword (DSLL2 Introduction: §27−§29), (ii) identification of the headword part-of-speech category (§30), (iii) decoding of the second/third basic form (§160−§165), and (iv) classification into the appropriate pattern for inflection and stress (§180−§196).

Although the initial phase of identifying the relevant headword seems relatively trivial, the results of the log file analysis presented in Section 2 show that users often query non-canonical word forms, which is why retrieving inflectional information from a dictionary should not be conditioned on comprehending the lemmatization principles used for headword selection. In addition to querying ambiguous inflected forms, such as *gospe* (inflected form of 'lady'), *hčer* (inflected form of 'daughter'), *dni/dnevi* (inflected forms of 'days'), *njih* ('them'), *matere* (inflected form of 'mother'), *brki* (plural form of 'moustache'), *starš*

(singular form of collective noun 'parents'), or *sabo/seboj* (instrumental form of 'oneself'), the list of most frequent queries in the Amebis Besana dictionary also includes words for which we can assume the users intended to enter an abstract canonical form, but chose the 'wrong' (non-standard) spelling, e.g. *imati* (instead of *imeti*, 'to have') or *pluča* (instead of *pljuča,* 'lungs'), or the 'wrong' (non-prototypical) grammatical features, such as number or gender, e.g. *psi* ('dogs' in plural), *smuči* ('skis' in plural), *dve* ('two' in feminine), *ona* ('she'), *vsi* ('everybody' in plural), *midve* ('us' in feminine dual), or *onidve* ('they' in feminine dual). With regard to DSLL2, these findings raise a particular concern with respect to its online version,[6] as looking up lexemes in a form different than the headword, such as an inflected form or a variant spelling, only gives results if the queried string appears as part of the grammatical information slot following the headword (e.g. there are no hits for querying *pluča*, the frequent non-standard spelling of the noun 'lungs'). On the other hand, adjusting the default settings to search through full dictionary entries, and not just the headword and its grammatical information, returns all dictionary entries containing the queried string, regardless of their relevance to the user (e.g. 78 dictionary entries for querying *psi*, the plural form of the noun 'dog').

Similarly, the second stage of identifying the part-of-speech category and other grammatical features of the headword needed for subsequent identification of the corresponding morphological pattern can also pose a challenge to non-professional users, as these can be given in different sections of the dictionary entry: either immediately after the headword, in the form of a qualifier with an abbreviation of the part-of-speech category or one of its features (e g. **finále** *-a m (â)* or **zanimív** *-a -o prid.*, where *m* denotes masculine noun and *prid.* denotes adjective), as part of the definition (**sêstrin** *-a -o (ē) svojilni pridevnik od sestra* 'possessive adjective of *sestra*' or **bíl²** *-à -ó in -ò opisni deležnik od biti sem (î ä ǫ)* 'descriptive participle of *biti*'), in the so-called qualifying explanation (**ánglo-** *prvi del zvez (â)* 'first part of phrases' or **si²** *členica* '~particle'), in a separate entry (**aloa** *gl. aloja* '*aloa* see *aloja*'), or this information is simply missing from the dictionary (**kamen...** *prim. kamn...* '*kamen…* see *kamn…*' and **kamn...** *prim. kamen...* '*kamn…* see *kamen…*'). After having identified the headword and the part-of-speech category of the lexeme of interest, the user should then consult the introduction section to find appropriate instructions on how to decode the abbreviated second or third (adjectives only) form listed next to the headword, e.g. *-a* in the *finale* example above. These instructions, however, demand a relatively high level of linguistic knowledge, which cannot necessarily be expected from non-professional users or non-native speakers, for example:

> Nouns and adjectival words are abbreviated in the following way: a) When the first word form ends in a consonant, the second word form is formed

---

6   http://www.sskj2.si

by adding the given part of the second word form, consisting of a vowel or *j, n* + vowel, to the first word form /…/. The second word form is formed in the same way, when the given part of the second word form is –*ih* /…/. If a longer part of the second word form is given or the ending is preceded by a consonant (other than *j, n*) due to changes in endings, the given form indicates which part of the headword it applies to /…/. b) When the first word form ends in a vowel, the second word form is formed by adding the given part of the second word form, consisting of *j, t, n* + vowel, to the first word form /…/, or the last vowel of the first word form is omitted, if the given part of the second word form begins with a vowel /…/. In the same way, the second word form of a noun ending in *-ega*, which is otherwise inflected by adjectival declension /…/. If the given part of the second word form begins in a consonant (other than *j, t, n*), the given form indicates which part of the headword it applies to /…/. (DSLL2 Introduction: § 161, translation by K. D.)

Another potential issue in the comprehensibility of morphological information in DSLL2 is information on inflection of the so-called cross-referencing headwords, pointing to an entry with a more standard-like spelling of the headword, when the two headwords do not inflect in the same way. For example, the entry for the word *croquis* (**croquis** *gl. kroki* 'croquis see croquis' points to the dictionary entry of it spelling variant **krokí** *-ja m (î)*), but the two headwords have different inflectional paradigms (e.g. *kroki+ja* vs. *croquis+a* in genitive singular). What is more, some dictionary entries also lack the abbreviated second word form needed for subsequent inflection pattern deduction, such as **múlda** *ž (û) jarek za odtok tekočine s ceste, tlakovanih površin* or **rímokatoličánka** *ž (î-â) pripadnica rimskokatoliške vere*.

In the last stage of the inflection deduction process, users then use the combination of the headword and its un-abbreviated second or third form(s) to select the appropriate governing scheme for inflection and stress (Figure 2) and its specific subtype, which also requires some knowledge of linguistic terminology (e.g. *base/ending stress, stress on different base syllables, short/long stress* and so on), consideration of exceptions and modifications signalled in footnotes, and understanding the meaning of special symbols, such as the symbol ~ (denoting the formation of the inflected form based on the nominative or infinitival base form or part of the base form), the symbol – (denoting either formation based on genitive or present base form or part of the base form, or the nominative masculine or feminine form for adjectives), and the symbol ' (denoting the place of stress).

Figure 2: An example of a governing scheme for inflection of nouns by first masculine declension (with footnotes) in DSLL2 Introduction.

The selection of an appropriate pattern can also depend on other inflected forms given in the dictionary entry (in addition to the default headword form and the abbreviated second/third form), but not always, as these can also signal particularities of an individual part-of-speech category or word forms that the lexicographer deemed to be potentially ambiguous, without any influence on the pattern deduction process (DSLL2 Introduction: §184–185), although the dictionary does not specify how users can distinguish between these competing interpretations. Similarly, a full list of word forms is given for headwords that cannot be placed in one of the patterns in the Introduction, as illustrated in Figure 3.

Figure 3: Full inflection paradigm given at the beginning of the dictionary entry for the pronoun *on* 'he' in the online version of DSLL2 (small font denotes qualifiers).

Although DSLL2 is considered to be the reference manual for inflectional and other morphological information on Slovenian lexica, it seems this purpose is not achieved in an optimal way. The four-stage inflection pattern decoding process presented above presents a challenge for dictionary users, requiring them to combine specific information from the dictionary entry and general instructions from the terminologically challenging introduction section.[7] Although such a method of encoding morphological information is understandable given the practical limitations of the paper-based first edition of DSLL, it is less justifiable in its second edition, published more than 40 years later in both print and online versions, especially given the fact that language professionals themselves pointed out the difficult decoding of inflection, stress and pitch patterns in the DSLL2 planning discussions (Perdih 2008: 18, 136, 142–143).

4   MORPHOLOGICAL LEXICON AS A COMPONENT OF AN E-DICTIONARY

The fact that a born-digital online dictionary enables a new approach to describing and presenting morphological information for Slovenian has first been recognized by the authors of the recent "Proposal for a Dictionary of Modern Slovene" (Krek et al. 2013b), who suggest storing morphological information as part of a separate database, an enhanced version of the Sloleks reference morphological lexicon of Slovenian language (presented in the chapter by Dobrovoljc et al. in this book), and visualising it in a separate section of the dictionary entry (the so-called *Inflection* tab). A similar solution has also been proposed by the authors of the "Draft Concept of the New Dictionary of Slovene Literary Language" (NDSLL; Gliha Komac et al. 2015) who speak of a lemmatization database with information on the formal properties of lexical units that would be displayed as part of the *Pronunciation and inflection* section of the online dictionary entry.
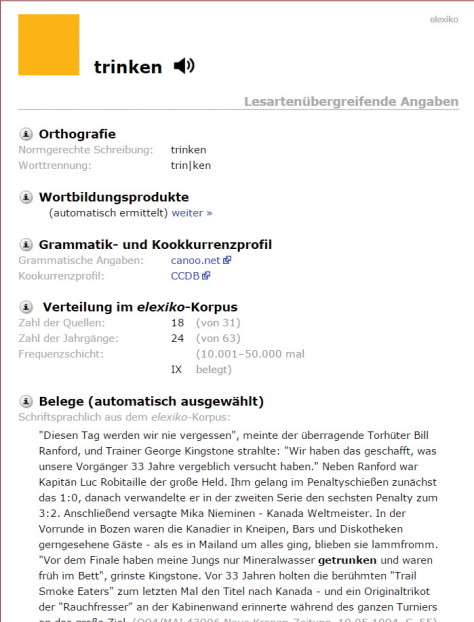
This idea of a separate, but integrated machine-readable morphological database (a morphological lexicon) has several advantages for both the quality of dictionary information and the resulting user experience (as discussed in section 4.1), but it also raises new questions on the relation between information stored in the lexicon and that shown in the dictionary (4.2), and the relation between the lexicon and the dictionary entry (section 4.2.).

---

7   This process is particularly problematic with respect to students and non-native speakers, who are believed to look up regular forms and patterns more often, since regular forms can only be decoded from the abbreviated patterns in the Introduction section, in contrast to irregular forms that are usually given in the dictionary entry itself.

4.1 Morphological lexicon as a source and navigator of dictionary information

In the context of building a future dictionary of modern Slovene, a machine-readable morphological lexicon fulfils two distinct roles. On the one hand, it is used as a key component in the development of different NLP applications for grammatical annotation of corpora and subsequent lexical data extraction (see the chapter by Erjavec et al. in this book). On the other hand, a morphological lexicon presents the primary source of information on the formal characteristics of lexical units in a dictionary, such as information on their part-of-speech category and other morphosyntactic features, or information on their inflection, derivation and pronunciation.

In related born-digital online dictionaries for other languages, inflectional paradigms are usually presented in full and without abbreviated forms, either through a hyperlink to an external dictionary of inflected forms (as with the Icelandic ISLEX multilingual online dictionary, the BFL lexical database for French or the Elexiko dictionary portal for German illustrated in Figure 4), or as part of the dictionary entry itself. For morphologically less complex languages, the latter solution usually includes listing inflected forms, pronunciations and related grammatical information

Figure 4: An example of hyperlinked morphological information in the *Elexiko* German dictionary for the verb *trinken* ('to drink', left) pointing to its conjugation paradigm in the Canoo morphological lexicon (right).

in the primary-level vicinity of the headword (as with the Collins English dictionary for learners or the ANW scholarly dictionary of contemporary standard Dutch), while other languages place this information on a secondary level accessed by clicking on an additional button or tab (as with the DAELE Spanish Learners' Dictionary or the Great Dictionary of Polish illustrated in Figure 5).

Figure 5: An example of embedded morphological information in the Great Dictionary of Polish for the noun *sadzonka* ('a seedling') in the *Odmiana* (Inflection) section of the dictionary entry.

Given that the Sloleks morphological lexicon is planned to include both standard and non-standard basic and inflected forms, labelled with the corresponding variation type and its compliance with the language norm (see chapter by Dobrovoljc et al. in this book for a detailed description), the morphological lexicon thus also functions as the pivotal source of information on potential spelling, pronunciation, inflection, derivation, syntactic or other issues related to individual lexical items. In addition to the lexicon providing the lists of all variant forms or pronunciations, their classification by specific variation type also allows for automatic selection and display of the relevant language issue explanation(s) in the norm-related section of the dictionary entry.[8] Using the same mechanism, specific tags

8 The norm-related section of the dictionary, proposed by Krek et al. (2013: 41), is designed as a style guide with user-friendly explanations of language issues in Slovenian. The explanations are based on the ontology of most frequent types of linguistic issues in Slovenian (Krek and Dobrovoljc 2011), and are thus designed as a set of universal explanations to be displayed with all lexical items related to a certain type of issue.

or notifications can be automatically displayed in different parts of the dictionary entry (for example, next to the headword or one of its variant spellings; next to a particular word form or pronunciation etc.) to alert users about specific issues or particularities and direct them to the related explanations.

In addition to being the source of morphological, grammatical and normative information, the morphological lexicon also has an essential role in displaying other types of dictionary information. It enables searching by all possible forms and spellings and therefore allows users to form intuitive search queries without having to consider the lemmatization, part-of-speech categorization and spelling principles used in the dictionary headword selection, as is currently the case with the reference DSLL2 dictionary and the Fran dictionary portal.[9] In a similar way, a morphological lexicon can enhance the comprehensibility of definitions by linking individual word forms with the relevant lexical units (see for example the hyperlinking mechanism in the definitions of Wiktionary and TheFreeDictionary), or by linking the dictionary to external language resources and tools, as in the case of the Sloleks web service,[10] where clicking on a particular word form or lemma takes the user to the list of all relevant concordances in the reference corpus (i.e. usages of the word form in context). Similarly, information on phonetic transcription in the background lexicon enables machine-generated speech synthesis of displayed word forms on the one hand, and automatic speech recognition of voice search queries on the other.

4.2   Relation between lexicon data and dictionary information

Despite the many technical and content-related advantages of keeping morphological information in a separate database, we must distinguish between original data in the lexicon database on the one side and the user-oriented dictionary information on the other, when planning its visualisation. One of the main advantages of a hierarchically-organized machine-readable system is the fact that it enables dynamic adjustments of information visualisation with respect to the type of language manual or the specific needs of its users. These include not only graphical design and technical solutions, but also the selection of the displayed information itself, such as the inclusion of data on non-standard language use, pronunciation or specific grammatical information, discussed below.

---

9   For example, the log file analysis of the Danish *Den Danske Netordbor* online dictionary (Bergenholtz and Johnsen 2005: 127–133) shows that the 19.5% of unsuccessful searches mostly include the passive and imperative forms of verbs, misspellings, spelling mistakes affected by pronunciation, and mistakes in writing multi-word expressions as one or several words.

10   http://eng.slovenscina.eu/sloleks

Most existing dictionaries of the Slovenian language include information on both standard and non-standard inflected forms. However, the latter are usually limited to a closed set of most common orthographical and morphological exceptions, such as the declension of nouns *otrok* 'child', *mati* 'mother', *hči* 'daughter', *gospa* 'lady/Mrs', and so on. A usage-based morphological lexicon, compiled to give an exhaustive description of formal characteristics of Slovenian lexica, would also include all frequent variant irregular patterns and modifications, such as the non-standard phoneme additions in declension, sound changes, etc. Experience in visualising the Sloleks morphological lexicon, which already includes several demo instances of such variant paradigms, shows that users prefer to see standard paradigms written in full, regardless of the frequency of usage of individual inflected forms, whereas the addition of full non-standard paradigms is too difficult to process, so the visualisation of these should be reduced to individual inflected forms occurring in corpus data. One of the first priorities of future user-experience research is thus to determine the frequency threshold, below which displaying non-standard language usage information no longer plays an informative or educational role, but instead acts as a disruption in the overall comprehensibility of the given information, regardless of the graphic design solutions.

A similar issue arises when visualising information on pronunciation, as the high frequency of stress placement variants in the Slovenian language results in extensive pronunciation paradigms; if we augment these by the alternative pronunciations of particular phonemes or different types of pronunciation transcriptions (accentuated or unaccentuated word forms; standardised or customized phonetic transcription), the display of all the combinatorics of all possible word forms quickly becomes overwhelming. It is thus important to prioritize pronunciation information according its relevance to dictionary users, for example, show the accentuated headword and its phonetic transcription by default, but embed other phonetic information, such as the full accentuated inflectional paradigm or its phonetic transcription (only rarely found in general dictionaries), on a secondary level accessed in a separate section or a special extension button next to the default, unaccentuated inflectional paradigm.

The third important aspect to consider when distinguishing lexicon data from dictionary information is the visualisation of grammatical information. The formal grammar used in the compilation of a morphological database, usually adjusted to meet the needs and limitations of automatic natural language processing, is not necessarily equivalent to the grammar description given in a general dictionary. In addition to terminological considerations, such as renaming particular grammatical features that might be less comprehensible for non-linguistic users (e.g. non-definiteness or biaspectuality), and an evaluation of their actual relevance for the user, this also includes elemental linguistic decisions on the inventory of part-of-speech and other morphological categories, as well as the criteria for their

selection with particular lexical items.[11] Using different approaches for different types of lexical databases is not problematic in itself, but it is important that the specific mappings between the two are systematized and well-documented, as this is a prerequisite for the full compatibility of fundamental language resources, such as a morphological lexicon, a lexical database or a dictionary, and their long-term usability in other language resources and tools.

4.3   Relation between lexicon and dictionary entry

Since a morphological lexicon is primarily intended to store information on the inflectional, derivational, normative and other morphological properties of lexical items, and not their semantic characteristics, lexical items with identical morphological, phonological and grammatical features are usually merged into one lexicon unit, regardless of potential differences in meaning. In this way, the Sloleks lexicon merges homonymous semantically distinct lexical items, such as *bor* ('pine tree') and *bor* (the chemical element), or *početi* ('to start') and *početi* ('to do'), into a single lexicon entry, while semantically equivalent, but formally different lexical items, such as *volivec* and *volilec, posebej* and *posebaj, zvedeti* and *izvedeti* (two different spellings of 'a voter', 'especially' and 'to find out', respectively), are separated into two or more distinct lexicon entries. When integrating a morphological lexicon into a general dictionary or its underlying lexical database, it should thus be remembered that given the different designs and purposes of both databases the relation between the lexicon and dictionary entries is not necessarily symmetrical nor static, as it primarily depends on the consensually defined criteria on what constitutes the basic unit (an entry) in each database.

One of the key dictionary design decisions, influencing the way the lexicon and the dictionary database inter-connect, is undoubtedly the selection of formal criteria used for distinguishing homonymy from polysemy. That is, defining what formal properties of two semantically distinct lexical units with identical spellings should be considered when deciding whether to describe them in separate dictionary entries (homonymy), or within the same dictionary entry with two or more different meanings (polysemy). According to Gantar (2015: 341), both theoretical and user-orientated lexicographical approaches to the issue of homonymy-polysemy distinction usually agree that differences in one of the following formal characteristics should be considered as sufficient criteria for homonymy to be chosen, regardless of the degree of semantic or etymological similarity between the two items: homographs belonging to different

---

11 An example of such differences in grammatical information in the morphological lexicon on the one side and a general dictionary on the other, would be potential merger of adverbial participles (currently stored as adverbs in Sloleks) with their original verbs, or elatives (currently stored as separate entries in Sloleks) with other degrees of comparison, etc.

part-of-speech categories (e.g. the noun and the adverb *naglas* 'accent/loudly', the noun and the adjective *žužkojed* 'insectivore/insectivorous'); homographs with different grammatical features (e.g. the masculine and feminine noun *prst* 'finger/soil', the masculine and neutral noun *čelo* 'cello/forehead'); homographs with a different inflection (e.g. the imperfective verbs *vesti-vezem* and *vesti-vedem* 'to embroid/to behave' or the perfective verbs *postati-postanem* and *postati-postojim* 'to become/~to pause'); or homographs with different pronunciations (e.g. *molíti-molím* and *móliti-mólim* "~to hand out/to pray' or *partíja-partíje* and *pártija: pártije* 'a political party/a match'). This is also in line with the entry selection criteria used in the Sloleks morphological lexicon, which separates all these items into two distinct lexicon units – the relationship between the lexicon and the dictionary entry is thus symmetric.

However, there is less lexicographic consensus on whether the formal properties to be taken into consideration also include: the differences in derivation (e.g. the homonymous noun *vila* 'a villa/a fairy', where the derived adjective *vilinski* is only associated with the second of the two meanings); the differences in part of the inflectional paradigm (e.g. the homonymous adjective *bučen* 'loud/of-pumpkin', where the comparative forms are only associated with the first of two meanings, or the homonymous noun *lisica* 'a fox/handcuffs', where the second meaning is only associated with plural forms); or the differences in specific inflected forms (e.g. the homonymous noun *tenor* 'the voice/the singer', where the two items only differ in the singular accusative form that depends on animacy). Given that the Sloleks lexicon has also been designed to be used in natural language processing applications, which are not yet capable of reliable semantic disambiguation of identical inflected forms with identical grammatical features (e.g. disambiguating the form *bučnega, lisic* or *tenorja* in all different possible meanings), the lexicon thus follows the principle of the maximum possible paradigm that merges such overlapping inflectional paradigms into a single lexicon entry, even if specific meanings only take on a limited subset of all possible forms. Regardless of whether or not these meanings are separated into independent entries in the dictionary, the relationship between the lexicon entry and the dictionary entry is thus inherently asymmetric, since a particular dictionary headword or one of its meanings only correlates with a subset of a certain lexicon entry (e.g. *lisica*, *bučen*, and *tenor*). This potential asymmetry of interlinked database entries should thus be given special consideration when designing the technical and visualisation solutions for an online dictionary.[12]

If the previous paragraph discusses relating one lexicon entry to several dictionary entries or meanings, it is equally relevant to address the issue of relating one

---

12 One possible solutions on how to display meaning-dependent morphological information, can be observed in the Great Dictionary of Polish, which considers all homographs as polysemous items, regardless of their diachronic connection, but requires the users to select the meaning of interest before displaying any additional information on grammatical properties or inflection.

dictionary entry to several lexicon entries. A typical example of this kind of database asymmetry are lexemes with variant spellings, e.g. *žiroračun* and *žiro račun* 'a giro account', *eventuelno* and *eventualno* 'possibly', *volivec* and *volilec* 'a voter'. The Sloleks lexicon stores these as distinct lexicon entries, whereas a dictionary usually considers them to represent the same lexical item if no semantic differences are observed, merging their description into a single dictionary entry with several spellings and related inflection or pronunciation paradigms. A similar issue arises when describing (potentially) semantically identical pairs of homographs with variant grammatical lexical features, as with *činčila* 'chinchilla', *sluz* 'slime' or *nadlaket* 'upper arm', which are used both as masculine or feminine nouns, or with *finale*, which is used both as masculine or neutral noun, without any change in meaning. Even if a dictionary considered these to constitute separate dictionary entries (i.e. in symmetry with the lexicon), displaying information for several lexicon entries within a single dictionary entry is nevertheless inevitable for lexemes with multi-gender inflections, as for example the neutral noun *oko* 'eye' that takes the feminine plural form *oči* in one of its meanings, or the feminine noun *ledvica* 'kidney' that takes either the feminine (*ledvice*) or the neutral (*ledvica*) plural form.

5   CONCLUSION

Both lexicographic tradition and empirical user research confirm that morphological information represents an indispensable part of lexica description in a general dictionary. In order to meet this information need, it seems that future dictionaries of the Slovenian language should break with the tradition of presenting inflectional information in the abbreviation system that was created for the print-based design of the first edition of DSLL, given the limited degree of comprehensibility and the general technological advances that have taken place over the past few decades. With respect to the rich morphology of the Slovenian language, keeping this information in the form of a separate database brings many advantages to dictionary makers and users alike. However, its integration into a dictionary must be designed and implemented in a systematic way, so as to ensure the dictionary's long-term compatibility with other language resources and tools, and to enable its dynamic adjustment to meet the varying needs of diverse user groups.

The Sloleks Morphological Lexicon and its Future Development

Kaja Dobrovoljc, Simon Krek and Tomaž Erjavec

Abstract

This paper presents Sloleks, the largest open-source machine-readable morphological lexicon of the Slovene language to date. We first briefly present its development and the formal grammar behind it, and then provide a detailed presentation of the types and structure of inflectional, derivational, grammatical and other included information, with a special emphasis on its formal representation within the standardized XML LMF framework. Given that Sloleks is a strong candidate to be used in the compilation of a new dictionary of modern Slovene, both as a source of morphological information and as a background resource for the language technology tools needed to create it, the second part of the presentation explores the most important aspects of its future development, in particular the expansion of its entry list, addition of pronunciation information, normative categorization of variants and a corpus-based re-evaluation of the existing inflectional paradigms. Such an extensive usage-based open-source morphological lexicon of modern Slovene with a unified system of morphological description will have a long-term use for both language technologies and for other born-digital reference works for the Slovene language.

Keywords: morphological lexicon, lexicon of inflected forms, machine readable dictionary, morphology, inflection, derivation, pronunciation, language standardisation

1   INTRODUCTION

When it comes to morphologically rich languages, such as Slovene, the description of morphological paradigms of inflected parts of speech is traditionally very important. For example, the first Slovenian grammar (Bohorič 1584) dedicates almost half of its content to word inflection, and morphological paradigms have a similarly prominent role in most of the later Slovene grammars. These mainly focus on systemic aspects of morphology, i.e. morphological patterns which they illustrate by means of examples. This in turn means that explicit, complete paradigms in grammar books are few and far between. On the other hand, dictionaries from the pre-digital age, mainly different orthography guides and later DSLL (*The Dictionary of Slovene Literary Language*), fulfilling their role as lexical enumerators, also contained data on inflection. The morphological descriptions in these reference works are significantly shortened; in addition to the headword, they are usually limited to one or a few inflectional forms, which are supposed to provide the user with enough information to deduce the entire morphological paradigm. Even when printed reference books were digitized, the data stayed the same.

The arrival of computers and advances in natural language processing soon established a need for accessible machine-readable dictionaries and lexicons of inflected forms (Atkins and Zampolli 1994). The first English machine-readable dictionaries designed for various language technology tasks were already designed in the 1960s (e.g. Boguraev and Briscoe 1987); the widespread digitization of languages in the 1990s, however, also paved the way for the creation of morphological lexicons for most other European languages.

Computers cannot work with only a pattern or a few word forms, which is why these lexicons – free from the space constraints imposed by the printed medium – typically contain paradigms written out in full and available in a machine-readable format. Morphological data, traditionally targeted at users of printed language reference books, were therefore given a new field of application, where the new "user" is the computer itself. Lexicons must therefore fulfil language technology needs in various computer applications – from spellcheckers and part-of-speech taggers to parsers, speech synthesizers, and machine translation software – and be simultaneously useful as independent morphological reference tools for language users. The contemporary machine-readable lexicon of the Slovene language should therefore fulfil both needs, and thus needs to be organized differently than morphological data in dictionaries and grammar books or the first computational lexicons.

In pursuing these two goals, the compilation of such lexicons stumbles upon two contradictory tendencies: when dealing with language technology applications,

the lexicon must be capable of representing the morphological characteristics of all the word forms present in authentic texts, including spoken discourse, allowing for simple machine processing of the data. However, when it comes to traditional usage, it must also provide effective information on inflection, pronunciation, and word derivation relevant for a human user, including normative aspects of the vocabulary. In the context of integrating a lexicon into a future dictionary of modern Slovene, the lexicon's content must be aligned with both poles: on the one hand with the morphological data produced by morphological taggers to automatically annotate text corpora (the data source of the dictionary), while also making sure that the lexicon aligns with the data in the lexical database used as the source of the dictionary.

When it comes to fulfilling the user needs associated with language reference books, the key problem in creating the reference morphological lexicon of modern Slovene lies in the fact that the existing language reference grammar books (e.g. Toporišič 2004), dictionaries (e.g. DSLL2) and normative guides (e.g. SP 2001) are not on the same page when it comes to examining morphological data, and at times even contradict one another (cf. Krek 2014a). This means that none of this work can be taken as a starting point – the whole concept needs to be redesigned from scratch. Additionally, these reference books were not created based on modern language data, meaning they are relatively detached from the linguistic reality of modern Slovene, although this is important for users of language reference books and for language technologies.

Computational morphological lexicons for Slovene have a relatively long history. At the start of the 1990s, the Amebis company started developing ASES, an electronic dictionary of the Slovene language, which also contains explicit morphological paradigms (Arhar and Holozan 2009). This database itself is not freely available; however, the data it contains may be found in various products the company offers, such as the Besana grammar checker, the Presis machine translation software, its system for natural language communication, and so on. Chronologically speaking, the first freely accessible computational lexicon of Slovene was created in the framework of the MULTEXT-East project in the 1990s. It contains over 15,000 lemmas and their inflectional paradigms in a tabular format (Erjavec et al. 1995).

During the first decade of this century, the development of speech technology (mainly speech synthesis) raised the importance of lexicons which – in addition to morphological data – also contain information on pronunciation, such as SIflex, SImlex (Rojc et al. 2002; Verdonik et al. 2002), LC-STAR (Verdonik et al. 2004; Verdonik and Rojc 2004), SI-PRON (Žganec Gros et al. 2006). The chief problem with all these lexicons lies in the fact that they are not freely available. The same goes for the morphological lexicon created during the same period at

the Fran Ramovš Institute of the Slovenian Language. There are in fact no data available on this lexicon, apart from the fact it exists (Naglič et al. 2005: 36).

A slightly more specific lexicon is available through the freely accessible machine translation system called Apertium; it contains just over 20,000 lemmas (Horvat and Vičič 2012; Vičič 2012). Even though it is basically derived from the MULTEXT-East lexicon, its content and format is somewhat different, since it is mainly used in the context of a translation system, and is therefore not useful as a general morphological lexicon for Slovene. Within the recently completed "Communication in Slovene" project, the morphological lexicon Sloleks (Dobrovoljc et al. 2013) was created. This is also the central subject of this chapter – because due to its size, accessibility, and use in Slovene language technology tools, it represents a logical stepping stone for the further development of a reference morphological lexicon for Slovene.

2   THE SLOLEKS MORPHOLOGICAL LEXICON

The following sections describe the content of the Sloleks morphological lexicon and its format, the types of data it contains and their organisation within an individual lexicon entry, and the design of its online interface.

2.1   Content

2.1.1 *Lemma list and paradigms*

The current version of Sloleks (Dobrovoljc et al. 2013) includes 100,805 entries, where an entry includes the basic form (the lemma) of the word, its inflected forms (the inflectional paradigm) and related morphological information. The list of headwords or lemmas has been compiled based on criteria set out in the guidelines for its construction (Erjavec et al. 2008), by first including the majority of lemmas occurring in the manually annotated ssj500k corpus (Krek et al. 2013c), all lemmas belonging to closed part-of-speech categories (prepositions, conjunctions, pronouns, particles) and a pre-selected list of morphological particularities, such as foreign proper names, homonymous verbs with identical lexical features and different inflections (e.g. *stati* 'to stand/to cost'), masculine nouns that inflect for (in)animacy in accusative singular (e.g. *delfin* 'a dolphin/ the butterfly stroke'), lemmas with irregular or variant inflections (e.g. *a child*), and so on. The remaining and majority of the lemmas were then selected from

the list of most frequent lemmas in the then reference corpus of written Slovenian FidaPLUS, containing 620 million words (Arhar and Gorjanc 2007).

In the second stage of Sloleks compilation, lemmas were assigned their inflected forms using a program for semi-automatic paradigm generation, developed by Amebis d. o. o. for the construction of the ASES lexical database (Arhar and Holozan 2009) and related languages tools. The Sloleks morphological lexicon thus includes almost 2,800,000 inflected forms, with a quantitative description per part-of-speech category given in Table 1.

Table 1: Number of lemmas and inflected forms in the Sloleks morphological lexicon v1.2.

Part-of-speech	Number of lemmas	Number of inflected forms
nouns	54,260	924,268
adjectives	26,612	1,571,970
verbs	10,242	260,826
adverbs	6,906	9,931
numerals	2,240	18,448
pronouns	169	6,182
prepositions	96	101
interjections	85	85
abbreviations	70	70
particles	68	68
conjunctions	54	54
multi-word units[1]	3	3
TOTAL	**100,805**	**2,792,006**

2.1.2 JOS Annotation Scheme

Grammatical information in the Sloleks morphological lexicon is based on the morphosyntactic specifications developed within the "Linguistic Annotation of Slovene" (JOS) project (Erjavec and Krek 2008)[2] aimed at annotating corpora to be used in human language technologies for Slovenian. The JOS annotation scheme is based on previous projects dealing with formal grammars of Slovenian, in particular the MULTEXT (Ide in Véronis 1994) and MULTEXT-East projects (which includes most Slavic languages), with the Slovenian MULTEXT-East 4.0 specifications being identical to the JOS specifications.

---

1   Multi-word entries in the current version of the lexicon have been included as part of its demo integration into the *Slogovni priročnik* online style guide (Krek et al. 2013a).

2   http://nl.ijs.si/jos/index-en.html

JOS specifications include 12 part-of-speech categories: noun, adjective, verb, adverb, pronoun, numeral, preposition, conjunction, particle, interjection, abbreviation and residual, with the latter not being used in the lexicon. With the exception of particles, interjections and abbreviations, most part-of-speech categories incorporate additional grammatical features, however, not all items belonging to a particular part-of-speech category necessarily display all possible features. The list of all possible combinations of part-of-speech categories, morphological features (attributes) and their values is given in the form of a precompiled tagset[3] containing 1,902 morphosyntactic tags, while specific guidelines for their assignment to words in context are described in the corresponding annotation guidelines (Holozan et al. 2008).

As Erjavec et al. explain in more detail in their chapter in this volume, the JOS morphosyntactic specifications have primarily been developed to facilitate the development of human language technologies for Slovenian, and thus sometimes differ from the traditional grammatical descriptions given the limitations of automated natural language processing applications (Ledinek 2014a: 34−48). It is thus usually the form of a word that influences its part-of-speech classification, rather than its syntactic function. A typical example of this principle are participles ending in -*n*, -*t*, or -*č*, which are always annotated as participle adjectives, regardless of their attributive (*ukradena denarnica* 'a stolen wallet') or predicative (*denarnica je bila ukradena* 'the wallet has been stolen') syntactic role. Similar simplifications have also been implemented with specific morphological features, where, for example, the *person* feature is assigned to present tense verbs (even if they are impersonal, e.g. *dežuje* 'it rains'), and the *definiteness* feature is assigned to all adjectives (even if possessive adjectives do not inflect for definiteness).

Implicitly, through the process of manual corpus annotation and compilation of the morphological lexicon, the JOS annotation guidelines also specify the basic principles for determining the base form (lemma) of inflected word forms. These principles mostly conform to the general lemmatization principles used in other existing Slovenian language resources, e.g. selecting the nominative singular for nouns, infinitive for verbs, positive indefinite masculine singular for adjectives or word numerals, and positive for adverbs, with a few irregularities.[4] The only exception are pronouns, for which the lemma depends on the type of pronoun and its lexical features (e.g. lemmas *vame, zame, čezme* etc. for accusative bound personal pronouns inflected for number, person and gender; or the lemma *se* for reflexive personal pronominal forms *sebe/se, sebi/se, sabo/seboj*).

---

3   http://nl.ijs.si/jos/msd/html-sl/msd.index.msds.html

4   For example, lemmatization with nominative plural for *pluralia tantum* nouns (*alimenti* 'alimony') or the only possible form (e.g. the noun *poštev* 'account' that is only used in accusative singular as part of the multi-word expression *priti v poštev* 'to take into account'). With adverbs, the comparative (*bolj, manj, prej, raje, več, večkrat*) and superlative (*najbolj, najmanj, najprej, najraje, največ, največkrat*) forms of some adverbs represent separate lexicon forms with separate lemmas due to their specific syntactic roles.

2.2   Format

To ensure wide usability of a costly language resource, such as the reference collection of inflectional, derivational and other morphological information about the Slovenian language, it is essential to publish it as an open-source resource and encode it in a standardized way that enables flexible data organisation, as well as data comparability across databases and languages.[5] The Sloleks morphological lexicon is thus encoded as an XML document using the Lexical Markup Framework (LMF) scheme, an international standard for encoding natural language processing lexicons and machine readable dictionaries (ISO 24613:2008), developed as a common model for the creation and use of mono- and multi-lingual lexical resources, to manage the exchange of data between and amongst these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources (Francopoulo et al. 2006: 1).

The LMF format consists of two main types of components, the *core package* and the *extensions* of the core package. The core package defines a structural skeleton, which describes the basic hierarchy of information in a lexical database, such as information on the language, the name and accessibility of the resource (the metadata of the lexicon), as well as information on the basic structure of a lexical entry, whereas extensions give further specifications on how to combine the core package components with additional components required for a specific lexical resource, such as a morphological lexicon.[6]

The adjustment of the LMF format for standardised encoding of morphological lexica for morphologically rich languages, which has been used as the basis for encoding Sloleks, is explained in Krek and Erjavec (2009), while the full list of possible XML elements, attributes and values, together with the description of their hierarchical structure, is given in the corresponding Document Type Definition (DTD) intended for the validation of the lexicon structure.

2.3   Lexicon Entry

The basic building block of Sloleks is the lexicon entry.[7] One lexical entry consists of the lemma and its inflectional paradigm, i.e. the full list of one or more

---

5   The first open-source morphological lexicons were encoded in a tabular format, which is inconvenient for storing information on variant inflected forms or pronunciations, and their complex relationships with other types of information.

6   While extensions define the expected types of information in a particular lexical resource type, their number and hierarchal organisation, they do not define their semantic content, as the standardised sets of categories used for linguistic descriptions, such as the standardised names of part-of-speech categories, features and values, are defined by the ISOcat Data Category Registry (http://www.isocat.org/).

7   Although the term 'lexical entry' is used more frequently, we use the term lexicon entry to differentiate entries in a morphological lexicon from those in other types of lexical databases with prevailing semantic information.

inflected forms with corresponding grammatical information. By default, each lexicon entry includes information on lemma, its part-of-speech category and at least one inflected word form,[8] while an optional array of additional inflected forms and other morphological information is added depending on the part-of-speech and lexical features of the lemma. In the following section, we briefly present the types of morphological information found in Sloleks, their hierarchal organisation and their XML LMF exemplification.

2.3.1 Entry Key

The lexicon entry key is defined as a unique identifier used for distinguishing individual lexicon entries, since a particular lemma (the headword) can appear in several lexicon entries, either with different part-of-speech categories (e.g. the adverb and the particle *ravno* 'straight/just', the adverb and the noun *stran* 'away/ page', the adverb and adjective *spet* 'again/tied') or within the same part-of-speech category (e.g. the perfective and imperfective verbs *zlagati* 'to lie/to fold', the participial and common adjective *poročen* 'married/marital', the feminine and masculine noun *prst* 'soil/finger'). Even though the entry key is primarily intended for machine processing purposes and not end-user visualisation, it is nevertheless designed so as to encode information on the part-of-speech category abbreviation and the lemma (a talking code), e.g. *S_avtomobil* for the noun 'car'. Whenever there are several identical lemmas within a part-of-speech category, an additional number identifier is added, e.g. *G_vesti_1* for the verb 'to embroid' and *G_vesti_2* for its homonymous verb 'to behave'.[9]

```
<LexicalEntry id="LE_ebc318126ea71205d05cd0ce85f86362">
        <feat att="ključ" val="R_pazljivo"/>
```

Figure 1: The entry key of the adverb *pazljivo* 'carefully' in the XML LMF format.

2.3.2 Lemma

The pivotal element of a lexicon entry to which all other types of morphological information within an entry attach is the lemma, or the entry headword. In the

---

8   In this paper, the terms inflectional paradigm and inflected word form are also used to describe one-word paradigms of non-inflecting part-of-speech categories, such as prepositions, as they are formally encoded in the same way.

9   Masculine and feminine pairs of surnames form a special category, as their entry key consists of information on gender instead of a number, e.g. *S_Novak_m* for male surname and *S_Novak_ž* for female surname. When a surname is homonymous with another noun of the same gender, the respective entry keys are extended by an additional number identifier, e.g. *S_Pavlica_ž_1* (for the indeclinable female surname *Pavlica*, and *S_Pavlica_ž_2* for the declinable female name *Pavlica*).

Sloleks morphological lexicon, the lemma is defined as the abstract canonical or citation form of a lexical item that unites all inflected forms with the same lexical and formal properties, and usually also the same meaning. The principles for determining entry headword in Sloleks follow the JOS lemmatization principles used in manual lemmatization of the training corpus ssj500k (Holozan et al. 2008) and the development of a data-driven morphosyntactic tagger and lemmatizer for Slovenian (Grčar et al. 2012).

```
<Lemma>
    <feat att="zapis oblike" val="pazljivo"/>
</Lemma>
```

Figure 2: The lemma of the adverb *pazljivo* **'carefully' in the XML LMF format.**

2.3.3 Part-of-speech and lexical features

In addition to the obligatory grammatical information on the part-of-speech category, most lexicon entries include one or more additional lexical features, i.e. grammatical features that are assigned at the lemma-level and belong to all word forms in its inflectional paradigms, such as type (common, proper) and gender (masculine, feminine, neutral) with nouns, type (main, auxiliary) and aspect (perfective, progressive, biaspectual) with verbs, case with prepositions, and so on. Like all other grammatical features in the lexicon, lexical features are given in the form of pairs of attributes (e.g. *gender* with nouns) and their values (e.g. *masculine, feminine* or *neutral*).

```
<feat att="besedna_vrsta" val="prislov"/>
<feat att="vrsta" val="splošni"/>
```

Figure 3: Lexical properties (type = general) of the adverb *pazljivo* **'carefully' in the XML LMF format.**

2.3.4 Inflectional paradigm

General information on the lexicon entry is followed by the inflectional paradigm, consisting of one or more inflected forms with corresponding information on specific grammatical features, usage frequency and compliance with the language standard (in case of variant inflected forms).

2.3.4.1 Inflected forms

In the case of uninflected par-of-speech categories, the inflectional paradigm[10] of a lexicon entry usually includes only one form, whereas the number of inflected word forms for other categories depends on the category itself, its lexical features and the degree of variability in language usage. Among the inflected part-of-speech categories, the shortest paradigms appear with adverbs and some pronouns, while adjectives display the largest paradigms, as they inflect for gender, degree of comparison, number, case and definiteness, with an average of 59 different word forms per lemma (see Table 1).

2.3.4.2 Inflectional features

Each inflected form is assigned a set of inflectional grammatical features. In contrast to lexical features, inflectional features distinguish individual forms in the inflectional paradigm of a lemma, and are therefore assigned at the level of (abstract) grammatical word forms, such as gender, number and animacy with nouns; degree of comparison with adverbs; form, person, number, gender or negation with verbs, etc. The set of inflectional features in Sloleks is based on JOS morphosyntactic specifications. However, it is not obligatory for all possible inflectional features within a part-of-speech category to be assigned to all lemmas belonging to the category, as their actual selection depends on the lemma and its lexical features.

At the same level, the lexicon also includes a mapping of all grammatical features to a position-based compact string encoding, the so-called morphosyntactic description (MSD) used in automatic morphosyntactic tagging of text corpora (see the chapter by Erjavec et al. in this volume).[11]

```
<WordForm>
    <feat att="stopnja" val="primernik"/>
    <feat att="msd" val="Rsr"/>
    /.../
</WordForm>
```

Figure 4: Inflectional features and the MSD of a comparative form the adverb *pazljivo* **'carefully' in the XML LMF format.**

10 The expression "inflectional paradigm" is used to denote all the inflected forms of the lemma, as determined by the JOS system, regardless of whether they are the result of morphological (e.g. declension) or formational (e.g. gradation) processes.

11 All the comparative forms of adverbs are therefore given the "Rsr" MSD, since – in accordance with the morphosyntactic specifications of JOS – the first letter of the MSD contains the part-of-speech (R: adverb); when dealing with adverbs, the second letter then indicates the type (s: general), and the third one the degree (r: comparative).

2.3.4.3 Variants

When a given set of grammatical features (an abstract grammatical form) is realized with more than one spelling, we consider these competing word forms to be inflectional variants. They are further distinguished by the so-called variant features, which currently include information on compliance with the current language norm (as set out in *Slovenian Orthography*, 2001). Inflected forms without any normative information are considered to be in compliance with the norm (e.g. the inflected form *gradu* of the lemma *grad* 'castle' in dative singular), while the "nestandardno" attribute value denotes incompliance with the norm (e.g. the inflected form *gradi* in nominative plural). If there is a variation between two or more standard forms, they are each assigned the "variantno" label (e.g. the forms *grada* and *gradu* in genitive singular).

2.3.4.4 Corpus frequency

In Sloleks each inflected word form is also assigned its frequency in the reference 1.2 billion-word Gigafida corpus, which has been extracted automatically by querying the frequency of occurrence of the combination of the given inflected form, its lemma and its MSD. The overall accuracy of the reference morphosyntactic tagger and lemmatizer used in the annotation of Gigafida (Grčar et al. 2012) is currently 91.34 %, but varies significantly depending individual types of lemmas or word forms (ibid: 92–94).

```
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljiveje"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="97"/>
</FormRepresentation>
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljivejše"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="2"/>
</FormRepresentation>
```

Figure 5: Variant comparative inflected forms of the adverb *pazljivo* **'carefully' with normative and corpus frequency information in the XML LMF format.**

2.3.5 Related forms

In addition to information on the inflectional properties of a lemma, Sloleks also includes information on its derivational connection with other lemmas or lexicon entries. The current list of derivational relations in Sloleks includes the following reciprocal relations: between a noun and its derived possessive adjective (*kruh* 'bread' and *kruhov* 'of bread'), between a verb and its gerund (*briti* 'to shave' and *britje* 'shaving'), between an adjective and a derived noun ending in *-ost* (*zarjavel* 'rusty' and *zarjavelost* 'rustiness'), between a verb and its adverbial participle (*začeti* 'to start' and *začenši* 'starting'), between a verb and its adjectival participle (*ujeti* 'to catch' and *ujet* 'caught'), between an adjective and the derived adverb (*navihan* 'mischievous' and *navihano* 'mischievously'), between an adjective and its elative (*lep* 'beautiful' and *prelep* 'too_beautiful, ~magnificent'), between an adverb and its elative (*glasno* 'loudly' and *preglasno* 'too_loudly') and between a lemma and its abbreviation (*gospod* 'mister' and *g.* 'Mr.').

```
<RelatedForm>
    <feat att="idref" val="LE_64ba3adcc4c42841599358c8
    6b738f1c"/>
    <feat att="besedna_vrsta" val="pridevnik"/>
    <feat att="lema" val="pazljiv"/>
</RelatedForm>
```

Figure 6: Related form (adjective) of the adverb *pazljivo* 'carefully' in the XML LMF format.

To summarize the above description of the Sloleks lexicon entry structure, Figure 7 shows the full set of information included in the lexical entry of the adverb *pazljivo* 'carefully', schematized to better visualise the hierarchical organisation of the original data in the XML LMF format.
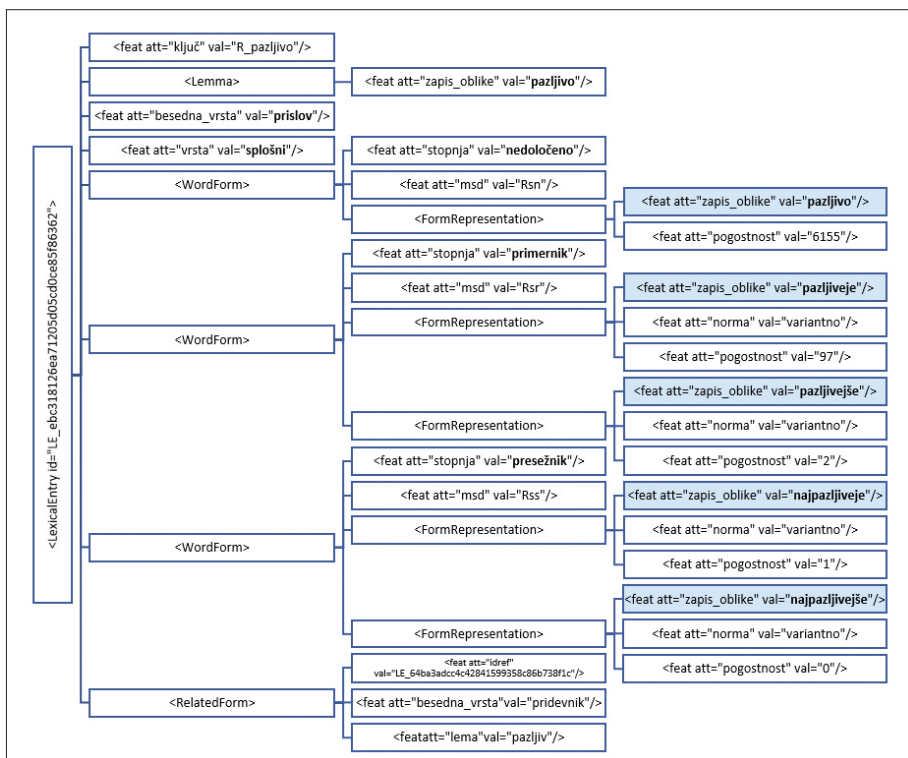
Figure 7: A schematic illustration of the full lexicon entry for the adverb *pazljivo*
'carefully' in the XML LMF format with inflected forms shaded in blue.

2.4   Visualisation

In addition to being used in various natural language processing applications, a structured collection of morphological information on Slovenian lexica that enables flexible modifications of the information that is displayed, and how it is visualised, represents an equally valuable language resource to be used as an autonomous inflection manual or integrated into other language resources, such as an online dictionary (see Dobrovoljc in this volume). An example of Sloleks lexicon visualisation has also been proposed as part of the Communication in Slovene project portal.[12]

As can be seen in the example of the visualisation of the lexicon entry for the adverb *pazljivo* 'carefully' (Figure 7) in Figure 8, the red-coloured lemma is followed by information on the part-of-speech category, lexical features and the overall

12   http://www.slovenscina.eu/sloleks

corpus frequency (a sum of frequencies for individual inflected forms included in the original database). This is followed by a separate display of the inflectional paradigm with corresponding grammatical and normative features, where specific combinations of inflectional features (grammatical forms) are separated by a line. Numbers in the frequency column include a hyperlink to the usage examples in the online corpus concordancer (corpus queries are generated automatically for the given word form, lemma and MSD combination). The bottom of the entry includes information on potential related lemmas (and their part-of-speech category), also in the form of a hyperlink to the corresponding lexicon entry.

Figure 8: Visualisation of the lexicon unit for the adverb *pazljivo* 'carefully' in the Sloleks web service.

3   GUIDELINES FOR FUTURE DEVELOPMENT

3.1   Expanding the entry list

As already stated in section 2.1., the Sloleks Morphological Lexicon currently contains around 100,000 of the most commonly used lemmas in Slovene vocabulary. Compared to the glossaries of other accessible morphological resources for Slovene, which are either smaller in size (Apertium, MULTEXT-East) or are not corpus-based (SP 2001, DSLL), it currently covers the largest percentage of general Slovene vocabulary. However, the planning of dictionary and other linguistic descriptions of modern Slovene on the one side, and the growing and diverse needs for its machine processing on the other, also necessitate its further expansion. This process is envisaged as three concentric circles, each representing the fundamental starting point of the next, although not necessarily drawing on the same methodological considerations.

Given that priority is given to the integration of morphological data into a digitally-born descriptive dictionary of modern Slovene, the first concentric circle of the further expansion of the Sloleks Morphological Lexicon represents its harmonization with the dictionary database entry list, i.e. the inclusion of the (missing) core lexical units of the Slovene language, including multiword dictionary headwords, spelling variants, and other lemmas or forms morphologically linked to the lemma of a given dictionary headword.

The second circle of expansion includes the vocabulary taken from the reference corpus of the Slovene language. Although some of the reference corpus vocabulary will not necessarily become part of the dictionary, depending on the dictionary headword selection criteria, it nevertheless forms an indispensable part of various language technologies – including those used in dictionary compilation – since the lemmatizers, morphosyntactic taggers, and lexical data extraction tools must be capable of correctly recognizing both headwords and their surrounding vocabulary. In accordance with the virtuous circle of linguistic annotation, the expansion of the lexicon improves the language model of the tools, which in turn improves the accuracy of corpus annotation.

By comparing the overlap of word forms (token types)[13] in the Sloleks Morphological Lexicon with the vocabulary in the Gigafida reference corpus, we find that Sloleks contains only 43% of all token types with a minimum frequency of five occurrences in the Gigafida corpus. As expected, this share increases by increasing the frequency threshold; however, the Morphological Lexicon still covers only 79% of the total 251,292 token types that appear at least 100 times in the corpus. Such frequency of an individual token type (i.e. word form, not lemma) in a balanced and representative corpus is already a strong indicator it should be formally described in an adequate morphological database.

A more detailed analysis of the list of the most common word form types in the Gigafida corpus not yet present in the Sloleks Morphological Lexicon indicates that the database would benefit from being expanded with the following vocabulary groups:

- various types of abbreviations (*p., s., j..; nan., dok., mr.; m2, cm3, a3; UV, MMS, VIP, SUV; VPS, SŽ,* etc.);

- borrowed nouns (*city, miss, fax, art, dj, bluetooth, mac, facebook, prix, alias, maestro, college, gay, styling, fitness, volley, weekend, hiphop*, etc.);

- non-inflected attributes (*turbo, online, anti, stereo, retro, audio, etno, latino, afro*, etc.);

---

13 In doing so, we intentionally compared only word forms written in lowercase letters, since we did not want to depend on the automatically added data about the lemma or the spelling particularities found in corpus texts (e.g. *slovenija, ljubljana*, etc.).

- non-standard word form spellings (*tud, kr, blo, brezveze, dobr, nevem, kao, jst, jap, tolk, nč, lahk, drgač, al, tm, zarad, mislm, pomoje, una, brezveze*, etc.);

- interjections (*živjo, bognedaj, jao, jp, hehe, he, hahaha, hahahaha, sviš, hehehe, khm*, etc.);

- foreign and Slovene proper nouns (*obama, ilirika, evroliga, barca, clio, patria, beverly, pomurec, messi, airways, michel, svena, sarkozy, coca, evrovizija, titanik, čedad, Wikipedia*, etc.);

- dialect or field specific vocabulary (*škrinja, zalüblenih, mojoga, škürec, zadvečerek, špas*, etc.);

- some commonly used vocabulary or loanwords (*drugouvrščen, mimoidoči, prida, kapitalov, superpokal, štoparski, fotogalerija, tričetrt, bogve, drugoligaški, didžej, avtohiša, enoprostorec, osemvaljnik, supermodel, drska, preska, četrtinski, požarnik, klaviaturist, klientelizem, kapetanski, avtoprevozništvo, označba, predizbor, napak, prismučati, nezemljan, brezplačnik, evroobmočje, streljaj, dvetretjinski*, etc.).

- For the purpose of natural language processing, frequently used foreign vocabulary should also be recorded, such as lexical items constituting foreign proper nouns (e.g. *the, of, and*, etc.).

After expanding the lexicon with the missing headwords from the dictionary and the frequent vocabulary found in the reference corpus, the third circle of expansion foresees the inclusion of specialized vocabulary for the requirements of specific language manuals or technological applications, such as typically spoken vocabulary, vocabulary from individual areas of expertise, dialect vocabulary, or other types of vocabulary from different registers. As opposed to the first two circles, which represent the universal core of a language's lexicon description, the third circle of expansion of the lexicon cannot be foreseen or guaranteed in advance; however, it is of key importance that the community be allowed to carry out the expansion independently, by providing it with the tools and sources necessary for such task – starting with an open source database of inflectional patterns for Slovene, as discussed in the following section.

3.2 Revising morphological patterns

One of the most important tasks linked to both the expansion and re-evaluation of existing lexicons for Slovene is the creation of a finite set of machine-readable

inflectional patterns for the language, which would enable the validation of inflectional paradigms of headwords in existing reference books, the assignment of paradigms to new lemmas, and the development of methods for their automatic recognition in text corpora (e.g. Šnajder 2013 for Croatian). When we look at the range of morphological lexicons for Slovene, one could assume that there already exist several similar collections of inflectional patterns. However, these are not available to the research community at large, and the principles behind their design, classification and compliance with actual language use are mostly not documented. What is more, the initial attempts to implicitly register the complete list of patterns based on the comparison of patterns available in larger accessible reference works, such as SP2001, Apertium, and Sloleks (Dobrovoljc 2014), also revealed non-systematic pattern selection and classification, as many errors, inconsistencies or incompatibilities with contemporary language usage were identified in all three language resources.

This confirms that any upgrade or further application of the existing morphological databases in Slovenian should also involve the creation of an updated, freely accessible list of formalized inflectional patterns for the language. However, in contrast to the traditional linguistic approaches to description of morphological patterns in Slovene, their use in language technologies requires the consideration of a few additional design principles. In addition to the strict separation of inflectional patterns on the one hand, and pronunciation patterns on the other (as opposed to simultaneous description of both orthographical and pronunciation changes during inflection in DSLL, see sections XXXVIII–XLIX), as well as machine-readable formalization of patterns in the form of algorithmic rules for paradigm generation – both aspects are discussed in detail by Dobrovoljc et al. (2015), and have already been implemented in the initial Sloleks design – future revisions of the existing inflectional patterns in the lexicon should mainly focus on their compliance with actual language use.

Updating morphological information based on tendencies observed in balanced and representative corpora of modern Slovene would not only ensure an exhaustive coverage of the frequently used vocabulary (regardless of its compliance with the existing codification norm), but also enable an important re-evaluation of morphological descriptions in existing reference grammar books and dictionaries, which were not based on such vast collections of authentic language use. As demonstrated by Dobrovoljc et al. (2015), who compared the DSLL2's schemes for dynamic stress and morphology with data occurring in the Gigafida reference corpus, contemporary language use reveals the inexistence of some supposedly systemic inflectional forms (e.g. the accusative dual *cerkvé* of the noun *cérkev* 'church'), as well as the unjustifiability of some theoretic presuppositions, such as the claim that the *e* comes between two sonorant consonants in the dual and plural genitive case only when the

second sonorant is *r* (*kamra: kamer*), since usage shows that *e* may be inserted even between other combinations of sonorants (e.g. *himna: himen*; *kolumna: kolumen*; *avla: avel*).

Such re-evaluations based on analysis of authentic language use are even more important from the point of view of complete paradigm attribution, i.e. the coupling of concrete lemmas with concrete inflectional patterns, where initial analysis of attributed patterns of comparison for adverbs in the Sloleks Morphological Lexicon and the SP 2001 Slovenian Normative Guide Dictionary (Dobrovoljc 2014) revealed that both reference works diverge from common language use. For example, some adverbs that demonstrate comparison by inflection in the Gigafida corpora (e.g. *smiselno, preudarno, poredko, enakovredno, korektno, športno*) are referenced without any inflectional paradigm in one or both manuals, whereas sometimes the paradigm for comparison is attributed to adverbs that do not exhibit such behaviour in common use (e.g. *arogantno, bistroumno, strahovito, zagonetno*, etc.). Even more surprisingly, such discrepancies occur in morphological patterns for exceptions, where the Slovenian Normative Guide, for example, gives the comparative forms *dražje, ožje* and *težje* for the adverbs *drago, ozko and težko* (even though comparative forms *draže, ože* and *teže* also appear in the corpus); the forms *krajše* and *kračje* are given for the adverb *kratko* (even though the second form is not present in language use); the adverb *gladko* has the forms *gladkeje, gladkejše, glaje* and *glajše* (even though *glaje* does not occur in the reference corpus and g*lajše* has only one entry), and so on.

3.3 Categorizing variation

Morphological variation, i.e. the existence of several formal possibilities of expressing the same grammatical form, is quite common in Slovene, and occurs at various linguistic levels: the spelling (*v naprej* or *vnaprej* 'ahead'), pronunciation (/*drsáuka*/ or /*drsálka*/ 'skater') or accentuation (*upokójenec* or *upokojênec* 'pensioner') of lemmas, as well as when selecting the morphological paradigm (*Luka: Luka or Luke* or *Lukata* for inflection of the male name Luka), the spelling or pronunciation of inflected forms (*college*: *collegea* or *college* for inflection of the loanword college), or word formation (*vanilija: vanilijev, vanilijin* or *vanilin* for forming an adjective from the noun vanilla).

Given that morphological variants in the existing version of the Sloleks lexicon are listed as word forms with identical lexical or grammatical properties, we are unable to systematically distinguish between them without additional specification of the expected differences. Consequently, since morphological lexicons are used for various purposes, it would be useful to assign the differentiation (variant)

features to the individual variants, along with their systematic classification. In doing so, we must stress that this kind of classification must not be confused with normative qualification (illustrated in Section 2.4.4.3): the first denotes a user's choice within the language system, while the second entails subsequent linguistic interpretation, which largely depends on social conventions and is thus also subject to change. Both sets of information are essential to a morphological lexicon; however, since classification enables a directed recall of individual variant forms or complete variant paradigms of one or more lexical units, while the information on their normative (non-)stigmatization is a key component for integrating the lexicon into language reference books, and can also be of value to language technology applications for text generation, such as machine translation software or speech synthesizers, that can benefit from information on the (non-)standard nature of individual variant choices.

The first attempts at systematic classification and normative qualification of variant morphological forms have already been made when establishing the design and workflow of the "Slogovni priročnik"[14] (Online Style Guide) web portal. The portal is intended as an online service for solving the most common language-related issues in Slovene text production, by juxtaposing information about the valid orthographic standard on the one hand and corpus data on the other (Krek 2012c; Krek et al. 2013a; Dobrovoljc and Krek 2013). The back-end mechanism, which connects the user's question to the relevant issue and its explanation (by visualizing the corpus and normative information for the exact queried word form(s)), takes all the necessary data from the Sloleks Morphological Lexicon, where lemmas or inflected forms, related to the language issue in question, have been adequately categorized. Each form (base or inflected) is therefore ascribed three types of categorization data: (i) the category of the issue, i.e. the type of morphological variation, which is based on an ontologically organized list of language-related issues in Slovene (Dobrovoljc and Krek 2011; Bizjak Končar et al. 2011), (ii) the type of variant within the category, and (iii) its normative value.

An example of such a categorization is shown by a fragment of the lexicon unit for the noun *Klemen* in Figure 9. At the first level, the lexical unit is already carrying the information about its link to language issue no. C1a3a (*Morphology > Nouns > Masculine Declensions > Nouns with Unstable Vowels > Slovene Proper Nouns*), while individual forms in the subsequent paradigm also include information about the specific variant they belong to (C1a3a_s_1, for example, is used to denote a paradigm that omits *e,* while C1a3a_s_2 a is used to denote a paradigm without this omission), as well as the information on their normative qualification (e.g. the *variantno* qualifier that marks a standard double).

---

14 http://slogovni.slovenscina.eu/
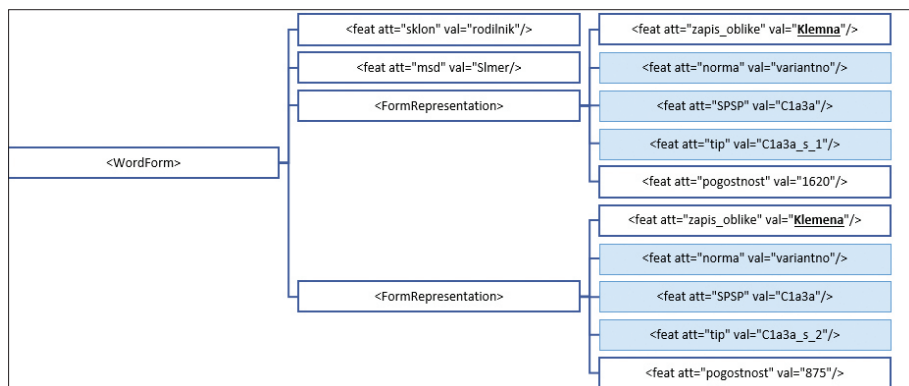
Figure 9: Part of the _S_Klemen_ lexicon unit with a category of variant declension.

This kind of categorization of morphological characteristics in the database thus enables a controlled recall of data within an individual lexicon unit, such as the list of all relevant linguistic issues related to the unit, or one or more forms of a given variant paradigm. On the other hand, it also enables an automatic recall of the list of all other lemmas which display the same kind of morphological, derivational or pronunciation variance, e.g. all Slovene proper nouns with an unstable vowel.

3.4   Adding pronunciation

The Sloleks Morphological Lexicon currently does not include data on the pronunciation characteristics of the word forms it contains, meaning that the word forms included are not accented. One of the priority upgrades to the existing version of the Sloleks Morphological Lexicon is thus the incorporation of pronunciation data, with the aim of providing a comprehensive description of both inflectional and phonetic characteristics of contemporary Slovene vocabulary. This is especially important from the point of view of speech technology, since Slovenian linguistic infrastructure currently lacks a freely accessible lexicon needed for the development of speech recognizers and synthesizers for various applications, such as subtitle generators, screen readers for visually impaired, natural language interaction systems, and the like.

The pronunciation information, based on a standard machine-readable phonetic alphabet, should be included at the level of both lemmas and inflected forms. In cases of pronunciation variation, a common phenomenon in Slovene, one orthographical word form can thus have several pronunciations assigned; similarly to

dealing with the variation of non-accented, orthographical forms, we can distinguish between them by using adequate qualifiers, which allow us to automatically recall the pronunciation of an individual form or all forms in one of the variant pronunciation paradigms (see Section 3.3 Categorizing variation). This approach is used for all types of pronunciation variance, regardless of whether we are dealing with phonemic (prevajalka: *prevajalka-prevajau̯ka*) or accentual variance (agencija: *agéncija-agencíja*) of all or just one of the inflectional forms in a given paradigm.

Just like in the current version of the Sloleks lexicon, adding pronunciation information would not change the fact that lemmas with the same spelling and pronunciation are separated into several independent lexical units if they display different expressive characteristics, i.e. if they fall under different parts-of-speech (e.g. the adverb and adjective *spet*), have different lexical properties (e.g. the feminine and masculine noun *prst*), or different inflections (e.g. the verbs *vesti*: *vedem* and *vesti: vezem*). Similarly, no changes would apply to homonymic pairs of lexemes with identical formal, but different semantic properties (e.g. the masculine nouns *bor* 'pine tree' or *bor* 'chemical element'), which would continue to be processed as one distinct unit of vocabulary with only one corresponding lexicon unit (the masculine noun *bor*), regardless of their meaning.[15] Since the Morphological Lexicon does not record tonemic accent, the same rule applies to pairs of semantically differing homographs that are differentiated only by their tonemic accent (e.g. the adjectives *bûčen* 'of a pumpkin' and *búčen* 'loud' thus share a common inflectional paradigm of the general adjective *bučen*).[16]

On the other hand, adding pronunciation information would change the treatment of lemmas with the same spelling, but a different pronunciation, e.g. *partija* (pronounced *partíja* 'the (Communist) party' and *pártija* 'the match') or *častiti* (*částiti* 'to buy somebody a drink' and *častíti* 'to worship'), which to now were considered a single lexicon unit due to their identical lemma, grammatical patterns and non-accented inflectional paradigms. By introducing semantically differentiating pronunciation information, both lexemes become independent lexicon units (*S_partija_1* in *S_partija_2*). However, it should be noted that current morphological analysers for Slovene do not enable semantic disambiguation of morphologically overlapping homographs within a given context, which is why word forms belonging to such homographs would be given an identical lemma and morphosyntactic tag. In turn, the corpus frequency information (see Section 2.4.4.4.) for identical word forms with identical grammatical features would be identical for both lemmas.

---

15 For the relationship between the lexicon and dictionary headword, see the paper by Dobrovoljc in this publication.

16 For the relationship between formally motivated lexicon units and semantically motivated dictionary units, see Gantar (2015) and K. Dobrovoljc in this publication.

4 CONCLUSION

The Sloleks Morphological Lexicon, together with its morphological, derivational, normative, distributional and other types of linguistic data, represents a common intersection point between the various language resources foreseen by the proposal for a new dictionary of modern Slovene (Krek et al. 2013b), such as reference, balanced, spoken, historical, and other types of linguistically annotated corpora. On the other hand, the data from the lexicon are equally useful in reference language manuals, such as (digital) dictionaries, online style guides, grammars and others. By introducing a systematic approach to the description and formalization of Slovene morphology, the Sloleks lexicon enables a uniform and consistent treatment of morphological phenomena within the fields of both language technologies and language resources. As such, it aims to overcome one of the key deficiencies of Slovene natural language processing and Slovene language teaching – from primary and secondary schools to teaching Slovene as a foreign language.

Future development of the lexicon should mainly focus on a significant expansion of its entry list, including multi-word units, usage-based revision of the existing morphological patterns and their attribution to individual lemmas, systematic linguistic and normative categorization of frequent morphological variation, and addition of pronunciation information. All these processes must be implemented with the current state-of-the-art technologies, many of which are already available for Slovene. The future development of the Sloleks lexicon should thus be understood as an ongoing process without a final endpoint, since languages are always accruing new words, which need to be both adequately described and efficiently processed. With this in mind, the widespread usability of the lexicon can only be assured by a continued open access to this resource. This not only justifies the investment into its development, but also gives the Slovene language the opportunity to survive in the coming digitized world.

Dictionaries and Learning Slovene

Tadeja Rozman, Iztok Kosem, Nataša Pirih Svetina and Ina Ferbežar

Abstract

This paper discusses which Slovenian dictionary or dictionaries would be the most suitable for native and non-native Slovenian speakers to use. Slovenian studies are presented that focus on dictionary use and the comprehensibility of dictionary information among Slovenian primary and secondary school students, as well as non-native Slovenian speakers. A brief overview is also presented of relevant findings from dictionary use studies conducted abroad. After this overview of the needs, abilities and preferences of dictionary users who are learning a language, the paper concludes with some suggestions for Slovenian dictionary makers.

Keywords: dictionary use, school dictionary, learner's dictionary, vocabulary acquisition, language learning

1   INTRODUCTION

Dictionaries are essential language resources, indispensable to both foreign language learning as well as native language acquisition. Dictionary use has a positive impact on the learning and retention of new words, and facilitates the improvement of knowledge with regard to the semantic characteristics and usage of words (Paynter et al. 2005: 35–37, 41–45). Moreover, the rich vocabulary that dictionary users can acquire is an extremely important part of individual's communication skills. In the context of education it is important to stress that dictionaries play an important role in student performance, as they can be used as an aid in understanding new material, and consequently contribute to better reading literacy (Paynter et al. 2005: 3–7; Pečjak 2012: 31). However, experts warn that the use of a dictionary which does not consider the development and language proficiency levels of children and other learners, as well as their specific needs, can have negative impacts (e.g. Wright 1998: 7).

In Slovenia there are no dictionaries targeted at non-native or young native speakers of Slovene. Similarly, there is also limited research literature that focuses on these topics. As a result, teachers of Slovene (as L1 or as L2) often use a general monolingual dictionary, i.e. the *Dictionary of Slovene Literary Language* (DSLL),[1] despite the dictionary targeting adult native speakers of the language. According to research (see Section 2), only a small proportion of teachers in primary and secondary schools are aware of the problematic nature of using such a dictionary in teaching, with most thinking that DSLL is totally appropriate for use with students of all levels. This is partly the consequence of a lack of dictionary choice, and predominant and symbolic role of this general monolingual dictionary, which is regarded as a fundamental language resource (as it claims to contain all the important information on Slovene words; see Stabej 2009; Rozman 2009). Another contributing factor is a lack of research focused on dictionary use in relation to language teaching, language acquisition and vocabulary retention.

This situation has slightly improved in the last decade, as a few studies into dictionary use in education have been conducted. The findings of such works are very important for the planning of dictionaries for young native and non-native speakers, taking into account the fact that dictionaries should always consider the needs and abilities of target users, as well as their dictionary habits (i.e. how they consult dictionaries). When planning dictionaries for young native speakers or older learners of Slovene, it is thus important to know which information they are more likely to need or consult (and will be more relevant to them), and how such information should be presented. To some extent, we

---

1   This is discussed in more detail in Section 2.

can draw on existing research on and experience in the compilation of school or learner dictionaries of other languages (see Section 3). However, this knowledge is not directly transferable to the Slovenian situation or language, mainly due to differences in how this society perceives dictionaries, standardisation and language teaching, as well as the particularities of its didactic methodology and education system.

The first part of this paper presents an overview of dictionary use research in Slovenia (Section 2), and then presents a review of the relevant international research studies (Section 3). In Section 4, we discuss concrete solutions based on existing research and our knowledge of the field, and also considering the characteristics of language development.

2 DICTIONARY USE RESEARCH IN SLOVENIA

Several studies have been conducted into the understandability of dictionary information among pupils and students in Slovenia, as well as dictionary use in Slovene language teaching, and language problems. This section provides an overview of the key results relevant for the planning of the dictionaries for non-native and young native speakers of Slovene.

2.1

The first large-scale survey was conducted in 2008 (Stabej et al. 2008). The survey included 409 teachers of Slovene and 3,427 students at different levels of education, from 4th grade of primary school up to 4th year of secondary school. The survey was two-fold: the first part was focused on the use of and opinions about monolingual dictionaries, and the second part aimed at detecting problems in language acquisition.

The responses revealed that the teachers used dictionaries quite frequently when preparing different types of teaching materials; mainly for teaching vocabulary, but also for teaching grammar, literary and technical texts, and when preparing and correcting homework and tests. DSLL was the most frequently used dictionary, with 96.8% of teachers reporting occasional use of DSLL in class. DSLL was also consulted when learning different syllabus contents (Table 1), even when using a dictionary was not envisaged by the syllabus or textbook. In addition, the teachers often prepared exercises on learning how to use DSLL.

Table 1: Percentages of teachers using DSLL in different teaching activities

Activity or topic	%
literary text	68.5
lexis and phraseology	65.3
technical text	56.2
orthography	52.8
group correction of tests and homework	39.4
proper pronunciation	37.9
grammar	32.3
text linguistics and communication	24.2
Other	2.4

The majority of the teachers reported encouraging students to make independent use of dictionaries in various activities (Table 2), especially those related to encoding, and occasionally directing students to the dictionary when encountering an unknown word.

Table 2: Percentages of teachers that encouraged the use of dictionaries[2] during different activities

Activity	%
writing	71.6
searching for synonyms and antonyms	62.3
preparing an oral report	61.1
text correction	56.0
searching for Slovene equivalents of foreign words	55.5
searching for unmarked equivalents	54.0
language exercises	37.2
reading	27.6
other	4.9

The teachers agreed that it is useful for students to learn how to use monolingual dictionaries, because this skill improves their communication skills, helps them with using language correctly, facilitates language acquisition and helps expanding their vocabularies. Overall, the teachers had good opinion of DSLL, giving the following reasons:

- they stated that it is useful for solving various language problems (especially those related to word meanings, spelling and pronunciation, slightly less with those related to stylistics, terminology, pragmatics and grammar),

---

2   The question asked about the use of DSLL, the dictionary part of Slovene Orthography, and dictionaries of foreign words.

- they agreed that it is normative,

- they believed that it is easy to understand and use,

- the majority (73.3%) considered it suitable for students.

Similarly, the students surveyed also thought highly of dictionaries:

- they stated that they help by providing the correct language use and thus solving language problems,

- they see dictionaries as normative reference works,

- they did not consider dictionary definitions to be too demanding, but agreed that the often difficult-to-understand abbreviations and symbols make dictionaries more difficult to use than necessary.

However, despite their positive attitudes towards dictionaries, a majority of the students did not like to use them (only 37% of primary school students and 30.3% of secondary school students agreed with the statement *Rad/a uporabljam slovarje*; 'I like using dictionaries') or simply did not use them at all (for example, DSLL is used by only 24.5% of primary school students and 16.5% of secondary school students). On the subject of independent dictionary use, the students reported using them mainly when doing dictionary-related exercises. When it came to solving language problems, the students mainly reported consulting dictionaries about the meaning and spelling of words. In both cases, the percentage of students using dictionaries was rather low. Similarly, using a dictionary proved to be one of the least favoured strategies when solving problems related to lexis, as the students preferred to ask a teacher or a friend, not complete the exercise, or search for the answer on the Internet.

Also interesting in this earlier study are the teachers' answers with regard to the types of language errors they most often find in their students' writing or speaking. By far the most frequent are spelling or pronunciation errors, followed by grammar and style errors, while less commonly observed are errors related to semantics, collocations, syntax and phraseology.

2.2

A similar survey, but much smaller in size, was conducted in 2013 (Čebulj 2013). The subjects were 75 primary school teachers (up to 5th grade). Most of the teachers (even those in 1st grade) reported using DSLL in class and teaching their pupils how to use it, and also using dictionaries as one of the strategies for explaining the

meaning of words.[3] The teachers did report that they often observe pupils having difficulties in using DSLL (especially problems with the order of the alphabet when looking up words), and a majority of them agreed that there is a need for a school dictionary.

2.3

As part of the Communication in Slovene project (SSJ),[4] a major survey on Slovene language teaching was conducted in 2010 (Rozman et al. 2010; Rozman et al. 2012). The respondents were 276 teachers of Slovene as L1 and 1,465 students (attending the last three grades of primary school or attending secondary school). Despite not including many questions related to dictionaries and language acquisition, this work does provide some highly relevant findings.

The teachers stated that they saw the acquisition of vocabulary during education as very important, so ideally they would dedicate more time to activities promoting this. In contrast, they would dedicate less time to reference works and how they are used, although they still considered these activities to be fairly important. Similarly, the students believed that a large vocabulary is the most important part of obtaining good communication skills,[5] and considered knowing about dictionaries and how to use them as less important, even less important than knowing how to use the internet. Consistent with this view were students' answers on the use of different language resources and information and communications technology (ICT): they reported using electronic resources, especially web browsers, much more frequently than dictionaries (especially paper dictionaries) when it came to solving language problems. These findings are also consistent with the results of Stabej et al. (2008), presented in section 2.1. On the other hand, the teachers, and especially the older ones, rarely used online dictionaries and ICT in class, although in principle they supported the use of these resources.

2.4

Also conducted during the SSJ project was a survey on the understandability of grammatical (morphosyntactic) information in DSLL (Rozman et al. 2010). The

---

3   The teachers using the dictionary as a source of information on the meaning of words, or, less frequently, pupils using the dictionary independently.

4   http://www.slovenscina.eu/

5   The question was: Which of the skills presented below is in your opinion important for speaking, writing and reading Slovene literary language? Eight answers were provided, and the respondents had to evaluate each of them on a scale of 1 to 6.

survey included 389 students attending 8th and 9th grades of primary school and 2nd and 3rd years of secondary school. The findings showed that newly compiled entries in which the grammar information was as explicit as possible were more understandable than DSLL entries,[6] which provide the same information in the form of abbreviations, or when the information in the entries is condensed immediately after the headword. The most useful factors with regard to improving understanding were those entry components that contained more explicit grammar information and were most relevant for the questions in the test used in the survey; the position of the information in the entry was not relevant. The examples of such information included non-abbreviated labels, specially highlighted explanations and dictionary examples.

2.5

Conducted between 2007 and 2009 as part of a PhD thesis, Rozman (2010) is a detailed analysis of syllabi and textbooks for Slovene for the last six grades of primary school and all years of secondary school. The analysis focused on the level of dictionary-related content in Slovene language teaching. It also included a survey on the understandability of dictionary definitions, conducted with about 607 students from three age groups: 5th and 6th grades of primary school, 8th and 9th grades of primary school, and 2nd and 3rd years of secondary school.

The ability to use a dictionary is one of the objectives of the Slovene syllabus that should be achieved at the end of primary school. Dictionaries are part of the syllabus from 7th grade onwards, although exercises involving dictionary use (especially the use of DSLL) are also found at earlier levels. The analysis in this study pointed out several problems in introducing dictionaries into the teaching process, most of which stem from the fact that due to its outdatedness, size, internal structure and less explicit nature of the information it contains, DSLL is often too demanding to allow the kind of consultation envisaged in such exercises.

The survey focused on comparing the understandability of DSLL definitions and of those written especially for the survey. These newly written definitions targeted maximum understandability and took into account the principles of explicitness and straight-forwardness, and avoided using abstract or specialised vocabulary, complex syntax and highly polysemous words. The testing confirmed the hypothesis that DSLL definitions are less understandable, especially to younger students, due to their abstract nature and overly demanding definition vocabulary. The survey also pointed to several features of definitions that affect their

6   Especially to students in primary schools.

understandability, most of them being linked to their abstract nature, structure, length and type, or to the structure of dictionary entries.

2.6

In 2010, a freely available corpus of student texts called Šolar[7] was built, containing authentic texts written by primary and secondary school students, which makes it a good source of information on their writing skills. An exhaustive analysis of the corpus (Kosem et al. 2012a) has been conducted for the purposes of the Pedagogic grammar portal,[8] although this analysis is only partly relevant for dictionary planning, as language errors[9] were categorised according to language problems (e.g. spelling, syntax) rather than at the level of individual words. The latter approach was used by Arhar Holdt and Rozman (2015), who focused on extracting information that could be used in the preparation of a school dictionary or vocabulary-related teaching materials. Their research was conducted on only part of the Šolar corpus, but still confirmed that such information would be useful for dictionary treatment of both content and function words. Among the identified features that would be particularly useful for dictionary users are linking dictionary and grammar information, putting a heavier stress on the collocational, stylistic and syntagmatic characteristics of words, offering the option to compare words with similar forms but different meanings, and pointing out similarities and differences between words with similar meaning but different collocational, stylistic or other characteristics.

2.7

All the studies mentioned above focused on students and teachers in Slovenian primary and secondary schools, and there is almost no research literature on Slovene as a second or foreign language. One exception is Rozman (2003), consisting of an analysis of English advanced learners' dictionaries and a short survey among 64 participants and 18 teachers of Slovene L2 language courses, conducted in the summer of 2003 by the Centre for Slovene as a second and foreign language (CSDTJ).[10] The results confirmed the need for a monolingual dictionary for non-native speakers of Slovene, and based on this, concrete suggestions on certain aspects of dictionary content were prepared. A similar survey, but on a

---

7 http://www.slovenscina.eu/korpusi/solar
8 http://www.slovenscina.eu/portali/pedagoski-slovnicni-portal
9 Only instances of language use corrected by teachers counted as errors.
10 http://centerslo.si/

smaller scale, was conducted by CSDTJ in May 2015; the survey contained ten questions and was completed by 15 teachers of Slovene as L2.

The main findings of both surveys can be summarised as follows:

- The vast majority of language learners use dictionaries,
- they mainly use bilingual dictionaries, with a combination of Slovene and their mother tongue,
- DSLL is the only monolingual dictionary they consult,
- DSLL is used by more advanced learners, speakers of other Slavic languages and linguists,
- dictionaries are used in different activities, most often when writing, translating and reading,
- the majority of learners would use a monolingual dictionary for non-native speakers, if available.
- many language teachers use monolingual dictionaries of Slovene in class, and during different activities, especially in translation exercises and those related to lexis (searching for meanings, examples, phrases, synonyms, word families etc.),
- the majority of teachers think that a monolingual dictionary for non-native speakers is needed,
- a monolingual dictionary could be used earlier in language learning (according to teachers, a monolingual dictionary for non-native speakers could be used at lower levels, e.g. A2–B1, whereas a general monolingual dictionary, such as DSLL, could be used at B2 level of Common European Framework of Reference for Languages),[11]
- such a dictionary for non-native speakers would be more suitable for different class activities, including writing and reading,
- many teachers believe that such a dictionary should above all contain simple definitions and many examples, and should be available in electronic format,
- teachers think that learners mainly need information on meaning, usage and grammar,
- the most frequent errors of language learners observed by the teachers are related to syntax and collocations.

---

11 The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (2001). Cambridge: Cambridge University Press. 15 July 2014. http://www.coe.int/t/dg4/linguistic/cadre1_en.asp. CEFRL describes language competencies at six levels, with A2 being the second level, and B1 the third level.

3   DICTIONARY USER RESEARCH OUTSIDE SLOVENIA

An overview of research on dictionary use around the world reveals a situation that is the exact opposite of that in Slovenia, as the research is dominated by studies of languages learners, while few examine dictionary use of native speakers. In addition, the subjects of a majority of the studies are students, mainly of foreign languages or translation, or else linguists or language teachers.

Technological progress has prompted a shift in research focus to examining the differences between the use of different dictionary media. As early as the 1990s, Leffa (1993) compared the use of electronic and paper dictionaries among primary school students, and found that they translated the focal texts better and faster when using an electronic dictionary. In addition, 80% of students preferred using electronic dictionaries. A similar preference was shown by L2 students of Spanish in Aust et al. (1993), which, among other things, pointed out that one of the advantages of electronic dictionaries over paper ones is the number of look ups that can be conducted within a given timeframe. Similar conclusions have been reached by Nesi (2000), Corris et al. (2000), Tono (2000), Laufer (2000), Winkler (2001), Laufer and Levitzky-Aviad (2006), Petrylaite et al. (2008) and Dziemianko (2010). Some of these studies have provided other interesting findings. For example, in her study with students of English as L2, Laufer (2000) found that the results of a test on understanding unknown words improved significantly when the students were presented with a combination of translations, definitions and examples. In her study, Winkler (2001) found that the skills needed for using an electronic or paper dictionary are sometimes very different, and that the difficulties that arise when using both also differ. Also relevant in the context of the current study are the findings by Chen (2010) on the use of pocket paper and electronic dictionaries.[12] The results showed that the subjects (85 Chinese learners of English) used pocket electronic dictionaries much more frequently than paper ones. However, there were differences identified in terms of dictionary use for specific activities, and these were linked to the amount of information that could be shown at one time on a page or a computer screen. More specifically, the subjects preferred using electronic dictionaries for reading, and paper dictionaries for translating and writing.

Many studies into dictionary use focus on identifying the types of entry information most often consulted by users. The most frequently consulted types of information are definitions and spelling (Béjoint 1981; Jackson 1988; Battenburg 1989; Harvey and Yuill 1997; Hartmann 1999; Kosem 2010; Verlinde and

---

12   Pocket dictionaries are usually small, portable dictionaries (a relatively small number of entries number of entries, containing short and simplified information on headwords).

Binon 2010; Lorentzen and Theilgaard 2012), with synonyms also being consulted quite frequently. Non-native speakers also frequently consult grammatical information, collocations, examples and idioms or phrases (Béjoint 1981; Harvey and Yuill 1997). Other types of information, e.g. etymology and pronunciation, are rarely consulted (Hartmann 1999; Kosem 2010). Especially worth mentioning is a study by Kosem (2010), conducted among 444 native speakers and 169 non-native speakers studying at Aston University. The results, presented in Table 3, show that the non-native speakers consulted nearly all types of information (spelling being the only exception) more often than the native speakers did. It should be noted, however, that certain types of information, such as examples and collocations, receive much more detailed treatment in dictionaries for non-native speakers.

Table 3: Use of different entry components by native speakers and non-native speakers (1 – almost never, 2 – rarely, 3 – often, 4 – almost always; from Kosem 2010: 162)

	Native speakers (average)	Non-native speakers (average)
definition	3.44	3.56
spelling	2.82	2.73
synonyms	2.63	2.91
examples	2.45	2.92
usage and grammar	1.72	2.16
phraseology	1.66	2.27
collocations	1.49	2.15
pronunciation	1.60	2.10

There are also several studies on the words looked up by dictionary users. For example, Béjoint (1981) found that 66% of students (non-native speakers) never looked up frequent words, and similar findings were later reported by Hatherall (1984), Bogaards (1998) and Nesi and Haill (2002). These findings were not confirmed by Verlinde in Binon (2010), who analysed 55,752 searches in Base lexical du français (BLF) and found that the users looked up frequent words quite often. Similar conclusions were reached by de Schryver et al. (2006), who analysed nearly half a million searches in a Swahili-English dictionary and found a certain correlation between the corpus frequency of the words and the frequency with which they were looked up. However, as this correlation was identified for only the top few thousand words on the frequency list, the authors argued that it is impossible to predict which words will be of interest to dictionary users. Trap-Jensen et al. (2014), analysing the log files of searches in the online version of the

Danish dictionary (Der Danske Ordbog),[13] found that function words and words with high corpus frequency were among the most frequently looked up (60% out of 1000 most frequently looked up words were found among the 10,000 most frequent words in the corpus).

Research also points to a few dictionary use strategies that often determine whether the search for or interpretation of dictionary information will be successful. One frequently mentioned strategy is that of 'choose the first definition', reported by Mitchell (1983), Tono (1984), Neubach and Cohen (1988), McCreary (2002), Nesi and Haill (2002) and Kosem (2010). A similar strategy in the use of pocket electronic dictionaries has been observed by Boonmoh (2012), with the students in the study consulting only the part of the entry visible on the screen. This means that it is important to order senses with the target users' needs in mind, and use different strategies of presenting information, e.g. providing a menu at the beginning of the entry to enable a quick overview of entry senses and quicker navigation through the entry. Another strategy, used by both native and non-native speakers, is the "kidrule" strategy, in which "a short familiar segment of the dictionary definition is taken out of context as an equivalent for the unknown headword" (Nesi and Haill 2002: 285). The strategy was first mentioned by Miller and Gildea (1987) in a study conducted among 10- and 11-year-old children, and was later also found to be used by both students and adults (Harvey and Yuill 1997; McCreary and Dolezal 1999; Nesi 2000; McCreary 2002; Nesi and Haill 2002). A separate group are represented by cases when the users encounter difficulties in dictionary use, also on account of inappropriate search strategies. For example, Selva and Verlinde (2002) report on user difficulties in finding relevant information in polysemous entries and long definitions. Similarly, Tono (2011) also observed users having difficulties with searches in long entries.

The most frequently used dictionaries in dictionary use research are those for advanced learners of English,[14] mainly because these dictionaries are the main sources of lexicographic innovation, and thus most interesting for detecting new trends in dictionary use. Among the innovations introduced by advanced learners' dictionaries are defining vocabulary (first used by the *Longman Dictionary of Contemporary English*), whole-sentence definitions (introduced by COBUILD), semantic indicators or signposts (introduced by the *Longman Dictionary of Contemporary English*), menus (first used by the *Macmillan English Dictionary for Advanced Learners*) and the inclusion of information from learner corpora. The main purposes of these innovations is to help users find the relevant information more quickly, and help them with any encoding tasks. It is thus not surprising that whole-sentence definitions, signposts and menus have gradually been adopted by some monolingual dictionaries for native speakers.

13 http://ordnet.dk/ddo
14 Especially advanced learners' dictionaries.

There is another important trend worth pointing out in this context, and this is that the online format is becoming predominant. In fact, it is difficult to find a contemporary dictionary without an online version that is released in addition to the paper one. In fact, several publishers have started to stop making paper versions (e.g. in 2012, Macmillan announced the end of their production of paper dictionaries, and has focussed solely on making online dictionaries; Rundell 2014). But the online versions of dictionaries have become much more than simply dictionaries offered in a new format; they have turned into portals offering access to several reference works (dictionaries, thesauri, and so on) and different types of information on language, e.g. blogs on certain aspects of language use, notes on frequent errors, multimedia content, etc. In this way, a dictionary is becoming a part of a language-didactic service. Interesting for educational use is the Wordsmyth portal,[15] offering access to children's, illustrated and school dictionaries for native speakers of English, as well as tools for solving anagrams and crosswords, and for making glossaries and quizzes. These are useful for both students and teachers, and can be used in class.

4 DISCUSSION

So what do we know about the needs, abilities and habits of dictionaries users, non-native speakers and young native-speakers of Slovene, and how can we use this knowledge in planning dictionaries?

4.1

Learning about dictionaries and how to use them is part of the Slovene syllabus, materials related to or including dictionaries can be found in textbooks, and Slovene teachers do not use dictionaries (especially DSLL) only when preparing teaching materials and marking student work, but also in class. Studies show that young native-speakers of Slovene use dictionaries in school when learning about different syllabus contents and during different activities, especially when producing text. In school, a dictionary is therefore not only a reference resource with various information on language, including normative details, but is also an important didactic tool. We can assume that a (descriptive) dictionary made with the needs and abilities of school students in mind would be even more widely used in education, and would have a considerably greater impact on the development of students' communication skills. In order to achieve this goal, we need to

---

15 http://www.wordsmyth.net

abandon the traditional notion of a dictionary as a separate resource, and think about the development of an online portal that would offer several types of lexical information. As shown by the analysis of Arhar Holdt and Rozman (2015),[16] we need to apply corpus analysis and the analysis of syllabi to detect the language problems of students and devise resources that will not only alert users to such issues, but also offer ways to solve them. Moreover, standard dictionary contents need to be accompanied with explicit (more "educational") explanations, and with information on grammar, orthography, norm, stylistics, collocational characteristics, differences between synonyms or related words, and options to compare different words or their meanings. Other useful content includes quizzes, exercises, multimedia material, lists of common language problems, lists of word families or semantic types and so on. All these types of information and tools will facilitate students' language acquisition, as it will be easier for students to link new information with existing knowledge and include it in their mental lexicon (see, for example, Rozman 2010: 32). In addition, a combination of information on words and exercises will facilitate the development of strategies for vocabulary acquisition (Paynter et al. 2005: 30–68). Explicit explanations for problems related to norm and usage, which go beyond the existing practice of presenting language use as black-and-white (right and wrong) (Stabej et al. 2008), are extremely important for improving students' understanding of the complexities of language (and with that, their communication skills).

4.2

Having such an informative and didactic value, as described in 4.1, the envisaged dictionary would also be more appealing for students' independent use. Research shows that dictionaries, even online dictionaries, are currently rarely used by Slovene students, yet the internet is frequently used to find language-related information. We do not know the reasons for this, but can assume that one is related to the low user-friendliness of online dictionaries and their entry structure, with the information in these often too condensed, poorly structured and difficult to understand.

Empirical studies, as presented in 2.4 and 2.5, focussed on certain components of dictionary microstructure, have shown that DSLL, as the most widely used dictionary in Slovene education, is often too difficult to understand, especially for younger students (i.e. those in primary schools). The way that grammar information is coded in the entries makes their interpretation very demanding for students, and the findings show that it is much more efficient if grammar

---

16 Similar practices can be observed outside Slovenia, e.g. Vocabulary.com and Merriam-Webster.com.

information is provided in the form of a label, example or in a specially dedicated section on usage. Nonetheless, a larger-scale survey is needed to determine not only which grammar information in the dictionary should be made more explicit, and how, but also to identify the best ways of presenting such information (on grammar and other characteristics of words).

As far as definitions are concerned, they need to be devised by considering the development level of students, as evidenced by the rewritten definitions in Rozman (2010), which were produced based on this approach and proved more understandable than those in DSLL (see Section 2.5). Students' vocabulary size and structure, as well as their understanding and knowledge of abstract meanings, the relationships between different words and meanings, longer and more complex syntax, morphology and word formation, all improve over time, partly due to mental and cognitive development and partly due to (language) education in school (see Rozman 2010). However, as Rozman's study focussed only on the understandability of definitions, it does not provide the answer to the very important question of whether definitions aimed at younger students are also suitable for older ones.

Rozman's study points to several characteristics that affect the understandability of a definition, such as: indirect definitions are better than direct ones; definitions with common words are effective in most cases, except when the words used are highly polysemous and reduce exactness and concreteness; and definitions should not contain rare (terminological) words, and should not be too abstract. Moreover, the study offers some suggestions how to approach the sense division of polysemous words, namely that the students have trouble understanding the meaning of the word in the dictionary if the entry contains closely linked senses with complex and abstract definitions.

4.3

Existing research provides some valuable guidelines for planning a dictionary suitable for students and when used as a teaching aid. Nonetheless, several questions remain. One of these concerns the treatment of function words, which has not been addressed by researchers other than Arhar Holdt and Rozman (2015), whose findings point to the need to replace or improve existing dictionary definitions with more functional or grammar-oriented ones.

Also missing is empirical data that would help with the creation of headword lists. Language acquisition theories suggest that during education an individual's vocabulary expands mainly in terms of multi-syllable, abstract and specialized

words, and later also with less frequent and more specialized words. However, limiting the headword list to or focussing dictionary treatment on these words is probably too narrow of an approach, as analyses of the Šolar corpus indicate that students have many problems with general words, especially during language production activities (Kosem et al. 2012a; Arhar Holdt and Rozman 2015).[17] This shows the need for further corpus-based and related research into students' language problems, and for a list of words used in textbooks and other school materials. Nonetheless, even without relevant research it appears that the headword list should include words that exhibit a certain level of semantic transfer, usage different from the regular patterns, words with variants, and words that are semantically or morphologically similar, i.e. words that are likely to cause problems for students (with such efforts also based on analyses of the language problems of adults).[18]

4.4 In sections 4.1–4.3 the focus was on native-speaking primary and secondary students of Slovene. What can be said about non-native speakers of Slovene and a dictionary that would meet their needs? First and foremost, non-native speakers are not a homogenous group – they differ in terms of their L1, proficiency level in Slovene, mode of learning (language course, study course, etc.), and location (in Slovenia or abroad). Language learners also differ in terms of needs and motivation, which are closely related to their learning interests and aims. Nevertheless, these differences can still be successfully addressed by a learners' dictionary, as evidenced by advanced learners' dictionaries of English (see Section 3), which are even a source of lexicographic solutions for dictionaries aimed at native speakers. Advanced learners' dictionaries of English are thus a good model for a dictionary of Slovene for non-native speakers, and possibly also for younger native speakers. However, current information on non-native speakers of Slovene is even more scarce that on young native speakers, as there are very few research studies in this area. In addition, existing learner corpora of Slovene are rather small,[19] and do not enable any comprehensive analysis of non-native speaker writing.

As far as Slovene as L2 is concerned, there is plenty of work for Slovenian lexicographers who can also benefit from the fact that the teaching of Slovene as L2 is a well-developed field. There are thus established methods of teaching, acquiring and learning Slovene vocabulary, documented in various textbooks and other didactic resources.[20] Having information on what vocabulary is taught to non-native speakers (and in what ways) would be of great help in preparing a dictionary

---

17 Similarly, international research studies do not provide a straightforward answer on which words, more frequent or rare ones, are more often looked up by the users (see section 3).

18 For example, see Bizjak Končar et al. (2011).

19 The only learner corpora of Slovene in existence are a corpus without annotated errors containing 32,117 words in 306 texts (Rozman et al. 2010), and a learner corpus called piKUST (Stritar 2012) and containing 34,873 words in 128 texts, as well as annotated errors (5,085 in total).

20 For example, in different workbooks and texts, available at the CSDTJ website.

for such users, as well as the related headword list. Another resource for devising a headword list would be lists of words found in textbooks and other materials, such as Sporazumevalni prag za slovenščino ('The Comprehension Threshold for Slovene'; Ferbežar et al. 2004), which describes knowledge of Slovene at level B1 of the Common European Framework of References for Languages. Sporazumevalni prag za slovenščino also groups words according to topics, and categorizes them into semantic groups pertaining to time, space, measurements and so on, which can be useful information for non-native speakers. Finally, the headword list could also draw information from the vocabulary used in language proficiency tests.

5   CONCLUSION

The overview of dictionary use research in Slovenia and a discussion on the needs of native-speaking school students and non-native speakers of Slovene, as presented in this work, make us wonder whether it is possible to make a single dictionary that would meet the needs of both types of users. The question is interesting in relation to the use of dictionaries in education and language learning. Acquisition of L1 and L2 vocabulary are two different processes, although they have several common aspects (Singleton 1999: 79–82; see also Jesenovec 2004). In order to identify the common aspects that can be addressed in a dictionary, more research and user studies are needed. It is also essential to consider the didactic aspects of such a dictionary, or dictionaries, as solutions related to this would be highly relevant and useful for both types of users.

There is another option worth considering, namely whether a contemporary general dictionary could be suitable for native- and non-native speakers. This opposes the general argument of this paper, although a few findings prevent us from completely dismissing this idea. Firstly, all the studies into general dictionary used DSLL, a dictionary that is outdated and found to be difficult to use even by adult native speakers of Slovene[21] (Kosem 2006: 26; see also Müller 1996 and 2009). Secondly, teachers think that a general dictionary is suitable for these two types of users, and although we do not agree with this view, we cannot deny that the use of such a dictionary in certain teaching situations can be useful. Finally, a general dictionary compiled with state-of-the-art methods would take into account findings about common look up strategies, address frequent language problems of users, and consider the needs of school students and non-native speakers, and thus could be much more suitable for language teaching and learning than DSLL. Moreover, if available in the form of a portal,

---

21   There is no empirical evidence to support this claim, as there are no studies into understandability of DSLL, conducted among adult native speakers.

the dictionary could provide didactic content (in a separate section), which may be less interesting for other users but would be very important for students and non-native speakers.

Neither of these questions (the possibility of compiling a single dictionary for both types of users; the use of a general monolingual dictionary) can be answered with a clear yes or no, as current lexicographic research in Slovenia does not offer enough evidence to support any answer. As there is a certain overlap in the needs of both types of users (young native speakers and non-native speakers), and as digital media formats offer the possibility of combining different lexicographic solutions, it makes sense to think about compiling a common dictionary database containing information relevant to all types of users, and information relevant to individual user groups.

Such a database then offers various possibilities, e.g. we can compile several dictionaries for different types of users, or a portal containing (carefully structured) information for all types of users, both, or something completely different that we have not yet considered. Another benefit of such an approach is that in the meantime more empirical studies can be conducted, which can provide much information on which to base our decisions on.

Creative Writers as Dictionary Users: Creating in Language and with Language

Vesna Mikolič

Abstract

In this article, we present a pilot survey among users of language manuals, with a focus on people who deal creatively with the language as part of their work, such as writers, scientists, journalists and advertisers. We were interested in how their language awareness is shown through their need for dictionaries and other language manuals. The results indicated that all the people observed in this study at least occasionally used language manuals, with both traditional printed and online versions being consulted. The usage partly differs from group to group, and it also depends on the age of the person involved.

Keywords: language manuals, dictionary, user survey, language awareness, creativity.

1   COMMUNICATIVE INTENTIONS OF THE CREATORS

The users of language manuals for professional purposes are a very broad and diverse group. In our short survey we captured those occupational profiles which are characterized by the regular production of texts, which are intended for the general public and are reflecting, to a certain extent, the authors' creativity. The works of literary artists, scientists, advertisers and journalists, as taken into account in the survey, differ in their purposes, even if creativity is a common element among them.

In the analysis of this creativity we can consider the theory of speech acts and the division of human speech activities on the basis of the four basic illocutionary roles, i.e. cognitive, communicative, executive and art-expressive (Mikolič 2007; Skubic 1995; 2005). The purpose of literature is art-expressive. Writers declare their subjective view of reality with their own aesthetic expressions. The more the writer's world is unique and multifaceted, the more valuable is their literature. This is of course on the condition that the literature is in itself coherent and convincing. The author is interested in the reader only in the second stage, wanting the literary work to have a life of its own. Scientists are also not primarily driven to carry out research by a desire to communicate knowledge to their readers, but instead by taking a creative attitude to the existing reality, in which the scientist sees all the time new ideas and new challenges, and as yet unexplored areas. The basis for scientific discourse is therefore also subjective and contains a great deal of creative thinking, because the scientist must be able to look at already known facts differently, with a new and perhaps previously inconceivable point of view. However, unlike literature, science belongs in the cognitive field, because the scientist's primary purpose is to explore unknown aspects of reality and their relationships to the whole, to expand and create new knowledge. Scientific activity thus has a primarily cognitive purpose. The scientist must confirm and externalize their subjective insights in the external-language reality by carrying out a survey based on evidence. Nevertheless, the scientist remains creative in their use of methods, with the possibility of linking disjointed and seemingly incompatible ideas into new, creative and insights. Journalistic and advertising discourse are communicative activities in the narrow sense of the term. Their basic feature is being oriented to the recipients, and their primary purpose is to communicate with the recipients, to convey a certain message to them. However, in journalistic speech the need to inform the recipient should prevail. On the other hand, advertising speech wants primarily to convince the recipient about the positive features of whatever the text is referring to. As such, when journalists function as advertisers they are looking for innovative and effective communication strategies. This is especially important for advertisers, where creativity is now the

central element of a successful advertising, with all other elements derived from this (Jewler and Drewniany 2005).

As we have seen, the categories of discourse examined in this work vary by the extent of creativity used, as well as the creative methods that are applied. Unlike a literary work, which tends to be more appreciated the more the author is coherent in their own subjective world and form of expression, scientist needs to externalize their initial innovative and subjective view of the problem reality in the course of their research. A similar difference can be found between journalists and advertisers. The latter, working in accordance with the purpose of advertising, have more creative freedom, although professional ethics dictate that advertisers should respect the truth of some objective conditions. Nevertheless, for all four categories of authors analyzed in this study we can say that in their language production they are, at least to a certain extent, subjective and creative; moreover, the texts that are produced on the basis of this creative view, are – sooner or later – intended for a wider audience. On this basis, we can predict the authors will have a special sensitivity for language, and thus we are interested in how this linguistic awareness is shown through their need for dictionaries and other language guides.

2 OBJECTIVES AND METHODS USED IN THE PILOT RESEARCH

With the rise of sociolinguistics in the 1960s came the assertion that language communication is always interactive and intended for an actual or potential recipient (Schiffrin 1987). Moreover, around the same time the field of user research began developing, which was interested in the users of language manuals. This is of course understandable, since such manuals were text-based ones, and thus explicitly and primarily intended for users. However, it is surprising that in the Slovenian context we have not paid much attention to this topic so far, except for a few works that have recently started to note the need for more research into dictionary users (Logar 2009; Stabej 2009). Undoubtedly this call for a change in focus, to a greater orientation on the user, can also be seen in some recent projects that are interested in users of the language more widely, and so aimed to carefully monitor their needs. For example, the researchers of the Communication in Slovene project[1] wanted to find out which aspects of the Slovene language cause problems to writers. Based on this work, the project developed a style guide and a number of other online language resources and tools, focussed on the needs of language users. Moreover, various online resources are also available to language users, such as the portals Fran and ŠUSS. Due to the rapid development of

---

1   http://eng.slovenscina.eu

communication technologies and the profound changes in the nature of communication in recent years, it is especially important to establish regular monitoring of user needs in these new contexts.

In this paper we thus present a pilot survey among users of language manuals that work in creative professions. The survey can form the basis for further studies in this area and the regular monitoring of the working methods of such users of the language, as well as their needs in terms of language manuals.

The initial question was as follows: How is the linguistic awareness of individuals, who in their careers deal creatively with the language, shown through their need for dictionaries and other language guides?

For this purpose in the spring of 2015 we carried out semi-structured interviews (half in person, half via e-mail) with 30 individuals. In these we asked the following questions:

1. Do you use the language manuals in your work?
   1. 1. If yes, which ones?

2. With what purpose or for what issues do you reach for them?

3. Do you know any online language manuals or tools? (dictionaries, corpora, etc.)
   3. 1. If yes, which ones?

4. Have you had any language problems or questions to which the language manuals have not given you the answer?

5. What do you hope for from a dictionary of the Slovenian language? What do you think it should contain?

The respondents in both the face-to-face and e-mail interviews were encouraged to freely share their true opinions. Some respondents used linguistic terminology associated with the use of dictionaries and other language guides more than others. In the subsequent analysis we present examples of their responses, while at the same time we have summarized their answers into linguistic categories, so that the results can be presented in tables and figures.

2.1   Description of sample

Among the 30 subjects that were interviewed, there were ten literary creators, ten scientists, five creatives/advertisers and five journalists, all of various ages (the youngest from 20–35, the middle group from 35–50, the older ones aged over 50)

and genders, and from different areas in Slovenia (from Koper, Piran, Ljubljana or Maribor). Among the literary creators were poets, prose writers and dramatists (three women and seven men). The scientists were from the areas of the humanities (but not linguistics), and social and natural sciences (six women and four men). With regard to the five journalists, there were two working in the print media (both women), two in radio (one woman and one man) and one in television media (a woman). Among the creatives there was one working in an advertising agency associate (a woman), a designer with the status of a freelance artist (a woman), two PR managers (working for large companies, a man and a woman) and a retiree (a woman), who was previously employed in the PR department of a large company.

Since we used a small sample, the findings naturally have limited value. However, some interesting characteristics with regard to the linguistic consciousness of the respondents and some of their similarities and differences were found. As mentioned earlier, in future work it will be necessary to monitor these features in a more in-depth manner.

3 THE RELATIONSHIPS BETWEEN THE WRITERS AND THE LANGUAGE MANUALS

It is interesting that we quickly recognized two extremes in this group of users; on the one hand, there were regular users of the language manuals, and on the other there were those who prefer to rely on their own language intuition, and do not use language manuals, or do so only in exceptional cases.

It seems that this is partially generationally determined. Users from the older group, and some from the middle one, regularly used the printed versions of *Slovene Orthography* (SO), the *Dictionary of Slovene Literary Language* (DSLL), and dictionaries of foreign words, and one respondent stated that they occasionally open Toporišič's grammar and Bezlaj's or Snoj's etymological dictionaries. Respondents from the older group, and some of the middle one, did not know about any online language manuals or resources. They tended to search the Internet only with the use of search engines, if interested in any specific language use.

The rest of the middle age group, and all of the youngest one, used language manuals very rarely, but among those they do use are printed dictionaries such as Verbinc's *Dictionary of Foreign Words*, Oxford's *Dictionary of English*, and online manuals like DSLL, SO, and Wikipedia. They did not know any other online dictionaries and resources (such as, for example, corpora).

The respondents stated that they reached for language manuals when they are writing and translating, and one author stated that they did so to enhance their

education. They used the language manuals for looking up spelling and grammatical information, sometimes to find the accurate meaning or the usage or formation of a neologism (i.e., they are curious as to whether the word already exists in the dictionary). They also tended to look up word-formational features in order to learn the language rules for the formation of neologisms. Finally, the respondents were also interested in stylistically characterized words, ambiguity and in rhythmic texts where the accent appears in a word.

However, in general the respondents did not remember having any specific linguistic questions for which the language manuals were not able to give an answer, although some argued that they sometimes they did not agree with a suggested spelling or that they could not find a word in a dictionary. One author also noted that dictionaries often omit some non-standard words that are retained in dialects, and are also part of the Slovenian language.

More revealing are the respondents' expectations for a dictionary of the Slovenian language, which can be summarized as follows: the dictionary should be easy to use and comprehensive, the descriptions and examples given should be extensive and originate from the living speech, and be accessible in both printed and online forms. These expectations are described in the following statements:

> "In dictionaries, for example, I miss many of the words that are in all respects completely Slovene, but they may be maintained only in one dialect. I am specifically interested in "linguistic archaeology", therefore I am in search of the hidden archaic remains, even treasure in one language. This could be for example the language character, the spiritual foundation of this character, this "spirit" of language, which can be seen in lexical roots and other phonemes, also syntax, etc. Linguists or etymologists prefer to avoid these components, and this is perhaps from a scientific perspective completely excusable - while literary writers are often looking for this "magical" vividness, because the language is the live tradition, the medium through which the literary writer "appeals" to spirituality, the spirit of already long-dead generations who created this language. Linguistics of course mainly remains silent about this "spirit", which cannot and must not mean that this is not present in the language. In my opinion the same is true for the essence of each language and its creative use, which mainly evade linguistics."

> (A respondent who is a literary writer from Ljubljana, born in 1958)

> „/…/ that there will be no artificially produced words in the dictionary, but it will follow the folk, beautiful Slovenian language, and it will not rearrange and invent words. /.../ I take the dictionary as some other opinion, and not as the absolute truth."

> (A respondent who is a poet from Piran, born in 1984)

4   THE RELATIONSHIP BETWEEN SCIENTISTS AND THE LANGUAGE MANUALS

The researchers all occasionally used language manuals. They sometimes used the printed versions of SO and DSLL, and knew and used several online language manuals and resources, such as the online DSLL, SO, PONS and other bilingual online dictionaries, dictionaries of English and other foreign languages (German, Italian), dictionaries of classical languages (ancient Greek), terminological dictionaries (geographical), the multilingual terminological database Evroterm, Google Translate, Amebis Presis, Besana, and Termania. However, they did not know about corpora of Slovene or foreign languages. The middle and younger generations of scientists used almost exclusively online language manuals, as well as visiting online services and forums for advice and opinions on language matters.

These scientists sometimes needed language manuals to find spelling and grammatical information, look for suggestions for synonyms and alternative terms, but in general their most common needs were related to the formation and use of terms. And thus in the translation of the technical literature into Slovenian these authors were looking for Slovenian counterparts for foreign terms (e.g. Eng. aspiration economy – *ekonomija učinkovitosti*), looking for the appropriate lexical roots for the formation of new terms (e.g. Eng. citizenisation – *državljenje*) and in the case of terminological doubles they were looking for the Slovenian denominations. Moreover, these respondents used language manuals, especially bilingual dictionaries, or dictionaries of foreign languages, when writing in a foreign language. Sometimes they look for a translation of a term, and sometimes only check the format of the words. The problems that they cannot solve with the help of language manuals are also linked with translation. For example, sometimes they are not able to find the relevant translation or they do not find a sufficient explanation to choose a new term. In particular the respondents noted that they are disturbed when the meanings given are too general, imprecise and not professional enough, with dictionaries often not taking into account the multidisciplinary use of a term. The respondents thus hope that a dictionary of the Slovenian language would come with many examples of use, that these would make use of more complex sentence structures and be presented in different contexts and areas, so that it is easy to see the various peculiarities of meaning and use. The researchers also noted the importance of a user-friendly and clear structure for the dictionary, while at the same time noting that the possibilities of new technologies should also be taken into account. Their expectations for a dictionary are well summarized in the following statements:

"I don't know, I have never thought in that way … I want that the dictionary is up-to-date, which means, that it contains also the most modern words, maybe professional terminology, and also foreign words …"

(A respondent who is a scientist working in the field of social studies from Koper, born in 1971)

"The dictionary should present different contexts of words. Also the form is important, which should be clear and manageable. At the same time the dictionary should be interactive, using all the opportunities that are enabled by new technologies."

(A respondent who is a scientist working in the field of humanities from Ljubljana, born in 1977).

"The basic concept of the dictionary should be interdisciplinary, so it can be used in a wide range of areas, because real life is wide and it is not limited to single disciplines. To solve most problems it is necessary to work in an interdisciplinary manner (e.g. for issues related to water and wood) and the appropriate terminology should be used in these fields."

(A respondent who is a scientist working in the field of social studies from Maribor, born in 1967).

5 THE RELATIONSHIP BETWEEN JOURNALISTS AND THE LANGUAGE MANUALS

The journalist respondents were also quite regular users of language manuals in both paper and online forms. They mainly use DSLL and SO, and also occasionally use the style manuals of Janez Gradišnik, bilingual dictionaries, various law manuals, local lexicon, and various encyclopaedias, and sometimes visit Wikipedia and forums, where their language problems can be solved. Only one respondent from this group stated that she also searched corpora (e.g. Gigafida). The youngest group of respondents exclusively used online language manuals and resources, although even they did not know about the most up-to-date online language portals, which collect several manuals and language resources together.

The respondents in this group stated that they used language manuals when writing articles. They were mostly interested in spelling and grammar, e.g. the declension of the foreign names and the use of upper-case. They often reached for language guides when they wanted to define the relationship between a new term and the linguistic norm. Sometimes they are also interested in terminology, especially legal terms. The journalists working in radio and television media

were also interested in the correct place of the accent in multi-syllable words. One of the journalists also highlighted the use of language manuals in her free time or with her family (e.g., helping a primary school child and student, or proofreading diplomas).

When these authors encounter something that is incomprehensible they tend to turn to a proofreader (if they have one) to ask for help, or to their journalistic colleagues. They stated that dictionaries often do not include newer lexis, and noted that they often could not find the words they were searching for.

With regard to what such users wanted from a dictionary of the Slovenian language, they wanted it to be user-friendly and concise, because due to the nature of their work where they do not have much time to resolve any language issues, and thus need answers to their questions very quickly, as noted in the following statement:

> "The tempo of my work is extremely fast and sometimes I simply don't have the time to looking into a specific question and find a solution."
>
> (A respondent who is a journalist from Koper, born in 1987)

In addition, they want the dictionary to be up-to-date and that its authors should thus monitor the development of new vocabulary, which can then be presented with the aid of new technology. In particular, the dictionary should pay attention to the needs of the language user, as journalists must be oriented to the needs of their readers, listeners and viewers, as illustrated by the following statement:

> "As journalists we have an obligation to bring specific and also technical issues to the reader, and this we can only do with understandable and stylistically appropriate language, therefore in dictionaries we want transparent explanations of the senses and the actual use of words."
>
> (A respondent who is a journalist form Koper, born in 1964)

6 THE RELATIONSHIP BETWEEN ADVERTISERS AND THE LANGUAGE MANUALS

The respondents who were advertisers/creatives were fairly regular users of the language manuals. Those in the middle and older groups regularly used the printed versions of SO, DSLL, dictionaries of foreign words, an English-Slovenian dictionary, and occasionally an etymological dictionary. Other users in the middle group, and all those in the youngest group, used the same manuals online, as well as dictionaries from the online portal of the Fran Ramovš Institute for the Slovenian language ZRC SAZU, *The Tongue Unleashed* (*Razvezani jezik*) – a free dictionary of the living Slovenian language, some terminological dictionaries

(e.g. for the theatre), bilingual online dictionaries, dictionaries of English and other foreign languages, and Google Translate, but they did not use corpora.

Language manuals were used primarily to find spelling and grammatical information, look up explicit meanings and stylistic characterizations, and for finding synonyms to avoid foreign words. One of the interviewees from this group stated that advertisers are aware of the trend for the excessive use of foreign words in advertising, which has been led by the desire to be special, different, and more visible.

Sometimes these respondents stated that they encounter an unsolvable language problem, because the dictionary does not provide sufficiently comprehensive explanations of meaning or enough examples of use, so they turn to authentic texts or asked a proofreader or translator. Their expectations for a dictionary of the Slovenian language are thus that it would give enough information about words, so that it can be used in a creative process of searching for creating advertising ideas. They are also aware of the impact of globalization on advertising, and in using the Slovene language also see the possibility of deviating from the standard examples. The Slovene language, with its own cultural specifics, gives them the possibility of engaging in different, creative thinking. At the same time they also want the dictionary to be user-friendly and concise. All of this is summarized in the following statements:

> "It should be manageable, functionally designed, and logical for the usage – 'simple and logical'."
>
> (A designer with status of a freelance artist from Koper, born in 1964)

> "The power, the weight of each word should be shown. /.../ More emphasis should be given to the cultural specifics, also to the etymology, so that advertisers would paid more attention to the relevant transfer of globalized advertising strategies and content in Slovenian culture. "
>
> (A retiree, previously employed in the PR department of a large company, from Ljubljana, born in 1946)

7   CONCLUSIONS

All of the speakers and writers surveyed in this work, who deal creatively with the language, no doubt have a developed and active language awareness. Although they are not mainly concerned with metalanguage issues and also do not think directly about what language manuals should be like, they often think about language and appropriate ways of expression, and thus often need to look for answers to problems that arise in different places, including language manuals.

There are clear generational differences with regard to the respondents' use of language manuals in this study. The older respondents only used printed language guides, those in the middle generation used both printed and online manuals, while the youngest group tended to use only online language manuals.

All of the groups mainly know and use the standard printed and online language manuals (SO, DSLL, dictionaries of foreign words, bi- and monolingual dictionaries). However, only a few know the more advanced web portals that offer a variety of language resources and manuals, such as the Communication in Slovene, the portal of language resources, Termania, Fran, FB-portal Language Slovenia, and so on. Moreover, only one interviewee knew about using corpora (see Tables 1 and 2).

Table 1: The language manuals use of the respondents

Do you use language manuals in your work?	Literary authors	Scientists	Journalists	Advertisers	Total
Yes	6	10	5	5	26
Rarely	3	0	0	0	3
No	1	0	0	0	1
Total	10	10	5	5	30

Table 2: Use of online language manuals and tools

Do you know any online language manuals or tools?	Literary authors	Scientists	Journalists	Advertisers	Total
Yes, I know them, including corpora and language portals	0	2	1	1	4
Yes, I know different tools, but I don't know corpora and language portals	3	7	1	3	14
Yes, I know them, but only the online versions of printed works	2	1	2	1	6
I know some, but rarely use them	3	0	0	0	3
I don't know any of them	2	0	1	0	3
Total	10	10	5	5	30

The respondents stated that they used the language manuals mostly to find spelling and grammar information. Moreover, the most common reasons for using

such tools were to find more details of the meaning of words, and to search for examples of use. These reasons are followed by interest in newer lexis, translation counterparts in Slovene, synonyms or professional terms (see Table 3).

Table 3: Reasons for using language manuals

With what purpose or what issues do you use these resources?	Literary authors	Scientists	Journalists	Advertisers	Total
To find explanations of meanings	4	3	2	2	11
To find explicit descriptions of meanings	6	8	0	5	**19**
To find out the stylistic characterization of a term	3	0	2	4	9
To find neologisms and newer words	2	6	5	3	**16**
To find spelling and grammar rules	6	8	5	5	**24**
To find synonyms	4	6	0	3	**13**
Ambiguity	1	2	0	0	3
Examples of use	5	6	3	5	**19**
Multidisciplinarity	0	5	1	0	6
Technical terms	0	9	3	0	**12**
Translation counterparts in Slovene	0	10	2	2	**14**
Etymology	1	2	0	2	5
Word-formational features	2	3	0	0	5
Accent	2	0	3	0	5
Total	36/10=**3,6**	68/10=**6,8**	26/5=**5,2**	31/5=**6,2**	161/30=**5,3**

With regard to the expectations that the respondents had as dictionary users, we can see that all groups felt that the existing dictionaries – as well as other language resources – were not always able to solve their linguistic problems. In particular such works tend to lack more extensive and authentic examples of use, and a wider range of vocabulary. Moreover, the also often lack newer and non-standard vocabulary. In addition, all of the respondents stated that the first characteristic that they expect from a dictionary is being easy to use. The key elements that the respondents felt are missing in current dictionaries, and which the hope to be included in future versions, are shown in Figures 1 and 2.
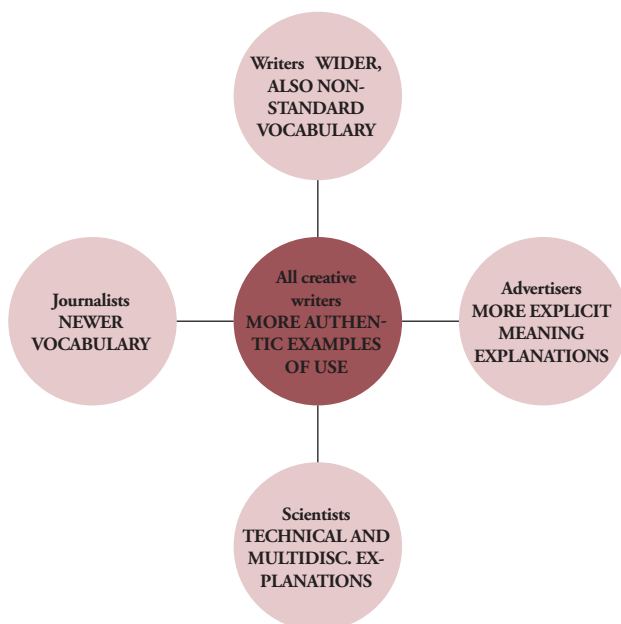
Figure 1: What do creative writers feel is lacking in current dictionaries?
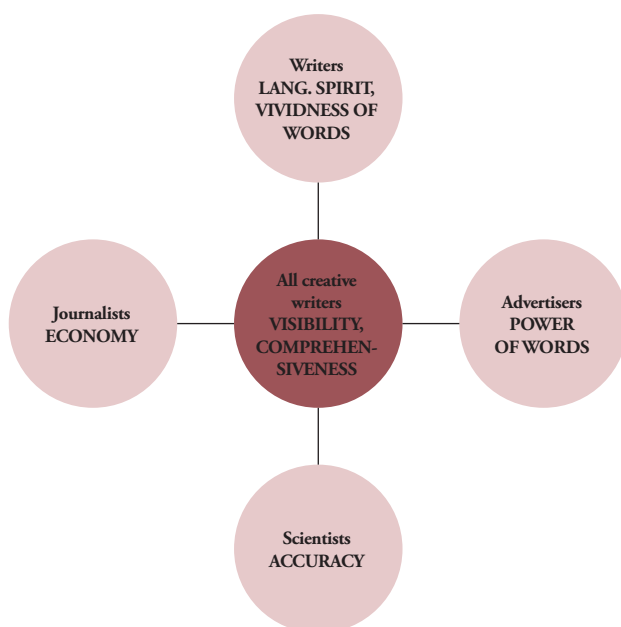

Figure 2: What do language creators hope for from a dictionary?

Regarding the use of language manuals and the expectations of dictionary users, although this study only had a small sample of respondents, it still found some interesting and notable differences among the groups. In Table 3, which shows the number of times each reason for using a language manual was cited, it can be seen that each respondent reported an average of 5.3 different reasons. The results also show that the scientists reported an above-average use of dictionaries (an average of 6.8 times) followed by advertisers (6.2 times), while the journalists are close to the average number (5.2 times). Finally, literary authors used the language manuals least (3.6).

Overall, it seems that the scientists knew most of the newer online language portals. Moreover, they reported using the language manuals primarily to form and understand professional terms, and thus want to obtain accurate explanations of words from the perspectives of different disciplines and areas. This group also emphasized the importance of making use of interactivity and other features of language manuals which are made possible by new technologies.

In contrast, the advertisers were primarily looking for inspiration for their creative ideas in the language, and thus in the language manuals used. They thus want to "feel" the words, their *power* and specifics, so that they can use the language in ways that deviate from the prevailing unified and globalized advertising patterns.

Literary creators are divided into two groups: they are either regular users of the language manuals or rely on their own language intuition and hardly use such works. The latter group are also the only one that stated that they very rarely or never use language manuals. However, all of these respondents were interested in the language, wanting to feel the spirit *of the language and the vividness of the words*, and thus they also wanted related details in a new dictionary of the Slovenian language.

All the journalists stated that they occasionally used language manuals. However, because of the nature of their work they cannot spend a lot of time using them. With regard to their expectations for a new the dictionary of the Slovene language, they stated that it should be user-friendly and *concise.* Journalists were also the respondents who most felt that current dictionaries lacked newer vocabulary items. Like some literary creators and advertisers, some journalists also rely on proofreaders to improve their writing. This is understandable, as larger media companies, as well as publishing houses, have organized proofreading services.

In conclusion we can say that the respondents examined in this work, who are both language users and users of language manuals, can actually be understood as language creators, who pay a lot of attention to language and its possibilities. The language infrastructure should thus follow their interests and needs, as this could then help to develop and expand their linguistic consciousness. The first condition for this, however, is a greater knowledge of these interests and needs, and so continued monitoring of user needs should be one of the main tasks of Slovenian lexicographers.

Reference corpora revisited: expansion of the Gigafida corpus

Nataša Logar

Abstract

The paper discusses the expansion of the Gigafida corpus, a reference corpus of Slovenian. In order to become an even better source of language data for a new explanatory monolingual dictionary of modern Slovene, the Gigafida corpus should first be supplemented with texts from the period 2010–15 and, if possible, 1990–95. In this respect, the issues of copyright and open access to corpus texts are important, as well as issues pertaining to the criteria for the text collection process and the proportions of text types. At the end of the paper, arguments are presented for increasing the number of textbooks in the corpus, and a proposal outlined for a new taxonomy which includes topic/domain categories.

Keywords: reference corpus, Slovenian, dictionary

1   INTRODUCTION

Corpus linguistics is founded on the idea that language is primarily a social phenomenon, and as such it manifests itself exclusively in texts, which can be described and analyzed (Teubert 2005: 108). Therefore, the focus of corpus research is primarily performance (and less so or not at all competence) and observation of the language in use, which then leads to the production of theory (and not vice versa) (Kennedy 1999: 7; Leech 1992: 107). In this context, corpus linguistics differs from research approaches to language that are based on introspection, and from linguistic conclusions without evidence (Kennedy 1998: 8). Corpus linguists are not interested in which words, structures or uses of the language are possible, but rather in what is more likely to occur in a particular language, what is more frequent and typical in it, as well as what is linguistically unique or special about it. In the last three decades, corpora have become a fundamental source of data for linguistic descriptions and justifications, particularly in any modern lexicography.

"The collection of linguistic data for the dictionary must correspond to the concept, to the design of the dictionary. The relevance of the data in relation to the concept is of fundamental importance," argued Vidovič Muha at a debate part on the new dictionary of the Slovenian language, which was held at the Fran Ramovš Institute of Slovenian Language in October 2008 (Perdih 2009: 35). In the same year we were preparing specifications for the collection of corpus texts within the framework of the Communication in Slovene project (Sporazume-vanje v slovenščini − SSJ),[1] with the aim of improving the previous reference corpus of Slovenian, i.e. the FidaPLUS corpus (Arhar Holdt and Gorjanc 2007), and defined the purpose of the new corpus as follows:

> Within the Communication in Slovene project there is a great number of objectives whose implementation will be based on the new corpus, including the pedagogical corpus grammar/.../ and orthography guide /.../. Slovenian lexical database will also be based on the corpus in the sense of data acquired from the corpus and its interpretations, as well as in the sense of dictionary examples. *(Korpus pisnih besedil: specifikacije /.../, December 2008: 12).*

The Gigafida corpus,[2] which was completed in 2012 (Logar Berginc et al. 2012), fully completed the pursued objectives, and with its use in preparation of the Slovenian lexical database[3] we also got the feedback on its lexical potential (Gantar 2009; 2010; 2011). Consequently, in the proposal for

---

1   http://eng.slovenscina.eu/
2   http://eng.slovenscina.eu/korpusi/gigafida
3   http://eng.slovenscina.eu/spletni-slovar

making a new explanatory monolingual dictionary of modern Slovene (Krek et al. 2013b), and as a starting point for the preparation of the headword list for the dictionary, it is stated that a "frequency list of the Gigafida corpus in combination with precise and relatively complex statistical analysis of the data from the corpus Kres, Gos and other databases" would be completed (ibid.: 24). The material for the new dictionary, as defined in Gliha Komac et al. (2015: 4), was very similar: "Linguistic data for making a headword list and editing of central parts of dictionary entries /.../ will come from corpus sources, mainly Gigafida, Kres, Nova beseda and partly Gos." We can therefore say once again (as in Logar et al. 2015) that the key Slovenian lexicographers in 2015 were united on the role of Gigafida and Kres in the Slovenian dictionary project, since both corpora adequately represent the lexical identity of written published Slovenian in the last 20 years (i.e. also Logar, 2014: 10 and others), although both also need to be upgraded.

The upgrading of Gigafida and Kres[4] is in the first place necessary because the last texts which were included in both were acquired on 29 May 2010, although some rather narrowly focused texts from the Internet were also obtained from the period from April 2010 to April 2011 (Logar Berginc et al. 2012: 43). Therefore, during the preparation of this paper, it should be noted that texts from books, magazines and newspapers produced less than five years ago did not exist in the Gigafida corpus. The second, perhaps more important reason for the update lies in a very modified and extended possibility of accessing the public word that changed public representation of the Slovenian language, transformed many genres that hitherto were bound only to the print, and with its associated editing processes, and brought new, specific kinds of written texts, namely the rise of new media online. And as we already wrote in Logar and Ljubešić (2013: 104):

> In defence of the necessity of building corpora – then namely corpora of *spoken* texts – Stabej and Vitez (2000) wrote: 'the fact is that the analytical picture of a certain language, which only covers the elements of written texts, is highly partial and incomplete' (79). And further on: 'if the ideal objective of a corpus-based linguistics is language comprehension, as attested in all dimensions of communication, only written corpus is insufficient' (80). The citation can be applied or it is necessarily to apply it to the texts, which a decade later are written for the 'new media'. To omit them in advance from the corpora, which represent the bases for linguistic description of a language in any dimension of communication would mean a disqualification of an important part of the language.

Krek (11. 11. 2013), during the concluding conference on the SSJ project, pointed out that during the preparation of the specifications for the Gigafida

---

4   Where it makes sense in continuation we refer to both.

corpus we were naturally not aware of the large increase in the use of social networks and Internet connected mobile devices that would occur after 2008, while at the same time that the reading of printed newspapers would decline. In the light of this new social reality, which has a strong influence on the language and its related descriptions, resources and technology, it is therefore necessary to rebuild reference corpora starting from good domestic and foreign practices, and plan adjustments where analysis of the corpus exposes its weaknesses.

In the following sections of the chapter we will therefore consider which segments of the Gigafida corpus should be upgraded as a priority to make it even more appropriate and relevant as a collection of linguistic data for the new explanatory monolingual dictionary of modern Slovene. Discussions on issues that require more extensive reflection (above all Internet texts) are presented in subsequent chapters of the book.

2   MODERN SLOVENE

2.1   Beginning of text collection: 1990

Language contemporaneity is a relative concept, and if we want to define the temporal dimension of texts covered by the corpus this concept necessarily requires some agreement. Consensus on the determination of the "contemporaneity" of the corpus depends on both extra- and intra-linguistic factors. Relevant for determining the starting and the finishing year of corpus texts are primarily any major changes to these. In practice, the most common reasons given for selecting the initial year of text collection (mostly rounded on a decade) are as follows:

a) time when the predecessor dictionary was published,

b) any significant socio-political changes in the language community, which brought about major changes in lexis, and

c) practical reasons, e.g. existence of electronic archives, success of the text collection process, and so on.

If we take a look at the state of modern corpora and general dictionaries of Czech and Slovak, which after 1989, due to social, political and economic events, changed or expanded their lexical funds (and even the statuses), similar to Slovenian,[5] we realise the following:

---

5   For example, see also a publication on Latvian by Zuicena and Miglia (2014).

a) The authors of a balanced reference corpus of Czech, prepared by the Institute of the Czech National Corpus of the Faculty of Arts in Prague, wrote in the first version of the corpus, which was made in 2000 (SYN2000,[6] followed by SYN2005 and SYN2010): "The SYN2000 is a synchronic corpus, which means that it covers contemporary Czech. Therefore it contains primarily texts that were created in 1990–1999", and the year 1990 was chosen for journalism and professional texts as a natural landmark of synchrony. The same was also true for the core part of the fiction corpus, with the exception of including books dating back further, ones that were still being reprinted and therefore affect contemporary Czech (whose author was born after 1880; for example K. Čapek and J. Hašek).[7] To this date, the most contemporary dictionary of Czech *Slovník spisovného jazyk českého* (B. Havránek et al.) is much older – and was published in four volumes in the years 1960–71, while the Institute for Czech of the Czech Academy of Sciences published it online in 2011.[8] The Institute for the Czech language is preparing a new dictionary entitled *Academic Dictionary of Contemporary Czech (Akademický slovník současné češtiny)*, but there are few publications discussing this, and these do not reveal its corpus-based methodology.[9]

b) The Ludovit Stur Institute of Linguistics of Slovak Academy of Sciences is also preparing a new dictionary, called the *Dictionary of Contemporary Slovak Language* (*Slovník súčasného slovenského jazyka*). Two volumes have already been published: the first in 2006 (A–G), the second in 2011 (H–L). It is designed as a large-scale dictionary with approximately 220,000 headwords, but its predecessor, i.e. *Dictionary of Slovak Language (Slovník slovenského jazyka)* was published four decades earlier, in the years 1959–1968 (Buzássyová 2009: 119). The primary material for the new dictionary is a lexicographical record with five million tickets and the *Slovak National Corpus*,[10] being edited since 2002 (ibid.: 124), containing texts from 1955 onwards (Šimková and Garabík 2014). In 2009 Buzássyová, who was the main editor of the dictionary (Perdih: 52), said the following:

> In theory /Slovak/ as a contemporary language is understood as from the 1940s, when Czechoslovakia split for the first time. Slovakia and its language then first took over all functions, such as the language of the arts, literature, spoken language, administration language, language for special purposes, but we do not originate from the 1940s, because that would not be realistic. /.../ We originate from the Second World War, which up to the 1960s was also covered by the previous dictionary.

---

6   https://ucnk.ff.cuni.cz/english/syn2000.php
7   http://wiki.korpus.cz/doku.php/cnk:syn2000
8   http://ssjc.ujc.cas.cz/
9   http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html
10   http://korpus.juls.savba.sk/

The decisions made by the Czech and Slovak corpus linguists and lexicographers, and the reasons given for them, confirm a very similar argument from ten years ago on the contemporaneity of texts in the first Slovenian reference corpus, FIDA, upgraded to FidaPLUS and then to Gigafida (Gorjanc 2005: 47–48):

> The corpus FIDA tries to provide comprehensive information on modern Slovene. It tries to cover the image of today's Slovene as comprehensively as it can /.../. FIDA corpus is a synchronic corpus; it includes texts published after the year 1990 /.../. The original idea about including texts after 1980 was changed at the very beginning of the construction of the corpus changed, because of two key reasons. The first one, purely pragmatic, is related to querying the available texts in electronic form; it has been shown that the culture of electronic archives began in the second half of the nineties, so various texts should be digitized before incorporating them to the corpus. The second is related to the indexed database of the Fran Ramovš Institute of Slovenian Language that somehow provides at least basic information on the status of the language from the eighties of the last century.

And in Logar Berginc et al. (2012: 127):

> In the process of defining the time of the text collection, the collectors / of the FIDA corpus / felt that the change of the political system in Slovenia affected the language to the point that this year can be taken as a starting point for the concept of 'synchronicity' of the corpus. /.../ The corpus therefore covered the ten-year period from 1991 to 2000, with some texts from the years 1989/90.

To summarise: we put the start of text collection for the Gigafida corpus and its future upgrade for the needs of the dictionary in 1990, for the following reasons: (a) the date of publishing the last volume of the *Dictionary of Slovene Literary Language* (1970–91; DSLL); (b) socio-political changes in the late 1980s, and especially after Slovenia gained independence in 1991, that have fundamentally affected the lexical image of today's Slovene, and (c) practical reasons, i.e. the existence of the electronic archives of publishers and others.

2.2   Texts after 2010 and in the first half of the 1990s

The time period covered by the texts in the Gigafida corpus started in 1990 and finished in 2010 (print) or 2011 (Internet). The number of words per year is shown in Figure 1.
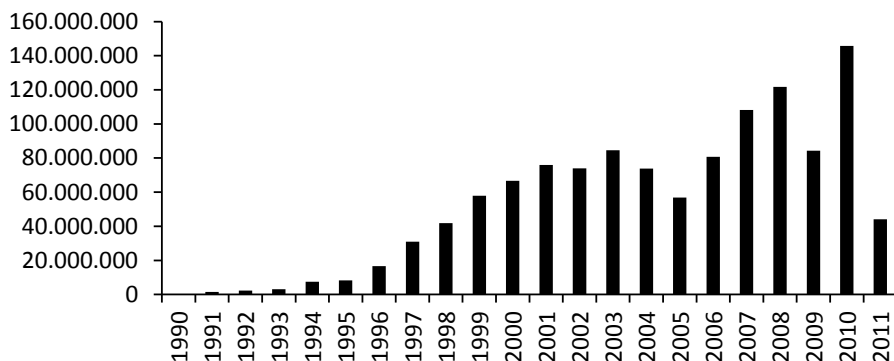
Figure 1: Number of words per year in the Gigafida corpus.
Source: Logar Berginc et al. (2012: 36).

Experience shows that texts from the media are principally acquired for a year or a few years back, and less so for the current year of collection, and thus the decreases in 2005 (the year of text collection for the FidaPLUS corpus) and in 2009 are as expected. These two years can be completed, if the period before the next text collection process is not too long. An upgrade with online texts can be carried out in real-time and throughout the duration of the project, and is then stopped. In this context, crawling for the period 2012–15, in order to build up a reference corpus, remains a key step that cannot be completely replaced by another process.[11] This fact, as well as the gaps in the range of printed texts that are a consequence of excessively long periods of non-updating corpora, certainly speak in favour of longer-term infrastructure solutions, such as the Web Archive of the National and University Library[12] or the long-term financing of infrastructure projects within the Centre for the Language Resources and Technologies at the University of Ljubljana.

At the same time there are very few texts in the corpus that would enable more detailed insights into the lexical collection of Slovenian in the first half of the 1990s. For the seven-year period of 1990–96 Gigafida contains, at first glance, an extensive 22 million words, but this actually represents less than 2% of the entire corpus. If the next project of upgrading Gigafida has the budget and time needed to allow the digitisation of selected texts from this period, then this would definitely be worth considering.

---

11 Theoretically we could make use of the online corpus of Slovene slWaC2 (Erjavec and Ljubešić 2014), but the collection of online texts for this was not guided or controlled to the extent that is desirable with Gigafida (more on this in the next chapter).

12 http://arhiv.nuk.uni-lj.si/

3 SLOVENE IN GENERAL WRITTEN USE

3.1 Appropriateness of the corpus for general dictionary needs and purposes

We have reported several times on the text collection process for the corpora in the "FIDA series" (Gorjanc 2005: 47–53; Arhar Holdt and Gorjanc 2007; Logar Berginc and Šuster 2009; Berginc Logar et al. 2012: 21–25). Generally speaking, the key points are as follows:

a) Purpose: corpora FIDA, FidaPLUS and Gigafida were constructed in order to show a comprehensive picture of the Slovenian language, as seen in public written texts. In this sense, Gigafida as the latest corpus in series is designed to meet various linguistic research aims, but the main focus (as usually observed for the general reference corpora) is its applicability to lexical and lexicographical purposes.

b) The criteria for the collection of texts, content and documents: Gigafida as well its predecessors FIDA and FidaPLUS, used clearly drawn criteria for text collection, details of which are presented in the references, along with other, related decisions.

c) "Chasing" the general use: The criteria for text collection from the corpus FIDA onwards resulted from both reception and production. In relation with the first – if possible – this was carried out through a wider influence sieve. By doing so, we took into account objective data on readership: the National Readership Survey (newspapers, magazines); library borrowing, book awards, circulation, popularity of websites, etc. We did not take into account the collection of specialised texts (scientific) in the third stage of collection, so there are just a few of these in Gigafida. It is difficult to estimate to what extent Gigafida actually shows the general written use of the language, but the collectors never lost sight of their main goal, which was to represent this kind of use as well as possible.

A total of 77% of the words in Gigafida come from texts published in print periodicals. As we were aware that this was likely to be the case, in the SSJ project we also took samples for Kres to obtain a more balanced taxonomic share between different types of texts (Erjavec and Logar Berginc 2012).

The Gigafida corpus is therefore a large corpus and one that is heterogeneous with regard to time, genres, authors, subjects, etc. Krek and Kosem (21. 9. 2013) wrote about this as follows: "As soon as more speakers actually read certain texts (irrespective of their 'weak style'), the greater influence these texts have on their language. And so it becomes more important that lexicographers equip the content

of the dictionary database with relevant information processed from these texts for different types of dictionary users." Based on this, it appears reasonable to continue following the principle of mainly gathering texts with greater communicational influence and with a lesser (or even none existing) role for highly specialised scientific texts, when upgrading the Gigafida and Kres corpora.

3.2   The issue of a "metacorpus"

Both introductory quotations from the two proposals for the future Slovenian dictionary (Krek et al. 2013b; Gliha Komac et al. 2015) with regard to the source for the glossary and the editing of lexical entries, mention using the Gigafida corpus in combinations with Kres, Gos (a corpus of spoken Slovene), Nova beseda and other Slovenian databases. In the last decade quite an extensive selection of different corpora of Slovene has emerged (see e.g. Erjavec 2013),[13] so the question of integrating these for the purposes of dictionary editing has also naturally arisen (see also Gorjanc in Perdih 2009: 47). Or, as we wrote in Logar et al. (2015): "For the future dictionary work /.../ it is not only important the question of which corpora will be used as data collections for editing dictionary entries and why, but also the question of which corpora *will not be* used and why."

Here we speak in favour of the choice that the corpus which will be the main dataset for the general dictionary must be already made with this intention, must be carefully documented and clear in its content and structure. Only in this way will the corpus as a sample allow generalisations, which will then be published as a general-language description and regulation. With regard to the main dictionary source (in our case Gigafida together with its derivative Kres), there are of course possible combinations with other corpus resources and databases (such is, for example, the lexicographical practice in the current format of the *Great Dictionary of the Polish Language*, see Żmigrodzki 2014: 2), but we must stress that this can only happen in a way that is explained to the users of the dictionary and explicitly prescribed in the editorial process.

4   COPYRIGHT AND OPEN ACCESS

Corpora FIDA, FidaPLUS and Gigafida had legal agreements with text providers arranged in a way that it was possible to publish the corpora publicly and with

---

13 http://nl.ijs.si

free access. The key point here is the contractual transfer of material copyrights of the text in a way defined in Article 22 of the Slovenian Law on Copyright and Related Rights (ZASP 2007). Since the case here was accessing the texts in digital form, the holder of the rights also transmitted the rights of electronic reproduction to the providers, as set out in the first paragraph of Article 23 of the ZASP and modification rights, as set out in Article 33 ZASP:

> Article 23:
> (1) The reproduction right is the exclusive right to store the work on a material medium or another medium, directly or indirectly, temporarily or permanently, partly or in whole and in any kind of way or in any kind of form.

> Article 33:
> (1) The right of modification is the exclusive right that allows that a certain original work can be translated, changed for theatrical performances, musically arranged, or be modified on other ways.

> (2) The right from the previous paragraph also applies to cases where the original work is not unchanged but incorporated or integrated into a new work.

> (3) The author of the original work retains the exclusive right to use his or her work in any modified form, unless this law or contract determines otherwise.

The contract between text providers and those preparing the Gigafida corpus contained an article according to which we were allowed to use up to 10% of the text in a manner as determined by the Creative Commons licence: recognition of authorship + non-commercial + share alike, known under the denotation CC BY-NC-SA.[14] This article has enabled the composition of the corpora ccGigafida (volume of 100 million words) and ccKres (10 million words) which are accessible in the form of a database.[15]

Open access to research data from publicly funded projects was supported by all the members of OECD by signing the *Declaration on Access to Research Data from Public Funding* (OECD 2004), and Slovenia signed this in 2010 (see also *OECD Principles and Guidelines for Access to Research Data from Public Funding*).[16] The initiative with strategic documents, reports and commitments was also supported by the European Commission, the European Scientific Council, the European Federation of Academies of Sciences ALLEA and other bodies. In this respect the European Commission's recommendation on

---

14 https://creativecommons.org/licenses/by-nc-sa/2.5/si/legalcode
15 http://hdl.handle.net/11356/1035 in http://hdl.handle.net/11356/1034
16 http://www.oecd.org/sti/sci-tech/38500813.pdf

accessing the scientific information and their archives from 2012 is important.[17] The latter reminds EU Member States about access to publications that are the result of publicly funded research – this must be open as soon as possible, preferably immediately, and in any case not later than six months after the date of publication for the social sciences and twelve months for humanistic sciences (L194/41).[18] In the final report of the project named *Open Data – Action Plan for the Establishment of a System of Open Access to Publicly Funded Research Data in Slovenia* (2010–2013), the researchers pointed out that open research information is

> a shared responsibility of all the participants in science, which cannot be left to only one segment, for example ethical principles, but requires clearly defined obligations for individual researchers, their institutions and administrations, professional and scientific associations and other representatives of scientific community, providers of data-related services and publishers (Štebe et al. 2013: XVI).

In the future making of a reference corpus of Slovene we will have to commit to this responsibility and prepare the corpus not only for its use in a concordancer, but also in the form of "CC", which will enable domestic and foreign researchers to develop high quality, robust and useful tools for processing of natural language, in our case Slovene (Erjavec 2009: 115; Erjavec 2014). The necessity of such tools for Slovene has been pointed out on several occasions (e.g. Krek 2012b).

5 RELATED CORPORA IN TODAY'S FOREIGN LEXICOGRAPHIC PRACTICE

Table 1 shows a list of currently formatting or recently formatted general dictionaries of Finnish, Estonian, Latvian, Polish, Czech, Slovak, Dutch, German and English with the structure of the corpus, which is (was) the basis for the dictionary (if such a corpus exists).[19]

---

17 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF

18 For more about open access see http://www.openaccess.si/

19 If for each language several general dictionaries are currently being compiled, we chose the one that is designed for web publishing; if there were several of these, as for English, the selection was random.

Table 1: List of dictionaries of nine foreign languages with the volume and contents of corpora from which they were formed or are still forming. Source: Completed and updated according to Logar (2014).

Language, dictionary, corpus	Corpus volume	Corpus contents
FINISH New dictionary of contemporary Finish / Kielitoimiston sanakirja	The dictionary is not corpus-based (Heinonen 2014).	/
ESTONIAN The Basic Estonian Dictionary (online edition in the making; Kallas et al. 2014) The Balanced Corpus of Estonian http://www.cl.ut.ee/korpused/ grammatikakorpus/	15 million	• newspapers and magazines: 33% • fiction: 33% • science texts: 33%
LATVIAN Dictionary of Contemporary Latvian / Mūsdienu latviešu valodas vārdnīca www.tezaurs.lv/mlvv The Balanced Corpus of Contemporary Latvian / Līdzsvarots mūsdienu latviešu valodas tekstu korpuss www.korpuss.lv	4.5 million	• newspapers and magazines: 55% • fiction: 20% • science texts: 10% • legal texts: 8% • other: 5% • written records of parliamentary meetings: 2%
POLISH Large Dictionary of Polish Language/ Wielky słownik języka polskiego http://www.wsjp.pl/ Nacional Corpus of Polish Language / Narodowy korpus języka polskiego http://nkjp.pl/	(in the planning stage) 1.5 billion (Górski in Łazinski 2012: 33)	• newspapers, magazines and press releases: 50% • fiction: 16% • spoken texts: 10% • non-fiction: 11% • web texts: 7% • didactic texts: 2% • other: 3% • nonaligned: 1%
CZECH Akademic Dictionary of Contemporary Czech / Akademický slovník současné češtiny http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/ akademicky-slovnik-soucasne-cestiny. html	Information about corpus-based design is not mentioned or clear.	/

Language, dictionary, corpus	Corpus volume	Corpus contents
SLOVAK Dictionary of Contemporary Slovak Language / Slovník súčasného slovenského jazyka http://slovniky.juls.savba.sk/ Slovak national corpus / Slovenský národný korpus (2013) http://korpus.juls.savba.sk/stats.html	829 million	• newspapers and magazines: 69% • non-fiction: 15% • fiction: 14% • other: 2%
DUTCH General Dutch Dictionary/ Algemeen Nederlands Woordenboek http://anw.inl.nl/search ANW Corpus / Algemeen Nederlands Woordenboek (ANW) http://anw.inl.nl/show?page=help_anwcorpus	102.5 million	• newspapers: 40% • web texts: 30% • fiction: 20% • newspapers, magazines and news portals – neologism: 5% • older texts, 1970–2000: 5%
GERMAN a) Project OWID of the Institute for German Language in Mannheim, http://www1.ids-mannheim.de/lexik/owid.html) Elexiko http://www.owid.de/wb/elexiko/start.html Elexiko-Corpus http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html b) DWDS: A Digital Dictionary of German Language / Das Digitale Wörterbuch der Deutschen Sprache (http://www.dwds.de/) Kernkorpus[21] (http://www.dwds.de/ressourcen/kernkorpus/)	2.7 billion	• newspapers and magazines: 100% • fiction: 26% • non-fiction: 22% • scientific texts: 25% • newspapers and magazines: 27%
ENGLISH Oxford Dictionaries http://www.oed.com/ Oxford English Corpus http://www.oxforddictionaries.com/words/the-oxford-english-corpus	2.5 billion	• web texts: almost 100% (novels, non-specialised and specialised magazines, newspapers, blogs, e-mail, social networks, etc.)

20 The dictionary is based on 15 corpora, with Kernkorpus as the most important one, due to its balanced and reference structure.

The table shows that the corpora that are datasets for present and comparatively interesting dictionaries of seven foreign languages (if we overlook the Finnish and Czech) are, according to their structures, very different. If we limit ourselves to only three key categories that were most criticized in the Gigafida corpus, i.e. the small volume of fiction, large volume of journalistic texts and seemingly non-normative web text, we obtain the data in Table 2 and Picture 2 (we omit the English corpus, for which the text type composition is not publicly available, but we add data for the Czech corpus SYN2010).

Table 2: The contents of corpora of seven foreign languages and Gigafida and Kres (in %) in the categories of fiction, newspapers and magazines and web texts. Source: Completed and updated according to Logar (2014).

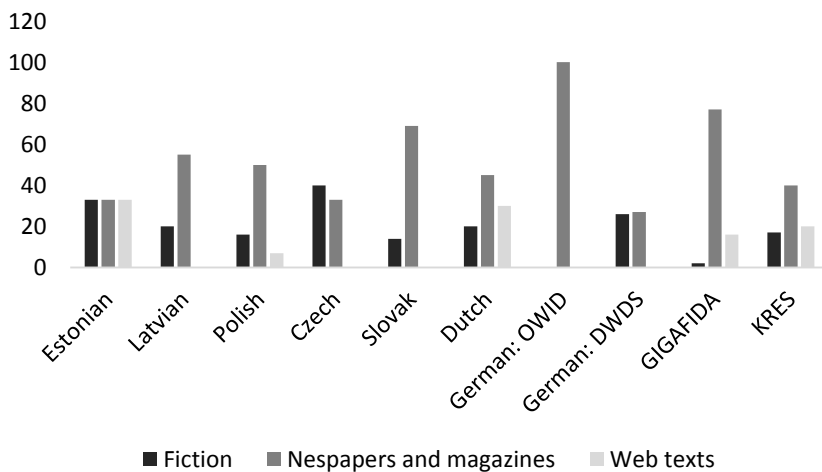|	Fiction	Newspapers and magazines	Web texts
Estonian	33	33	33
Latvian	20	55	0
Polish	16	50	7
Czech	40	33	0
Slovak	14	69	0
Dutch	20	45	30
German: OWID	0	100	0
German: DWDS	26	27	0
GIGAFIDA	2	77	16
KRES	17	40	20

Figure 2: Contents of corpora of seven foreign languages and Gigafida and Kres (in %) in the categories of fiction, newspapers and magazines and web texts.

In Table 2 and Figure 2 we can see the following: on average more texts in the corpora come from newspapers and magazines; Gigafida has relatively little fiction, but has the largest share of journalistic texts, although the German corpus surpasses it here and the Slovak corpus is also close. Gigafida is approximately in the middle with regard to web texts. In relation to the other corpora, the components of Kres are rather average.

6 TARGETED COLLECTION OF TEXTS FOR THE PURPOSE OF THE DICTIONARY

6.1 Specialised lexis

Ledinek (2014b: 2) summarised the key issues related to the inclusion of terminology in general dictionaries as follows:

> Questions like, what is the terminology in the concrete monolingual dictionary of middle range, what will be its presumed part in the dictionary, which fields of expertise will be (in greater extent and systematically) included and what will be the way of terminology qualification (baseline) of terminology lexicon, are fundamental questions of a dictionary concept.

There is no doubt about whether to include a terminological lexicon with approximately 100,000 entries in the general dictionary or not, the question is what professional lexis and their typical text environments should be included, and in what way. The exact percentage of specialised lexis to be included in the general dictionary is debatable, but one thing is clear: to make possible any kind of collection and selection, the corpus which will form the basis for the dictionary has to be prepared in a way that it will demonstrate the state of terminological – i.e. de-terminological – lexis that is part of general language. If we leave aside the fact that such a lexicon is already reflected in the newspaper and magazine part of the corpus, as well as in the news portals part, it makes sense to follow two principles to achieve this objective when updating the Gigafida corpus:

a) the principle of *non-inclusion* of specialised texts (scientific magazines and monographs, doctoral dissertations, articles from scientific conferences, etc. precisely those that are most interesting for LSP corpora; cf. Logar, 2013: 47–52), and at the same time

b) the principle of *integration* of the popular professional works and textbooks to the level of secondary school.

We already wrote that in the final collection we avoided scientific texts, while great attention throughout the collection period after 1997 was focused on

obtaining popular professional books (manuals, guides, etc.) from various fields of human life, as well as magazines that present scientific knowledge to laymen (often younger readers). Gigafida contains almost 900 manuals from 84 different publishers and among the magazines at least 50 of them focus on some kind of expertise (e.g. motoring: *Avto Foto Market*, *Avto Magazin*, *Avtokatalog*, *Motorevija*, *Motokatalog* and *Mobil*; computing: *Connect*, *Joker*, *Moj mikro*, *Monitor*, *PC & mediji* and *Računalniške novice*). In this context it is possible to follow previous good practices and experience. The situation is different with textbooks, catalogues and didactical books, where new collections should be more systematic. Gigafida contains 103 such works, which were released by five publishers: National Education Institute Slovenia, National Examinations Centre, Rokus Klett, DZS and Ataja, but a review of included textbooks (and workbooks) shows that the scope of obligatory elementary education is covered irregularly:

- Mathematics (6 textbooks or workbooks)
- Slovene (13)
- English (1)
- History (8)
- Biology (7)
- Environmental Sciences (2)
- Physics (1)
- Chemistry (4)
- Society (4)
- Natural Sciences (1)
- Natural Sciences and Technology (1)
- Arts (1)
- Musical art (8)
- Sports (1)
- Home economics (3)

At first glance it is therefore clear that in Gigafida the obligatory school programme is not properly covered by the textbooks it includes, and there are even less works for the programs of secondary and high schools. According to the syllabus for elementary schools produced by the Ministry of Education, Science and Sport,[21] there are still missing textbooks for geography, state and civic culture, engineering and technology. From this perspective it is necessary

---

21 http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/devetletka/predmetniki/Pred_14_OS_4_12.pdf

to complete the corpus, preferably with a tendency to capture textbooks, workbooks and texts related to pupils and students of all school subjects that are part of general and vocational programs (at elementary schools, gymnasiums, and vocational secondary schools). Moreover, it would be useful to obtain information on textbooks and similar materials used for after-school extracurricular activities, particularly those with large-scale participation, and try to include this material. In this way, an upgraded Gigafida – assuming the cooperation of the text providers – would appropriately cover the terminology that almost everybody encounters during the education process. From such a corpus a collection of terms with a more comprehensive range would be extracted, and this could then be applied to the dictionary concept using a coordinated lexicographically-terminographic process.

6.2 Topic coverage

The collection of texts for reference corpora is directed by several criteria, including the diversity of text topic. In the collection of texts for the Gigafida corpus, we worked from the following list (Logar Berginc et al. 2012: 15):

- current events
- economy, politics
- education
- nature, home, pets
- people, family, men, women, children, youth
- health, food
- business, finance
- leisure, music, movie, entertainment, fashion
- sport, tourism
- culture, art
- religion, spirituality
- computing, motoring, etc.

When we used the topic modelling method to compare Gigafida with the first version of the web corpus of Slovenian slWaC (Logar Berginc and Ljubešić 2013), we found out that of twenty topics the two corpora have eight in common, seven partly in common and five different (ibid.: 92):

Characteristic to the Gigafida corpus are topics of settlements and road traffic (particularly in terms of traffic accidents), events (especially in terms of their announcement, description), television and radio programmes, individual sports and employment. In the slWaC corpus standing out are movies, music, travel and tourism, foreign policy (especially EU, Croatia), and classified ads.

From Tables 1 and 2, presented in the next chapter, we can summarise similarities and differences between the topics in the Gigafida corpus and the latest version of the slWaC$_2$ corpus, formed in 2014 (Erjavec and Ljubešić 2014):

a) Thirteen topics are common to both: human, men, woman, family life; society, other; sports; internal policy; education; finance; local politics; law; publications, culture, art; motoring; health; ICT and food.

b) Three topics are partially shared: economy (Gigafida) – economy, development (slWaC$_2$); events in the local area (Gigafida) – events (film, music, theatre) (slWaC$_2$); animals, nature, living environment (Gigafida) – living environment (slWaC$_2$).

c) Four topics are different:

- Gigafida: war, terrorism, crime; TV and radio programmes; traffic; media;

- SlWaC2: travel, tourism; online shopping; religion and internet.

Topic weaknesses of the Gigafida corpus, as indicated by this analysis, are its lack of texts about film, music and related events, travel and tourism, classified ads, external politics related to the EU, online shopping and world in general, and – surprisingly – religion. With the exception of the last one we can conclude that these are topics that in recent years have appeared quite frequently in the online media, which speaks in favour of the integration of web texts (with these topics) in the reference corpus. The analysis also confirmed the over-representation of TV and radio programmes in the Gigafida corpus (which will have to be reduced by an extended de-duplication process in the future) and the adequacy of the list of topics, which was prepared prior to the collection, although the process of updating the corpus should add the topics of law, traffic, living environment and Internet.

7   ADDITIONAL TAXONOMIC CATEGORIES

Gigafida's taxonomy is quite simple: the texts are on the first level separated into *printed* and *Internet* (see below), and then printed into *books* and *periodicals*. Literary works are divided into *fiction* and *factual texts*, and periodically printed texts

into *newspapers* and *magazines*. The category *other* is diverse (and provides only 0.67% words in the Gigafida corpus) and contains texts such as records of the meetings of the National Assembly of Republic of Slovenia, subtitles and post-production texts of the Slovenian National Television.

print
      book
            fiction
            factual texts
      periodical
            newspapers
            magazines
      other
Internet

For a general corpus search it seems that such a taxonomy is sufficient, but for lexicographical purposes it would be helpful if this would be complemented and/or further analysed. In this regard we have already indicated the need for a separate category for textbooks and similar texts, and in the next chapter we will think in this way about online texts written in non-standard Slovenian (blogs, forum posts, tweets, and comments on news portals). So far, analysis also shows that additional corpus labelling may help the lexicographers in deciding on:

- annotation of field specific lexis,

- annotation of style specific lexis.

7.1   Corpus metadata and field specific dictionary labels

Labels for field specific words or specific meanings of words, i.e. labels like *agriculture*, *motoring*, and *banking*, are closely linked with the question of terminology included in general dictionaries. If the Gigafida corpus would be at least partially labelled with topic categories, this could warn the lexicographer about a potentially sector specific meaning of the entry that he/she is editing, and at the same time such label in the corpus would allow additional sub-corpus searches. As we have already observed in Logar and Ljubešić (2013: 80), several foreign corpora have thematic categories attributed to the factual texts:

a) In the Czech National Corpus SYN2010[22] factual texts are divided into:

- religion

---

22 http://ucnk.ff.cuni.cz/english/syn2010.php

- law

- art

- economics

- technology

- natural sciences

- humanities and lifestyles

b) In the Croatian National Corpus[23] the layout is:

- scientific texts:
  - ○ life sciences
  - ○ technical science
  - ○ biomedical sciences
  - ○ biotechnical sciences
  - ○ social sciences
  - ○ humanistic science

- professional texts:
  - ○ travel
  - ○ reviews
  - ○ media
  - ○ criminology
  - ○ sports
  - ○ politics
  - ○ ecology, bioethics, etc.

c) In the British National Corpus[24] under the informative texts can be found:

- world politics

- trade and finance

- art

- religion and philosophy

- leisure etc.

---

23 http://hnk.ffzg.hr/struktura.html
24 http://www.natcorp.ox.ac.uk/

Topic division, though not fully implemented, is, for example, also typical of the reference corpus *Oxford English Corpus*,[25] which consists of twenty parts, mostly named according to topics, e.g. computer science, environment, leisure, military, and transport. These parts are further sub-divided into sub-topics or sub-sections (sport, for example, has about 40 of these).

To achieve a complete collection of topic categories, which could be used with the texts of the upgraded Gigafida, several approaches are possible and can also be combined with one another: we could select the typology of one of the foreign corpora or rearrange the collection of topics that guided the collection of texts. A sensible approach here would be to have in sight the results of comparisons between the Gigafida and slWaC corpus obtained with the method of topic modelling and before finalising the topic scheme – to obtain key words for every corpus document with the method of TF-IDF (*Term Frequency – Inverse Document Frequency;* Salton and Buckley 1988). With the resulting topic scheme we would then manually mark the training set of documents, perform machine learning and then automatically label the corpora.

7.2   Corpus metadata and stylistic dictionary labels

The output of stylistic labels in the current version of the lexical database for Slovenian showed that the editors qualified the meanings with the following annotations in five groups (Krek et al. 2013b: 94–96):

a) **time**: *less frequent use, the word is very rarely used in this sense in contemporary Slovene, obsolete*[26]

b) **connotation**: *to express emphasis, figurative meaning, dissenting, it expresses impairment, pejorative, usually with disapproval*

c) **context**: *in journalistic jargon, ad texts, often in classified ads, particularly in sport, in Christianity, in a political context*

d) **pragmatics**: *as a proverb, with disapproval, euphemistically, usually as insult, rough and slightly vulgar*

d) **register**: *in very informal situations, in informal situations, in speech, in an informal school speech, informally*

To determine the *connotation* and *pragmatic labels* lexicographer must evaluate the text environment, where tools such as the Sketch Engine[27] (Kilgarriff et al.

---

25   http://oxforddictionaries.com/words/the-oec-composition-and-structure
26   These are just few examples from the preliminary drafting stage.
27   http://www.sketchengine.co.uk/

2004) can be a great help, while current corpus metadata may help in the time-frequency, contextual and register labels.

A. Time and frequency

Oldness or *obsolescence* of the vocabulary cannot be seen directly from corpus metadata (year of publication) since only texts issued after 1990 (mainly after 1996) are included in Gigafida. This means that the time labels can be provided by a lexicographer only on the basis of a review of the direct textual environment of the word in combination with an analysis of the frequency relationship between synonyms. On the other hand, Gigafida, with texts from a 20-year period, is relevant enough to allow reasonable annotation of labels such as *increasing use*, *decreasing use* and so on.[28] Here we must also be attentive to the frequency trend, and the fact that we should combine the increase or decrease in the frequency in a specific time period with the dispersion of sources, relative frequency depending on the number of words per year and frequency of possible synonyms. The tendency towards the transition from the labelling of timing to the labelling of frequency is in fact already seen in the preliminary set of labels in the current lexical database (e.g. *less frequent use*, *the word is rarely used*).

B. Context

Current contextual labels are diverse. They are partly linked to the analysis of a contex, which already existing corpus metadata also helps with, although to a lesser degree (e.g. lexical units from the records of the meetings of the National Assembly), and additional labelling of the corpus based on this would not help. Contextual labels are partly associated with the topic (see above, and particularly in *sport*, *Christianity*, and *political* contexts), about which we already wrote in Section 6.1.

C. Register

Register labels, the same as contextual ones, derive partly from the analysis of the context. It appears that this is primarily about identification of informal speaking situations, which can occur in all types of text, e.g. in *fiction*, in the dialogues of people in *magazines* and *newspapers*, in citations, interviews, half-literary genres or literary feuilletons. Two types of text in the Gigafida corpus were primarily spoken (records of meetings of the National Assembly and television subtitles), and both are labelled as *other* and named in the taxonomy, which directly helps a lexicographer with determining register. The third interesting source for register labels, which is also named, is the *Internet*, particularly texts

---

28 A chart would be most obvious in this respect.

that are to be found on news portals, and, more precisely, the texts of comments under news stories. The news sites included in the current Gigafida corpus are 24ur.com, rtvslo.si, siol.net, arhivo.com, govori.se, najdi.si (news), n-tv.si, pozareport.si, primorske.si and revija-reporter.si. The first three portals are mentioned by name, the rest have a common naming, *Internet – news*. When upgrading Gigafida with Internet texts (see next chapter) it would also be helpful to assign a separate taxonomic category to text comments as well.

8   CONCLUSION

Thirty-two researchers from eight institutions of scientific research and one publishing house cooperated in the building of the Gigafida corpus (Logar 2014: 4). The "FIDA series" corpora, which emerged over a period of almost two decades, are examples of good practice, which have followed the standards of European corpus linguistics. Therefore, when preparing the new reference corpus of Slovenian it would be good to start where we left off with Gigafida, taking into consideration the amendments which were brought into the language and text production by a new digital social reality, and the proposed improvements that were raised by the assessments of the final version of Gigafida and Kres. In this paper we did not define the structure of the future corpus of modern Slovene language. Likewise, we did not propose lists of texts that are missing in specific topics, and did not determine web sites on which it would be reasonable to perform crawling, or prepare a new taxonomy. A more concrete document must thus respond to these and related issues, such as the specification of methods used to collect texts, which is possible and sensible to prepare only when the project is approved and its time and financial frameworks are known.

The relevance of linguistic data with regard to the dictionary concept is fundamental, as we wrote in the introduction. Neither of the two existing conceptual proposals for the new dictionary of Slovenian has yet been finalised. One proposes a product "in the sense of a basic and comprehensive lexical handbook for Slovenian in the digital age", that will respectively be "conceptually, as well as from the database point of view designed completely from scratch" (Krek et al. 2013b: 20), the other will "continue the tradition of the *Dictionary of the Slovenian Language* in the sense of modern linguistic theory and in the sense of description of language use" (Gliha Komac et al. 2015: 1). Gigafida suffices to enable this baseline, but in accordance with the findings shown here and in the following chapter it can be – and should be – extended. Subsequent adjustments

will be then determined by the final dictionary concept. It will then depend on the transparency and consistency of the lexicographical process how the resulting data will be interpreted, and to what extent it will be taken into consideration, exploited or ignored.

The expansion of the Gigafida corpus: Internet content

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar and Vesna Mikolič

Abstract

The paper discusses the expansion of the Gigafida corpus, a Slovenian reference corpus, to include Internet content, i.e. web pages and user-generated content (tweets, blogs, forums and comments on news portals). The resources and tools available which are best suited to achieve this objective are discussed, and the web crawling methodology used for this purpose is also presented.

Keywords: reference corpus, Slovenian, dictionary, Internet content, web crawling

1   INTRODUCTION

In Logar Berginc et al. (2012: 45) we opened the chapter entitled "Web Text in the Gigafida Corpus" with the finding that the written language is becoming less commonly used in the form of the printed word, and more commonly seen in electronic media. The chapter presented data showing that as of October 2007, 66% of the respondents, aged between 12 to 65 years, were using the Internet (RIS survey).[1] The most recent percentages are – as expected – even higher: according to an analysis by the Statistical Office of the Republic of Slovenia, in the first quarter of 2014, 97% of Slovenian households with children and 70% of households without children had Internet access, and during this time 72% of all people aged 16 to 74[2] years old were using the Internet. It may be added that

> 81% of these persons /.../ were using the Internet every day or almost every day. The largest percentage (87%) used it for sending or receiving e-mails and finding information about goods or services. / 58% of respondents in the first quarter of 2014 participated in online social networks (in the first quarter of 2013 the figure was 53%) (ibid.).

Another important finding was that 66% of the users accessed the Internet via mobile phones or other mobile devices (e.g. a tablet). The Internet is thus accessible anywhere, and not just for reading, watching and listening, but also for writing and publishing texts, images, music, and so on. Widely available public platforms that rely on language – once limited to print, radio and television – are now open to contributions from virtually everyone, and this has brought a new kind of Slovenian into public use: texts showing linguistic characteristics that were previously primarily used for speech in private and informal situations.

Editors of modern reference corpora of different languages include web texts into their work in various different ways. The overview presented by Logar Berginc and Ljubešić (2013) noted "a common tendency for including texts from the Internet in the reference corpus, although to what extent this may happen in the future is not yet clearly defined, but if the corpus already contains or will contain texts from the Internet, texts of different genres should also be included" (ibid: 103). Consequently, on the one hand we have for example the *Oxford English Corpus*, from which *Oxford Dictionaries* arise,[3] that is almost entirely composed of texts from the Internet, and on the other hand, for example *The Slovak National Corpus*, on the basis of which *Dictionary of Contemporary*

---

1   http://www.ris.org/
2   http://www.stat.si/StatWeb/glavnanavigacija/podatki/prikazistaronovico?IdNovice=6560
3   http://www.oxforddictionaries.com/

Slovak Language,[4] is currently being prepared, that does not contain any online texts (see more in Table 1 in the chapter Reference Corpora Revisited: Expansion of the Gigafida Corpus).

As will be seen later in this chapter, we see texts from web pages, comments on news sites, blogs, tweets and forum messages as a significant part of the public written Slovenian, which is why we argue they should be included in the corpus that will be the basis for the future reference dictionary of our language. As such, lexicographers should be interested in lexicons that are used in different circumstances by all the speakers of Slovenian, not just journalists, translators, writers, and so on. We should therefore pay special attention to (semi)public written online communication that is determined by circumstances such as (non)interactivity (a)synchronicity, physical (non)presence/absence of the interlocutor and other situational factors, resulting in a highly interactive form of communication with more elements of the spontaneous spoken language, and with (adapted for computer communication) paralinguistic and prosodic elements (Crystal 2001). The task of a corpus as a lexical resource *must* therefore be also to capture this linguistic reality, so in this chapter we illuminate this issue from four angles:

a)   the initial state of the Gigafida corpus (compared with the slWaC$_2$ corpus, the online corpus of Slovenian),

b)   diversity of online text genres and reasons for their inclusion in the corpus (or exclusion from it),

c)   resources and tools that are already available for a future upgrade of the Gigafida corpus (the JANES project),[5] and

d)   the most appropriate methodology of web crawling, including the possibility of building a subcorpus that would be regularly updated.

2   GIGAFIDA AND SLWAC$_2$: EXISTING STATE, COMPARISON, BINDING POSSIBILITIES

Web sites that were included in the Gigafida corpus and technologies for their collection are described in more detail in the already mentioned chapter in Logar Berginc et al. (2012: 45–67), so we shall only note that integrating web content into the Gigafida corpus was "methodologically speaking, the first such major attempt in Slovenia that could formulate guidelines for the future construction of Slovenian reference corpora and indicate some interesting comparative linguistic

---

4   http://slovniky.juls.savba.sk/

5   http://nl.ijs.si/janes/

analysis" (ibid.: 45). Gigafida therefore contains texts from 10 news portals and a total of 91 introductory web pages (29 corporate web pages, and 42 cultural, state, research and university institution web pages). The web was crawled in the period April 2010 – April 2011, and it contributed more than 185 million words to the corpus, of which 63% come from news portals (24ur.com, rtvslo.si, siol.net, etc.), 30% from institutional web pages (gov.si, uni-lj.si, sazu.si, ijs.si, etc.), and 7% from corporate web pages (eles.si, gorenje.si, and kolosej.si, among others). The procedure to capture texts from web pages followed several steps: selection and preparation of the programme for three regimes of crawling (daily, monthly and one-off), boilerplate removal, language detection, and finally the detection and the removal of duplicates and near-duplicates. It turns out that in order to achieve seemingly simple tasks, i.e. that of including web texts in the reference corpus, a fairly complex methodology is required, that – along with the criteria of selecting of web sites and rating of the obtained results – we successfully tested and adapted for use with Slovenian (more on the latest methods of crawling are reported in section 5, below).

During the integration of web texts in the Gigafida corpus – in 2011 – a new and methodologically similar corpus of Slovenian emerged, the corpus slWaC (Erjavec and Ljubešić 2011),[6] which was upgraded to $slWaC_2$ in 2014 (Erjavec and Ljubešić 2014). $slWaC_2$ contains 1.2 billion words from texts acquired from over 37,000 web domains or 2.8 million URLs. The methodology of the construction of the two versions of slWaC is presented in detail in the references mentioned and in Logar Berginc and Ljubešić (2013: 87–89).

The existence of two large corpora of Slovenian has prompted some comparisons that have shown what both of them contain, as well as what their deficiencies are (as much as a comparison of the two entities can reveal in this regard). A comparison based on the frequency profiles (Rayson and Garside 2000) of Gigafida and $slWaC_2$ showed (Erjavec et al. 2015b: 40) that in the latter there are several texts related to computer science, the Internet and the use of web contents, while Gigafida contains more texts that are typical for newspapers, on subjects such as sports, domestic politics, the economy and crime.

To the already published comparative data (ibid., and in Logar Berginc and Ljubešić 2013), we now add data from more recent comparisons between Gigafida and $slWaC_2$, obtained by the topic modelling method (Blei et al. 2003; Sharoff 2010) – though here only paying attention to possible weakness of the Gigafida corpus. Tables 1 and 2 show the 20 most common topics for Gigafida and $slWaC_2$, respectively.

---

6   http://nl.ijs.si/

Table 1: Noun lemmas, which most likely belong to one topic, and the occurrence of the topics in Gigafida.

Subject	Frequency*	Noun lemma
people, family, life in general	4,835	otrok leto dan čas ženska življenje človek družina oče moški roka prijatelj glava žena mama mož sin starš hiša
sport	4,034	tekma mesto leto ekipa zmaga točka igra sezona igralec prvenstvo klub liga prvak trener minuta konec pokal krog reprezentanca
domestic politics	3,639	predsednik vlada država stranka svet minister leto zakon volitev predlog poslanec vprašanje komisija član odbor zbor seja politika ministrstvo
society, OTHER	3,631	človek življenje svet čas odnos način stvar država vprašanje družba primer beseda delo moč stran problem resnica leto občutek
shows, performances etc. in the local area	2,865	ura društvo leto prireditev dan sobota dom član vas mesto občina skupina nedelja šola srečanje gost obiskovalec dvorana delo
war, terrorism, criminal acts	2,669	leto vojna država policija policist človek vojska dejanje orožje dan napad vojak žrtev sodišče oblast zapor kazen čas mesto
TV and radio programmes	2,622	film leto glasba oddaja tv poročilo skupina serija dan pesem festival čas koncert predstava program vloga gledališče del novica
traffic	2,481	cesta pot dan nesreča leto ura voda voznik morje vozilo mesto meter promet letalo kilometer čas vožnja kraj avtomobil
economy	2,470	leto odstotek država podjetje cena trg plača izdelek rast razvoj delo gospodarstvo proizvodnja delavec število področje strošek mesec sistem
education	2,363	šola delo leto otrok program področje znanje študent projekt izobraževanje univerza fakulteta učenec razvoj starš organizacija učitelj center zavod
finances	2,292	milijon evro tolar leto banka družba podjetje odstotek delnica milijarda dolar denar vrednost cena prodaja delež dobiček trg sklad
local politics	2,228	občina leto prostor gradnja objekt cesta projekt območje delo zemljišče mesto milijon stanovanje okolje podjetje načrt denar voda tolar
animals, nature, living spaces	2,153	žival barva prostor vrsta pes voda hiša gozd del material les vrt tla drevo konj čas leto vrata oblika
law	2,145	zakon člen sodišče postopek pravica primer podatek organ odstavek dan oseba podlaga pogodba delo odločba sklad stranka določba zadeva

Subject	Frequency*	Noun lemma
publications, culture, art	2,087	leto knjiga delo razstava stoletje cerkev mesto čas muzej svet ime zbirka umetnost avtor jezik zgodovina del slika beseda
motoring	2,021	m sit avtomobil motor km cena vozilo eur d e l model leto avto x n g r h
health	1,942	bolezen zdravnik bolnik zdravilo telo človek zdravljenje leto koža težava dan zdravje primer rak bolnišnica bolečina kri celica čas
media	1,788	naslov stran številka medij novinar revija dan nagrada pošta časopis leto ime delo informacija oddaja članek televizija bralec vprašanje
information and communication technology	1,491	računalnik sistem uporabnik podatek program slika stran naprava uporaba kartica telefon zaslon internet omrežje model oprema tehnologija možnost storitev
food	1,437	vino voda olje rastlina minuta meso sladkor g sol hrana zelenjava jed žlica okus sadje mleko krompir sok list

* "Frequency" in the second column signifies the occurrence of individual topics in the corpus.

Table 2: Noun lemmas, which most likely belong to one topic, and the occurrence of the topic in slWaC$_2$.

Subject	Frequency	Noun lemma
people, family, life in general	3,929	otrok dan čas leto človek ženska roka pes življenje stvar prijatelj moški glava mama ura družina starš svet konec
society, OTHER	3,266	človek življenje svet čas način odnos stvar družba otrok ljubezen beseda primer vprašanje resnica pot občutek ženska moč problem
domestic politics	2,626	vlada država predsednik stranka zakon svet leto predlog minister član poslanec komisija vprašanje zbor odbor politika skupina pravica mnenje
travelling, tourism	2,524	pot mesto dan cesta ura leto čas vrh morje voda smer gora meter del dolina kraj gozd hotel stran
economy, development	2,360	podjetje področje razvoj sistem projekt delo leto trg storitev okolje država cilj organizacija program izdelek znanje rešitev sodelovanje tehnologija
finances	2,265	leto evro odstotek milijon podjetje banka država cena družba denar trg vrednost rast milijarda sredstvo delnica plača prodaja mesec
sport	2,232	tekma ekipa mesto igra leto točka zmaga sezona igralec minuta prvenstvo klub liga konec tekmovanje prvak rezultat trener pokal
shows (film, music, theatre)	2,139	film leto glasba skupina album pesem festival koncert skladba čas oder nastop predstava nagrada vloga dan zasedba oddaja zgodba

Subject	Frequency	Noun lemma
education	2,072	šola otrok leto delo program študent učenec znanje starš ura izobraževanje fakulteta univerza čas študij delavnica področje učitelj dan
health	2,059	telo bolezen koža zdravilo težava zdravljenje zdravnik dan leto bolnik bolečina človek zdravje celica primer čas kri otrok učinek
online shopping	2,042	stran podatek uporabnik naslov storitev vsebina račun pošta cena ime nakup internet številka informacija izdelek naročilo ponudba dan paket
law	2,016	člen zakon sodišče postopek pravica odstavek oseba pogodba primer dan stranka podlaga sklad organ delo določba odločba podatek pogoj
local politics	1,937	občina leto projekt društvo območje mesto delo prostor sredstvo objekt program član center gradnja zavod organizacija področje okolje ministrstvo
religion	1,887	leto cerkev človek vojna bog življenje dan mesto čas smrt vojska svet država oče maša ime beseda vera stoletje
publications, culture, art	1,826	leto knjiga delo jezik razstava avtor medij beseda fotografija nagrada zbirka revija del umetnost zgodba naslov čas svet dogodek
information and communication technology	1,781	računalnik naprava sistem slika program telefon fotografija podatek uporabnik uporaba video zaslon stran aplikacija dokument model kamera različica oprema
motoring	1,697	vozilo avtomobil motor barva vožnja voznik model kolo avto del leto oblačilo znamka cesta hitrost obleka sedež oprema sistem
living spaces	1,624	voda prostor energija hiša material sistem površina odpadek zrak objekt naprava del uporaba stanovanje temperatura okno okolje les plin
food	1,471	hrana voda olje rastlina vino mleko okus meso zelenjava vrsta jed sadje izdelek oseba količina dan kislina žival sladkor
World Wide Web	0,520	piškotek dan nastavitev seja mesto namen stran storitev uporaba informacija podatek oglaševanje klik gumb primer ura facebook možnost novica

Three topics can be identified that are of particular importance when selecting URLs to obtain new web texts to upgrade the Gigafida corpus (tweets, forum messages, comments on news sites and blogs are discussed in the next section). These are topics that the current Gigafida corpus, with mostly printed texts and only a small and narrowly selected set of web texts, has poor coverage of, and

these thus needed to be examined if we want to describe their distinctive lexicons in a dictionary. For slWaC$_2$ (but not for Gigafida) the typical topics are *travel, tourism, online shopping* and the *World Wide Web* (see last row in Table 2). The topic *religion* is in this respect surprising, because it is the only one that could be better integrated into Gigafida by means of printed texts (wherein the response of the text providers is crucial).

At the end of such comparisons the question of the direct inclusion of the web corpus of Slovenian, slWaC$_2$, into the new Gigafida corpus arises. From the perspective of a more focused and controlled, as well as time-predictable and equable, form of text collection, with the explicit purpose of inclusion in the reference corpus, this question would be better answered in the negative, but it is not necessary to keep future upgrades of both corpora completely separate. On the contrary: as will be shown in section 5, these corpora are closely connected by their method of construction. Furthermore, the existence of two corpora of modern Slovene is also useful in terms of synergies, and as a demonstration of their differences and deficiencies.

3 WEB TEXT GENRES AND DICTIONARY SOURCES

On the Internet, the most influential medium of the 21st century, we are faced with a variety of communication environments or areas that apply all four basic functions of text (Skubic 1995; Mikolič 2007): cognitive, communicative, executive and art-expressive. There are also various discourse/speech communities that determine the characteristic language choices people make in the context of a specific discourse/speech.

Some of the features of web texts are tied to (more) informal speech situations, these are often manifested in texts in non-standard form (e.g. slang, jargon, vernacular language, and dialect). On the other hand, other web texts correspond to the concept of public communication in the narrow sense of the word (Škiljan 1999), and are written in accordance with standard language norms. The language heterogeneity of the Internet has caused changes in the language and the expansion of its lexis, so it is necessarily to find out which texts must be an integral part of any corpus that will be the basic source for a dictionary of modern Slovene (and at the same time, we can find out which texts it is possible, at the moment, to reject).

A description of the variety of online genres and their key factors is actually a rather difficult task, due to the extensiveness and uncontrollability of the material,

and the small number of studies of such genres and their target audience (Crowston 2010: 17, 26).

Nevertheless, on the basis of the analysed literature and related material, it can be seen that, for the analysis of online text genre variety and also for establishing the selection of web texts for the corpus, there are two key criteria, as presented by Herring et al. (2004):

- authorship or the relationship between the sender and the recipient (one or more authors, a formal or informal relationship) (see also Oblak et al. 2005),

- functions and the associated internal and external structure or form of the text, as well as multi-codes and updates (see also Bishop 2009; Crowston 2010).

The language choices of online authors depend on both these criteria, particularly in relation to conformity with the norms of standard language, or deviations from them.

From the perspective of the author, web texts are basically divided into:

- classic websites (HTML) with one single author or source of the texts,

- online community genres ("web-based community genres", Bishop 2009) with more than one author of the texts,

- blogs/blog writings (blogging) as an intermediate genre between one- and two-way communication.

a) Classic websites are mainly characterized by one-way communication (the most common exception here are media sites, which may include the forum messages, comments, or blog writing of the readers). The source of a website's text source is known or easily determinable. The relationship between sender and the recipient is mostly formal, and since texts address the general public they are mostly written in accordance with the standard language norms.

Among classic websites we place the following:

- Web portals (as well as Wikipedia, Wikisource, Wikiversity, Wikibook, and so on),

- media sites,

- commercial and corporate websites,

- websites of governmental and non-governmental organizations and local government bodies.

Personal websites are less formal and may be closer to the form and the purpose of blogging or the genres of online communities (such as on Facebook).

b) Online community genres ("web-based community genres") are related to collective-action oriented websites or interactive text forms of computer mediated communication (CMC), in which several authors collaborate. These genres are determined by the dominant actors, communication environment or topic, and the internal structure. The language choice here also determines the nature of the interactions among the actors. They are very diverse and often also anonymous, so the expressiveness of the texts is rather varied and they mostly include elements of vernacular, informal language genres. These genres are increasingly replacing speech communication, and so they are often manifested by written spoken language, and in some applications also by spoken text.

Among the online community genres we could place texts from various online tools and social networks, such as:

- forum messages (users are discussing a certain topic)
- Twitter, Facebook, Myspace, and LinkedIn (texts such as tweets, statuses or thoughts, status comments, photos, videos, hyperlinks, interest groups, event creation, invitations etc.),
- Instagram (publishing photos and hyperlinks to Twitter or Facebook),
- Ask.fm (users create an account and other users then ask them questions, also using hyperlinks to Twitter or Facebook),
- Snapchat (a mobile application through which users share thoughts, videos, photos etc. with their friends. Their messages disappear in a few minutes),
- Viber (a mobile application for smartphones, through which communication takes place via the Internet, can be written or spoken, and includes mobile contact list),
- web chats (diverse categories of "rooms" where users with same interests connect with each other).
- Comments on journalistic articles, videos and so on (comments can then develop into a discussion on a specific topic, usually between users, unknown to each other, and this functions according to the principles of a forum).

c) Blogs/blog writings are most often part of the journalistic genre, intended for a wider audience, and are often also in direct interaction with the readers. Usually there is one single author of a blog, and these can be written by professional (journalists) or unprofessional writers, so the language choice depends on the

communication competence of the author and especially on the target audience that the author wants to reach.

According to Domingo and Heinonen (2008) we can distinguish the following types of journalistic blogs/blog writings, which are differentiated by professionalism of the writers and degree of institutionalisation of the environment:

- citizen blogs (written by unprofessional writers outside media institutions),

- audience blogs (written by unprofessional writers within media institutions),

- journalist blogs (written by journalists outside media institutions),

- media blogs (written by journalists within media institutions).

In terms of *function*, texts are classified in groups of broad text genres and narrow text types and according to the following common properties: the purpose or influential role, recipient, reference and external and internal structure of the text (comp. Mikolič 2013; Nidorfer Šiškovič 2013). According to these properties, we also analysed the text in a web environment, where we referred to Crowston (2010) who summarizes the key typologies of Internet genres considering the purpose and form. Based on the findings of this, we can try to describe web genres in the context of the following groups (as summarized by Mikolič and Rolih 2015):

1. *Conversational and at least partially private text genres*: e-mail, web chats, tweets and other genres of social networks (e.g. Twitter, Facebook) and forum messages. The adjective "private": for these genres is based on their greater range of private language elements or content components of private communication spheres (Škiljan 1999), sociolects and idiolects (Skubic 2004).

2. *Promotional, advertisement and commercial text genres:* banner adverts, link collections, online shops, marketing and sales websites, personal websites, often with the purpose of self-promotion and marketing. The aim of these genres is to influence the consumer behaviour of the recipients. Due to their appeal to the general public, the language used in these genres generally does not depart from standard language norms, except when stylistic effects need to be achieved.

3. *Reporting/news and broadcast journalistic text genres:* journalistic texts of various genres, the online editions of print media, contributions associated with lifestyle (e.g. recipes, tips in the form of tutorials, guides for a healthy body, and so on). These are genres in which deviations from the standard language norms have only the stylistic role. The exceptions are the comments below the contributions on the major news portals, in which the authors do not usually follow such norms.

4. *Program text genres:* technical data/assistance/support, problem reports, and frequently asked questions (FAQ). These texts are messages from the operators or programmers of web pages. The text opens by discussing a problem and leads the user to a solution. Since it is a professional text aimed at the general public, the language is mostly consistent with standard language norms.

5. *Academic text genres* (accessible at sites such as Google Scholar): technical and scientific texts, written in line with standard language norms.

6. *Official and officiated text genres*: records of the meetings of state bodies, legislative websites, stock market websites, published policies, and so on; online administration, e-applications, etc. The purpose of these genres is to inform the general public about the key procedures, rules and laws in the country, and to enable working with the administrative authorities through the use of online forms. The language in these texts thus does not deviate from the standard language norms.

7. *Literary and semi-literary text genres*: these are belletristic texts, which are characterized by compliance with the standard language norms with an intentional deviation from it. The most common semi-literary web text genres are blogs and web diaries.

Undoubtedly, most online text genres – although still under-explored in both Slovenia and internationally – are very active in terms of their implementation and the development of language.

Due to the rapid development of online tools, some web genres may quickly become out-of-date (at this moment, for example, we are seeing the decline of web chats) and others will emerge, with similar or completely different intentions and linguistic characteristics. Therefore, in the preparation of the dictionary descriptions, we should not only *consider* the online linguistic reality, but also regularly *follow* it.

Of the various web text types described above, the new version of Gigafida should at least consider the content that has a known author or source and it intended for the general public. These texts should include those from large, mainstream websites and personal websites with large readerships, professional writers' blogs, the tweets and Facebook pages of individuals and institutions that have a great impact on general linguistic use (based on number of followers and media responses). Therefore, in terms of function, these texts include some of the conversational, promotional, advertising and commercial text genres, and all of the reporting/news and broadcasting, official and officiated and literary and semi-literary web text genres. As mentioned before, the main conditions for conclusion must be a high level of influence and large readership, and that the text's genre should be evident from the taxonomic categories.

4   USER-GENERATED CONTENT

A special challenge in contemporary lexicography is the vocabulary in user-generated content, published by regular people, and not professional writers. This kind of computer-mediated communication (CMC) is heavily characterized by varying degrees of interactivity, synchronicity and physical detachment. The more the selected medium is interactive, the more elements of spoken language it displays, including the CMC-adapted paralinguistic and prosodic elements (Crystal 2001). The most common features of this kind of language are non-canonical spellings, colloquial and regional expressions, foreign-language elements, non-institutionalised abbreviations, as well as neologisms. These make such texts extremely valuable for lexicographic purposes, but they are at the same time very difficult for automatic processing (Sproat idr. 2001), which is why the development of tools that can handle noisy texts from the web is currently one of the most active research topics in natural-language processing.

In contemporary linguistics, paradigms that consider non-standard language variants in computer-mediated communication as a sign of imperfect or impoverished communication abilities have become a thing of the past, since a number of studies have demonstrated that users adapt their language to maximise the potential and the functionalities of the medium in order to meet their communication needs with the least time and effort required, displaying their identity and spontaneous speech along the way (Herring 2001).

The discrepancy between the language as a living organism, and its static description that calls for research into non-standard language, has been addressed by several Slovenian linguists who have analysed the language of text messages, forum posts as well e-mail (cf. Kalin Golob 2008; Jakop 2008; Michelizza 2008). However, this kind of research is still not receiving enough attention by the mainstream linguistic community, and, as a consequence, the Slovenian linguistic landscape lacks a comprehensive description of the non-standard language varieties, as well as sufficient, publicly available collections of such text types.

JANES, the basic national research project, aims to close this gap and develop the resources, tools and methods need for the analysis of CMC (Fišer et al. 2014a). This section presents the interim results of the projects relevant for the construction of a modern dictionary of Slovene.

4.1   The JANES corpus of Slovenian user-generated content

The current version of the JANES corpus contains four types of user-generated content: tweets, forum messages, news comments and blog posts. Tweets have been

harvested for the past two years with a custom-built tool called TweetCat (Ljubešić et al. 2014). One-off crawling of forum messages and news comments was performed using designated crawlers and text extractors of some of the most popular or influential forums and news portals, based on their traditions, forms of text production and the number of users. Blogs were adopted from the de-duplicated version of the slWaC 2.0 corpus (Erjavec and Ljubešić 2014) by using the string "blog" in the domain name as a positive filter. This is only a temporary solution, as the lack of an internal structure of blogs makes it difficult to distinguish between the language of the main text of the blog and the language of the readers' comments on it. A designated crawler and text extractor for blogs that takes this into account will therefore be developed for the next version of the corpus.

All the texts along with the unified metadata are merged into the JANES corpus and formatted in a bespoke XML, thus enabling corpus structuring, metadata labelling and Unicode character encoding. The corpus is also annotated. Sentence segmentation and tokenization was performed with the standard mlToken library for Slovenian which is part of the ToTaLe (Erjavec et al. 2005) tool chain. Next, word forms were normalized with a character-based machine translation approach that was trained on 1,000 manually normalized key words obtained from the tweet corpus with respect to the reference corpus KRES (Ljubešić et al. 2014). Finally, the corpus was morphosyntactically tagged and lemmatized with ToTaLe, which was originally developed for standard Slovenian.

The JANES v0.3 corpus comprises 161 million tokens, most of which come from tweets (38%), followed by forum messages (29%), blog posts (24%) and news comments (9%). The corpus is already a useful resource for lexicographic work, since it is complementary to the reference Gigafida corpus in terms of content, is substantial in terms of size, and diverse in terms of the text types included. Further enhancement of the corpus by increasing the number of text sources, especially forums and news comments, would of course be highly desirable. It also needs to be noted that while the JANES corpus is limited to public CMC, lexicographers would benefit greatly from the private communication on social media, such as Facebook, which has 750,000 Slovenian users, as well as the new apps that are becoming popular with younger users, such as Instagram and WhatsApp, but also multimedia and video technologies, such as YouTube, Skype and FaceTime, that are taking the place of the traditional text messaging and on-line chats, as seen with MSN Messenger.

4.2   Non-canonical language in the JANES corpus

While it is true that the JANES corpus contains user-generated content, not all of it is written in non-canonical language. Quite the contrary, a quick manual

examination of a small sample of random tweets has shown that a large majority of them are in fact perfectly standard, which may seem surprising at first but since Twitter is used as a popular information dissemination channel, not only by individuals but also by news agencies, public institutions and companies, it is only natural that such communication is carried out in standard Slovenian.

In order to be able to focus on the analysis of non-canonical language, we have developed an approach to automatically measure the level of standardness of the input text at two levels: technical and linguistic (Ljubešić et al. 2015). Technical standardness considers capitalization, use of punctuation and spacing, while linguistic standardness takes into account spelling, lexical choice, word order and so on. A training set of tweets, forum posts and news comments was manually annotated for both standardness levels and scored from 1 to 3, with 1 meaning very standard and 3 very non-standard.

About 30 features that could serve as indicators of technical and linguistic standardness were defined at the character level (e.g. ratio of punctuation written to text length), string level (e.g. ratio of capitalized words written to text length) and word level (e.g. ratio of out-of-vocabulary words written to the Sloleks lexicon). The training set and the features were used to train a linear regressor that assigns a technical and linguistic standardness score to all texts in the corpus, enabling lexicographers to limit their searches to the desired level of standardness.

5   COLLECTING INTERNET CORPORA

Crawling is a process of automatically gathering documents from the web with the purpose of generating search engines indexes, retrieving other information from the web, or building corpora. High recall is the key factor in the former case, while the latter case strives toward acquiring clean linguistic content. Here, it is better to lose parts of the retrieved documents than to get a larger but very noisy corpus, which would contain elements such as the headers and footers of web pages, navigation elements etc. besides continuous text.

There are two basic approaches to crawling linguistically interesting data. The first, *generic* approach uses the same procedure for all documents. Its main advantage is easy implementation and wide scope in terms of text source and type. However, there are also disadvantages: data collected in this way contains more noise, has less structure and (almost) no metadata. For example, titles and subtitles are not identified, nor is the author, time and date of its publication. The second method is *target-oriented*, adjusting the implementation of crawling to individual document sources. The advantages of this approach are less noise, better

structure of collected documents and more metadata, while its weakness lies in having to adjust the crawling script for each source separately, which is time-consuming and also likely to stop working if the source modifies its platform.

The generic approach is used when building large collections of texts based on a common top-level web domain (e.g. ".si", ".uk") or the same language (e.g. slWaC corpus). The targeted approach is better suited for smaller textual collections built for specific research purposes where the structure of a text and its metadata are of key importance (e.g. the JANES corpus, described in the previous section).

Crawling typically starts with a pre-defined set of web documents, and continues with the crawler gathering new documents from hyperlinks in the existing set. The problem here is how to limit the set of collected documents to avoid gathering texts in the wrong language or genre given the purpose of the corpus compilation. There are two basic approaches when selecting which documents to crawl. The first is based on restricting URL addresses, e.g. to the domain ".si" or "med.over.net", while the second works with a list of keywords that define the target discourse domain, such as environment, tourism, cuisine etc. In this case, collecting URL addresses suitable for crawling is typically done through a search-engine API. When crawling documents for general web corpora, restricting URL addresses (e.g. for Gigafida) works best, whereas for specialised corpora keyword lists are more appropriate. Two well-known tools for the latter approach are Boot-CaT (Baroni and Bernardini 2004) and WebBootCaT (Baroni et al. 2006).

Web documents exist in a number of formats. The most important are *HTML* documents, which are problematic because a significant part of their content may refer to the appearance of the web page. Moreover, parts of these documents often have identical content, and in the case of textual corpora this signifies noise. Another format of documents that also contain linguistically interesting data, but is much more rarely gathered and processed, are PDF documents. The problem with collecting text from PDFs is that this format is meant for printing, so the text is encoded as characters with their positions in the page, making extraction of quality text often challenging. The following sections will thus primarily focus on describing how HTML documents are processed, while for PDF documents the extraction of content would need to be adjusted.

Another document type comes from web platforms where text is directly sent to the recipients as individual messages, similar to SMS messages or emails. By far the most well-known such platform is Twitter, a system that enables sending short messages to one's followers. Twitter also offers API scripting plugins that can be used for crawling tweets by individual authors or topics. As shown in the previous section, we gathered tweets for the JANES corpus with the TweetCat

tool (Ljubešić et al. 2014), which was purpose built for compiling tweet corpora of smaller languages. This tool, with the help of an initial language-specific word list, identifies users tweeting predominantly in the focal language (in our case Slovenian), and then via their friends and followers gradually enlarges its user base and collects their tweets together with the tweet metadata.

5.1   Procedures with generic crawling

As noted above, generic crawling is most often used when the goal is collecting a large quantity of text (more than one billion tokens) or when the human resources for collecting data are limited. The process of generic crawling for linguistically relevant data, as is also implemented in the system used for building the slWaC corpus, consists of several basic steps. The initial step is generating a *list of websites* to be crawled first. For languages with a relatively small number of speakers, such as Slovenian, this typically means a few better known websites in the language. The second step is *crawling*. Technically speaking, this step is performed by running multiple threads and searching for hyperlinks in a breadth-first approach, where the list of websites to be crawled next is updated dynamically by identifying hyperlinks from websites that have already been crawled. When a document is collected, the next step is to determine which *character encoding* is used. This piece of data should be documented in the metadata of an HTML document, yet in reality it is often missing or an incorrect encoding is declared. Determining the correct encoding system is thus mostly based on comparing the distribution of bytes in the textual part of the document to the distribution of bytes in a pre-determined set of documents with known encodings.

With generic crawling, it is not possible to define the document's structure in advance, which is why a generic program, such as jusText (Pomikálek 2011) or Boilerpipe (Kohlschütter et al. 2010), has to be used. Due to its generic nature, this step creates a document structure which does not go beyond the paragraph-level nor does it collect metadata. Typically, it also does not remove all non-textual noise from the document. Next, the *language of the document* needs to be identified. This step is necessary when building a corpus, since the web is a multilingual environment. An efficient tool for this step is the langid.py script, written in Python (Lui and Baldwin 2012). The last step is *removing (near) duplicates*, since identical or nearly identical textual content is often published on multiple URL addresses. Removing (near) duplicates is most often based on calculating the intersection of word n-grams from two documents. A typical heuristic suggests that if 7-grams of two documents overlap in more than half of the cases, one can be removed as a near duplicate.

The six steps described above are mostly executed separately, which makes crawling far from optimal. The only exception is SpiderLing (Suchomel and Pomikálek 2012), which has combined the steps from crawling to language identification into an integrated process, in which individual steps communicate with each other to optimise the quantity of the crawled data and the final size of the corpus.

5.2   Procedures with target-oriented crawling

Target-oriented crawling is used when fairly little data needs to be crawled, or when there are sufficient human resources to carry out the necessary steps. This type of crawling comprises three basic steps. Specialised corpora are most often built based on a certain content and not a specific web domain. The first step is thus identifying web domains or their parts which are likely to contain plenty of sought-after content. The technical as well as legal limitations of individual sources need to be taken into account, e.g. does the website prohibit crawling (with the use of robots.txt), does it offer API scripting plugins to collect data (e.g. Twitter), and does it perhaps even allow for the entire database of texts to be downloaded (e.g. Wikipedia). The latter two options substantially ease the process of data collection, while the use of technologies such as POST and AJAX requests makes writing extractors very difficult. The next step is *crawling*, which mostly gathers all or as many documents as possible from the chosen domains. The most complicated and time-consuming is the process of writing extractors, i.e. scripts used by programmers to describe the schema of a certain type of HTML documents. This often needs to be done separately for each source, especially if its structure is very complicated, e.g. when gathering news articles and comments on these in a chronological order.

5.3   Monitor corpora

The web is particularly suitable for building monitor corpora, since its content is constantly being updated. Once the crawl platform is set up, it is simple to gather new data. This holds true for generic crawling and somewhat less so for target-oriented crawling, since individual sources may change the structure of their website, causing the original target-based extractors to stop functioning correctly.

The best tool for continuous crawling of the web are search engines, especially Google, but also local search engines (e.g. Najdi.si in Slovenia), since they are continuously trawling the web, searching for new texts. Although it is difficult

to imagine using such highly intensive processes for linguistic purposes, this can serve as the upper bound of what could be collected, and it depends on each project and the needs and abilities of its researchers how often the chosen content should be re-crawled. For researchers in lexicography, a monitor corpus would certainly be a valuable tool to detect larger and more sudden lexical changes caused by events and phenomena widely covered by media reports. and thus prompting the interest of speakers – the potential users of a dictionary. Once the first version of such dictionary is complete and made available, its authors might like to add continuous updates to its contents. In this case, building a monitor corpus and defining methods to detect new lexemes, semantic changes or changes in the characteristic context of words becomes even more important or, rather, of key importance.

6   CONCLUSION

In modern linguistics the paradigms that are used to show non-standard language versions of written communication on the Internet as a reflection of failure or pauperism of communicative abilities have somehow survived, because the analysis of language used on the Internet identifies users' ability to adapt to the electronic media or the ability to utilise the media to meet their communication needs, as they endeavour to shorten and simplify the written communication, and especially to adjust the writing to their identity (Herring 2001).

Nowadays Internet communication actively complements and changes the characteristics of the Slovenian language written for the public, to the point where a modern dictionary can no longer ignore it. In this chapter we tried to show how the web part of Gigafida can be upgraded both in its volume as well as in its topic and genre terms, and warn that such new texts should be placed into the corpus in a transparent manner (i.e. with more elaborated taxonomic categories). Some of the online genres are written in a non-standard Slovenian, which confronts corpus linguistics with an additional language technology challenge: overcoming the barriers to its automatic processing. The resources and tools with which we can help ourselves in this task are already arising in Slovenia, and different methods of crawling the Web are already being tested. The aim of the Gigafida corpus, as expanded in the proposed way, is therefore to include publicly available written production of Slovenian on the web in a broader sense; leaving the process of selection and interpretation of data from such a corpus for the needs of the dictionary to the actors in the next phase of this process, and thus to lexicographers.

Language Technologies and Corpus Encoding

Tomaž Erjavec, Peter Holozan and Nikola Ljubešić

Abstract

This chapter provides an overview of the levels of basic automatic linguistic annotation that should be applied to corpora to be used as the basis for lexicographic analyses of contemporary Slovene, as well as for other purposes. We give an overview of existing research in this field and then focus on a concrete set of open source and mainly language independent tools and their models for Slovene, and give suggestions for their improvement. A short description of the proposed corpus encoding process is also provided.

Keywords: linguistic annotation, corpora, annotation format

1   INTRODUCTION

This chapter deals with linguistic annotation of corpora which could serve as the basis for lexicographic work, and also provides suggestions for the format of the corpus annotations. The chapter does not cover all tools that are useful for lexicographic work, but only those that generate annotations to be included in the corpus and then used as a source of knowledge by down-stream programs, from concordancers to synonym extractors. In addition, the focus is predominantly on programs that have been developed for the Slovene language. The following levels of annotation will be taken into account, and are listed in typical order of appearance in the processing chain:

1. **Tokenisation**, which divides the text into individual tokens, either words or punctuation. This step can also identify token types, such as numerals, abbreviations, URLs, emoticons and emoji. Sentence segmentation is often performed in the same step.

2. **Normalisation**, which transforms (translates) non-standard word forms (found, for example, in historical texts and in user-generated content) into standard ones. This is useful for easier searching and better performance of the annotation tools developed for standard language.

3. **Morphosyntactic (or part-of-speech) tagging**, which assigns to each word token a morphosyntactic description (MDS), e.g. *Ncmdn* which decomposes to a common noun of masculine gender, dual number and nominative case.

4. **Lemmatisation**, which assigns the base form to a word, used for dictionaries or lexical look-up.

5. **Parsing**, which gives a syntactic analysis to each sentence of the text.

Apart from parsing, all the above levels of annotation are necessary for a corpus to be useful in lexicographic work. Some other levels of annotation might also be useful, but these are difficult to place at a fixed position in the processing chain, as this depends if they require (or can use) all or some of the above annotations. Depending on the method, these further annotation tools can use raw or tokenised text, which could furthermore be tagged or even parsed. Some of these tools have already been developed for Slovene, even if only as prototypes:

6. **Named entity recognition**, which identifies proper names in the text and classifies them, for example, into personal names, geographical names, and names of companies or institutions. Additionally, some systems identify numerical and other expressions and classify them, such as into currencies, dates and so on.

7. **Term extraction**, which identifies potential terms in the text. However, it should be noted that what constitutes a term is fairly problematic, as this depends on the subject area, target audience and the like.

8. **Semantic information annotation**, which labels words or phrases with their meaning by relying on a semantic and lexical resource; it can also link them together according to their semantic roles. Although such annotations could be extremely useful for lexicographers, the complexity of the annotation causes existing programs to generate results that may not be accurate enough to be of value.

Generally speaking, annotation (and other) language technology tools come in two varieties:

- Tools that use handwritten rules, which require a lot of human work but may give (depending on the level of annotation) very good results. Such tools are often used for text segmentation, e.g. into tokens or terms, and traditionally also for morphological analysis. For some levels of annotation, most notably morphosyntax and syntax, the number of necessary rules becomes extremely large, which makes their development and debugging very difficult, costly and error prone.

- Tools that learn a language model from training data, i.e. manually annotated corpora or other language resources. Machine learning methods are being developed at a fast pace, yet in order to generate quality models we typically need extensive language resources – and building these is a time-consuming and expensive process. On the other hand, once a training dataset has been built, it can be used to train and test various machine learning tools with the best one chose for the task at hand.

Both types of tools generally use background language resources, especially lexicons.

2 OVERVIEW OF TOOLS FOR THE SLOVENE LANGUAGE

Tools for annotating Slovene language texts at all the above-listed levels have already been developed, although a number remain at the prototype stage. This section will only focus on those which are still being maintained and, for the most part, which are freely or openly available. As such, one of the earliest tools for morphosyntactic annotation of Slovene (Jakopin and Bizjak Končar 1997) will not be included.

2.1 Tools developed by Amebis

The tools developed by Amebis are not openly accessible, yet their system of annotation tools and background resources – not only adapted to Slovene but written exclusively for it – has the longest tradition of development. These tools were initially developed for the Besana grammar checker (Holozan 2012) and Presis machine translator (Romih and Holozan 2002). They are written in the C++ programming language and work in 32- as well as 64-bit versions. Their structure slightly differs from the classic one: tokenisation is still performed first, but its key feature is the treatment of special tokens such as Web and e-mail addresses, phone numbers and emoticons as one unit.

The next step is the tagger, which uses a lexicon to annotate words with all possible combinations of lemmas and morphosyntactic tags (there are currently 7.6 million elements in the lexicon). The tagger also recognizes special tokens such as Web addresses, chemical formulas and emoticons. Simultaneously, it searches for potentially misspelled words and typical non-standard forms; some of the latter are already included in the lexicon with special morphosyntactic tags. The last step is performed by the analyser, which chooses the most likely pair of the lemma and morphosyntactic tag for each word. At the same time, it performs syntactic analysis and lists word meanings taken from the Ases database (Arhar and Holozan 2009), both by using an interface language developed at Amebis (Holozan 2011). If necessary, the analyser can also modify tokenisation or textual segmentation, e.g. when tackling examples such as "*ga*" as the pronoun "*him*" in Slovene, or "*ga.*" as an abbreviation equivalent to "*Mrs.*" in English. Examples such as "*Prišla je še ga. Micka.*" (*Mrs. Mary also came.*) are easy to handle, but cases such as "*Videl sem ga. Micka ga je tudi videla.*" (*I saw him. Mary saw him also.*), are much harder and may cause a naïve tokenizer to fail and identify the first instance of "*ga.*" as one token (abbreviation), and the text as one sentence, while the Besana tool would correctly identify two tokens (pronoun followed by a comma) and two sentences.

Amebis' tools rely on handwritten rules and data from the Ases database for their operation. The most important concept from the database is verb templates (Holozan 2011), which comprise data on valency. Many proper names have also been entered; they are divided into 20 categories (which enables named entity recognition). A special script tokenises, lemmatises and assigns MSD tags according to the specifications of the "Communication in Slovene" project and implemented in the Obeliks tagger (more on this in one of the following sections).

The tools by Amebis were also used to annotate two large reference corpora, Fida and FidaPLUS. The main obstacle to the wider usability of Amebis' tools

is license ownership. The tools are currently not open source, and thus an agreement needs to be signed with Amebis for their use.

2.2   The To(Tr)TaLe tagger

The ToTaLe tagger (Erjavec et al. 2005) was developed at the Department of Knowledge Technologies at the Jožef Stefan Institute within the framework of several projects. The tool implements a pipeline composed of three modules: a tokenizer, which also segments the text into sentences, a MSD tagger and a lemmatiser. A module called mlToken is used for tokenisation; it is a multilingual tokenizer that uses language dependent lists, e.g. of abbreviations or rules on how to write numbers, to adapt to a particular language. MSD tagging is done with the TnT tagger (Brants 2000), a relatively old trigram-based tagger which uses models trained for a specific language on a manually annotated corpus; it can also use a background lexicon. The current model is trained on the jos1M corpus (Erjavec et al. 2010; Erjavec and Krek 2010), and uses tokens from the FidaPLUS corpus (Arhar and Gorjanc 2007) as a background lexicon. The lemmatisation module uses a program called CLOG (Erjavec and Džeroski 2004), which assigns the base form to each word form according to its MSD tag. This program also relies on an automatic lemmatisation model, based on a training dataset, which consists of a list of triplets (word form, MSD tag, lemma). The training set for Slovene has been generated by combining the tokens from jos100k, manually checked tokens from jos1M and selected words from FidaPLUS. ToTaLe is available online and has been used to annotate most corpora that can be accessed through the noSketchEngine (Rychly 2007) concordancer installed at nl.ijs.si (Erjavec 2013).

ToTaLe is written in the Perl programming language, and the same applies to modules for tokenisation and lemmatisation. Although Perl is no longer a very popular language, it can still be used with all main operating systems. However, the TnT tagger is not open source, and is only available for non-commercial use as an executable under Linux, which is why in its current state ToTaLe cannot be made openly accessible nor used on OS Windows.

A tool called ToTrTaLe has also been developed, and this differs improves on ToTaLe in two important ways. First it includes an (optional) transcription module and, second, unlike ToTaLe, which expects raw text as input and outputs a tabular file, ToTrTaLe expects a TEI-compliant XML file at input and also returns a TEI XML file as output. The transcription module is intended for modernising historical word forms in older (Slovene) texts; by working on normalised forms, the MSD tagger and lemmatiser produce much better results, as they are both trained to process texts in contemporary Slovene. For

modernising the tokens, the transcription module uses a tool called Vaam (Reffle 2011), which uses handwritten rules on how to modernise historical Slovene word forms. To date, only the IMP corpus of historical Slovene (Erjavec 2015) has been annotated with ToTrTaLe.

Both tools give relatively good results, but their maintenance could be improved. For example, it would be worth re-training the models for the Slovene language, since better resources have since appeared, most notably the Sloleks lexicon (Arhar 2009; Dobrovoljc et al. 2013) and the ssj500k corpus. Moreover, the programs that implement individual modules are, by now, rather outdated. At the very least, TnT should be replaced by a newer tagger, which should be open source and system independent.

As mentioned earlier, normalisation in the context of modernising historical Slovene words has already been implemented in ToTrTaLe. However, the rules were written manually, and since their implementation automatically trained normalisation models using character-based statistical machine translation (Scherrer and Erjavec 2013) have been shown to perform better. A standard tool that can implement this method is Moses (Koehn et al. 2005), a statistical machine translation system that was, for the task of modernising Slovene words, trained on word pairs of a historical (non-standard) word and its modernised (normalised) version. This approach is useful not only for modernising historical words, but also for standardising contemporary texts with non-standard orthography, such as texts in computer mediated communication. Character-based statistical machine translation has already been tested for standardising words in Slovene tweets (Ljubešić et al. 2014), and has produced promising results. An issue related to normalisation is which texts to normalise: if normalisation is also used on texts with standard orthography, it is likely that completely standard words would be "normalised" as well, doing more harm than good. We have trained a system that uses machine learning on a small sample of Slovene tweets and other user-generated content from the Web, all manually annotated with their level of standardness to estimate (and annotate) how non-standard new texts are (Ljubešić et al. 2015). Normalisation could then be used only on texts that have been automatically annotated as non-standard.

2.3   Obeliks tagger and parser for Slovene

As part of the "Communication in Slovene" project, a tool called Obeliks (Grčar et al. 2012) was also developed. As with ToTaLe, the tool tokenises the input text, segments it into sentences, adds morphosyntactic tags and lemmatises it. It uses a module with handwritten rules for tokenisation, a purpose-built machine

learning tool for morphosyntactic tagging, and the machine learning LemmaGen program (Juršič et al. 2010) for lemmatisation. The MSD tagger is special in terms of not relying solely on a model automatically generated from a training corpus, but also using handwritten expert rules, which filter hypotheses generated by the model, and combining the results of the lemmatiser and the tagger, assuring that they are not contradictory. Obeliks has been trained on a manually annotated corpus (Arhar 2009; Krek et al. 2013c), and gives the best results among those tools for Slovene that are publicly accessible. At the moment, the tagger's main problem is probably its implementation in the C# programming language, which is designed to work on Windows and cannot be easily used on other platforms, such as Linux.

Obeliks was also used to annotate the Gos corpus of spoken Slovene (Verdonik and Zwitter Vitez 2011), the KoRP corpus of public relations texts (Logar 2013), the Šolar developmental corpus (Rozman et al. 2012) and the Gigafida corpus; the annotations from Gigafida (as well as the texts themselves) are also part of the KRES, ccGigafida and ccKRES corpora (Erjavec and Logar 2012).

A parser for Slovene (Dobrovoljc et al. 2012) was also built within the above-mentioned project, where the well-known MSTParser dependency parser (McDonald et al. 2006) was trained on the dependency annotated part of the ssj500k corpus. The parser gives relatively good results, but – as is common for any linguistic annotation, especially parsing – its accuracy depends heavily on the genre of the text – the more the genre differs from that of training dataset, the poorer the results. An evaluation of the parser showed that its accuracy also depends substantially on the type of dependency relation, since this ranges from 54% to 96%.

2.4   Other tools

Named entity recognition (NER) for Slovene is supported by two tools. An NER tagger was developed (Štajner et al. 2012) which uses machine learning based on conditional random fields, with the model trained on ssj500k. The tool is available under an open license, but is rather difficult to use since its installation and use are relatively poorly documented. The second tool (Ljubešić et al. 2013) is based on StanfordNER (Finkel et al. 2005), which also works with conditional random fields and was also trained on ssj500k, but in combination with the slWaC corpus of the Slovene Web (Ljubešić and Erjavec 2011). The latter allows the tool to collect more accurate information on features and their distribution, which proves to be very efficient in decreasing the number of false positives as well as improving recall. The models are openly accessible, and StanfordNER is well maintained and documented.

Term extraction for Slovene has been implemented and studied through a number of experiments (Logar and Vintar 2008; Vintar 2009; 2010; Logar et al. 2013), yet these are not available under an open licence nor maintained. The tools are mostly based on a combination of linguistic knowledge about terms (especially which patterns of MSD tags can represent terms) and mathematical knowledge about the distributional features of word sequences in corpora. Identifying terms by using machine learning methods has not been tried yet for Slovene, and there are also no openly accessible training sets that could be used for this purpose.

3 GUIDELINES FOR THE FUTURE ANNOTATION OF CORPORA

3.1 Improving annotation schemes

Before focusing on how to improve the tools or corpora used for training, it is necessary to discuss annotation schemes that (manually) annotated corpora are based on. The design of these schemes should be re-thought and tested to improve the accuracy of the tools, while also preserving or even improving the linguistic informativeness of the individual levels of annotation.

Grammatical information about individual words from corpora such as ssj500k, Gigafida, and KRES, as well as from the morphosyntactic lexicon Sloleks, is based on the morphosyntactic specifications developed in the project JOS "Linguistic Annotation of Slovene" (Erjavec and Krek 2008). This system originates from and is in line with the MULTEXT specifications (Ide and Véronis 1994), or its subcategory MULTEXT-East. The MULTEXT-East 4.0 specifications (Erjavec 2012) cover 12 languages, including almost all Slavic languages, and are, for Slovene, identical to the JOS specifications.

The JOS specifications define 12 parts of speech: noun, adjective, verb, adverb, pronoun, numeral, preposition, conjunction, particle, interjection, abbreviation and residual. The majority of these contain information on their morphosyntactic features, either lexical ones, such as is the noun common or proper or the verb auxiliary or main, and inflectional ones, such as number or case. All valid combinations of a part of speech and its features and encoded as strings (MSD tags), where each position in the string represents a certain attribute; its value is expressed through a one-letter alphabetic character. For example, the meaning of the string *Ncmsn* is *part of speech = Noun, type = common, gender = masculine, number = singular, case = nominative.* String encodings as well as features (i.e. attributes and their values) are available both in Slovene and in English. JOS comprises 1,902 different MSD tags, which are listed in the specifications together

with corpus examples. MSD tags, such as *Ncmsn*, are then used in morphosyntactically annotated corpora, and also in the morphosyntactic lexicon Sloleks to define paradigms of word forms for individual words.

JOS tags are used in many corpora, but the full set of the tags and their features may not be the most suitable for all applications. It is possible to opt for a pruned tagset, excluding features where morphosyntactic taggers are most error-prone (e.g. grammatical case), or all inflectional features if lexical features suffice for the purposes of the project. Such alternatives, which reduce the size of the JOS tagset and increase the accuracy of taggers, have already been used: a detailed study in Krek (2011) suggested several options on how to reduce the tagset, while Erjavec (2013) reduced the set to 32 tags, which are limited to the part of speech and some of its lexical features. However, more studies would be needed to determine what the optimal tagset for each specific purpose would be.

More recently another interesting possibility has appeared, as with the ongoing Universal Dependencies project (Nivre et al. 2105) specifications and treebanks for many languages are being developed, including for Slovene. In addition to defining syntactic relations, the project offers a universal set of morphosyntactic features (with optional language-specific extensions). Although the drive toward universality inevitably leads to lower adaptability of the scheme to individual languages, its subsequent comparability between many languages may outweigh individual cases of poor performance.

There is an even greater need for additional studies on the set of syntactic tags and relations from the JOS and "Communication in Slovene" projects (Erjavec et al. 2010; Arhar 2009), since these have not yet been thoroughly tested. One of the issues is parses with multiple roots, which pose a problem for automatic parsers (Javoršek 2015). However, both theoretical and practical recommendations from the Universal Dependencies project should also be taken into account in the further development of a syntactic annotation system.

Current categories for named entities as used in the ssj500k corpus have also proven deficient. Štajner et al. (2012) showed that by dividing the "other" category (i.e. for those named entities that are neither personal nor geographical names) into names of organisations and "other" not only gives a more fine-grained set of categories, but also improves the overall quality of annotation. This conclusion is in line with findings from the authors of the Czech corpus of named entities CNCEC, which has no less than 62 categories of named entities (Ševčíková et al. 2007). Here it was found that reducing the number of named entity categories also led to worse results in annotation. Therefore, further increases in the number of named entity categories should also be considered for Slovene.

3.2   Improving the accuracy of tools

The tools described above differ in the quality of annotation as well as their ease of use. It would be useful to increase the accuracy of annotation in all of them, since every error is problematic in two respects. On the one hand, low precision brings noise to the dataset, since lexicographers also obtain incorrect results to their queries. For example, when querying a certain lemma, the words that were incorrectly annotated with this lemma will distort the overall picture regarding concordances, collocations, word sketches etc. However, here the lexicographer can at least go through the examples, decide which are correct, and discard the rest, a time-consuming but doable task. The second issue is low recall, which is worse. In this case, lexicographers cannot obtain some of the results they are interested in, because the tools fail to discover them: if a word is completely or mostly incorrectly lemmatised, it will not be found when searching its lemma or there will be few results. The main goal of automatic corpus annotation should thus be improving both the accuracy and recall in all processing steps – especially the initial ones (tokenisation, MSD tagging, lemmatisation), since every error gets multiplied further down the processing chain, making lexicographic analysis much more difficult to carry out.

It is becoming obvious for most annotation levels that better and quicker results are achieved through machine learning rather than with handwritten rules. But machine learning requires high quality manual annotated datasets for training, and such datasets are also needed for testing the quality of the tools, regardless of whether they use machine learning methods or handwritten rules. To improve the quality of the tools, it would thus be useful to increase the size as well as the diversity of manually annotated corpora. Here, it would not be necessary to annotate entire texts – methods of active learning could pick out examples that would be most useful in helping improve the trained model. Depending on the level of annotation, it would make sense to also expand supporting data sources, particularly lexicons and lexical databases, since these can offer linguistic information in a refined form.

Also noteworthy is the conceptual model of annotation or expanding support sources that build on a "virtuous circle": additional manually annotated corpora train the programs for better annotation, which results in a better basis for the next cycle of manual annotation, and this circle or rather spiral can be repeated multiple times.

The accuracy at all annotation levels heavily depends on the tokenizer, since its errors are transferred into all further annotation steps, while the errors in tokenisation directly block the possibility of finding incorrectly tokenised words.

Therefore, a lot of effort has already been put into building a reference tokenizer for Slovene (Krek 2011), which is, to an extent, already implemented in Obeliks, although its functioning could be further improved. For example, the tokenizer does not recognize the already mentioned "*ga.*" as an abbreviation in sentences such as "*Spoštovana ga. Micka!*" (*Dear Mrs. Mary!*). But one also needs to be aware of the fact that every change to a tokeniser that was previously used to produce existing (also manually) annotated corpora results in incompatible resources, which has a negative effect on annotation as well as on extracting grammatical information from corpora. Such cases come to light when, for example, corpora annotated with ToTaLe are used together with those annotated with Obeliks. A study of Web-specific vocabulary, where keywords of slWaC in comparison with the KRES reference corpus were examined, has found many "key words" to be exactly those that are tokenised differently in ToTaLe versus Obeliks (Erjavec and Ljubešić 2014).

When improving morphosyntactic tagging, it is useful to carry out experiments on which methods or combinations of methods truly generate the best results. It has already been shown, for example, that the use of meta-learning, which combines the results of Amebis' rule-based tagger and the TnT statistics-based tagger, gave better results than either tool used separately (Rupnik et al. 2010).

3.3   Improving the technical side of tools

In addition to improving the accuracy of the tools, other technical improvements could also be made, i.e. simplifying their installation and use, as well as improving the ease of their integration.

A general recommendation is to use open source tools that are independent of the language and the computer platform, are based on machine learning, well documented, and maintained on one of the platforms for revision control, such as Git, which have an active group of developers and users that can communicate through a forum, report errors or send suggestions for improvements. Two examples of such platforms are Moses and – to some degree – StanfordNER. Even though purpose-built tools for Slovene would have the advantage of being better fitted to the specifics of the language (or its theoretical linguistic framework), it is arguable whether this outweighs the amount of work needed for their development and maintenance. Such tools may produce good results at a certain stage in their development, but it is very likely that continued progress in machine learning will bring increasingly better results. It is therefore more reasonable to put the effort into developing annotated corpora of Slovene that can serve as quality training sets rather than into complex rule-based tools built solely for Slovene.

One possible approach to building annotated corpora for lexicography is deciding that the accessibility of tools is irrelevant as long as they are successfully used for corpus annotation, yet this makes expanding and also maintaining the corpora more difficult. Additionally, it would not be possible to use these tools for other purposes nor annotate other corpora that are not related to this particular lexicographic project or study. Using closed and proprietary tools also prevents the results of annotation from being checked or reproduced.

The next issue is the connectivity of individual tools, both regarding their import and export formats as well as their limitation to certain computer platforms. The use of the above-mentioned implementation of Obeliks, currently the best openly-accessible MSD tagger for Slovene, is limited to the Windows OS. This makes it incompatible with the Linux environment, which is traditionally much better equipped with open source taggers and other tools. However, there is a growing number of platforms which enable users to set up and run online workflows, such as WebLicht (Hinrichs et al. 2010). These systems implement individual programs or modules in such a way that they run as Web-based services, possibly in computer clusters, while the execution of the workflow as a whole (e.g. tokenisation → MSD tagging → lemmatisation) is controlled by a central server, which calls these Web services as specified by the workflow. Perhaps this model of annotation is where the future lies, but the current solution – especially for processing large corpora – is still execution on local computers that are connected into clusters and typically possess large processing as well as memory capacities. It is thus crucial for annotation tools to be independent of the operating system or platform on which they are being executed. In practical terms, this requires them to be written in one of the standard open source programming languages, such as Java or Python.

Besides platform independence, compatibility of import and export formats needs to be specified, e.g. as done WebLicht; more on this in section 4.

Another interesting challenge for scientists and developers is the architecture of system for text annotation. Most current implementations function by choosing the best candidate at every step of annotation. Yet the best candidate from one step may turn out to be the wrong choice when more information becomes available, as only later steps in the processing chain could correctly disambiguate among potential candidates. For example, only when taking into consideration the syntax can the system determine that the first "*ga.*" from "*Videl sem ga. Micka ga je tudi videla.*" is not an abbreviation but a pronoun followed by a period that marks the end of the sentence. A newer trend in this field are systems that use Bayesian networks (Finkel et al. 2006) instead of a simple pipeline of taggers. In the former, each tagger represents one variable of the system, allowing it to make approximate assumptions that determine the best tags globally.

3.4   Proposed chain of annotation tools

To annotate corpora of Slovene language for lexicographical purposes, only those levels of annotation were chosen where existing tools are already readily available, rather than all possible or potentially useful ones that are yet to be developed. The following paragraphs describe a suggestion for a chain of annotation tools. Out of all tools presented above only fully open source tools, both in terms of software and models of the Slovene language, have been chosen. For each tool some fairly simple suggestions for improvement are given.

- **Obeliks**: tokenisation, MSD tagging and lemmatisation. It would be useful to implement the systems in one of the standard programming language and re-train its morphosyntactic and lemmatisation models. Word normalisation could be added instead of having it implemented in a separate module.

- **Moses**: normalisation of word forms. The program should support several normalisation models, at least one for modern non-standard Slovene and one for historical Slovene. If the text needs to be normalised and, if so, which model to use could be either decided automatically according to the content or based on the metadata of the text.

- **MSTParser**: shallow parsing. The existing model for syntactic analysis could be used, but it would be useful to implement a conversion from the JOS scheme to the Slovene version of Universal Dependencies and train the parser on the latter, too. Corrections in certain parts of the training corpus ssj500k would be useful, as well as increasing the size of genres that are currently poorly represented but seem to be syntactically different from those already in the corpus. Experiments could also be made with some other, more contemporary parsers to see if better results could be obtained.

- **StanfordNER**: named entity recognition. The size of the dataset used for training (at the moment only ssj500k) could be increased and, more importantly, made more heterogeneous.

- As mentioned, there is still no maintained and openly accessible tools for annotating terminology, which is why term extraction should probably be programmed from scratch. The existing patters of morphosyntactic tags that represent potential terms could also prove useful.

One question is still open: how to link the above-listed tools, which are quite heterogeneous. For efficient automatic annotation of large corpora, the best solution is the installation and parallelisation of a chain of taggers on high capacity

Linux servers or clusters of such servers, where the conversion of their input and output formats should be implemented in such a way that they are compatible. One possibility is directly using the TEI format, but a stand-off format would be more appropriate to make the taggers work more efficiently; more on this in the following section.

4   ANNOTATION FORMAT FOR CORPORA

The structure of a corpus may be very complex, both regarding its metadata as well as linguistic annotations. Slovene corpora mostly follow the TEI guidelines (TEI 2013), which cover nearly all the above-mentioned annotation levels, as well as some others. The guidelines are well maintained under the auspices of the international TEI Consortium. The TEI format is also supported by a plethora of tools for building custom-made XML schemas and converting from and into various formats, such as from Word to TEI or from TEI to HMTL. The names of TEI elements have been translated into Slovene and are being applied by a number of users in the field of digital humanities (Erjavec et al. 2004; Ogrin et al. 2013).

In TEI the majority of linguistic tags are written directly as XML elements, using e.g. <w> for a word and <name> for a name. The advantages of such an in-line approach are the transparency of elements and simple formal validation of the format; the tags as well as the text can both be simply corrected. This solution has also several weaknesses: the elements need to be correctly nested (XML primarily supports tree structures), and with an increasing number of elements the XML becomes more difficult to understand and control. Moreover, the files with in-line elements can become quite large. These are the reasons why annotation schemas intended primarily for automatic annotation more often use the stand-off approach, where the base text remains unchanged. Instead, the annotations generated by individual tools point to the corresponding parts in the text or one of the element layers. Such an approach is used by the above-mentioned WebLicht (Hinrichs et al. 2010), which has built a shared corpus format called TCF. The same approach is also defined by the MAF standard, which is used to annotate morphosyntax (ISO 24611, 2012).

Although the stand-off approach is technically simpler and offers greater flexibility, it makes it harder to discover errors and link together individual annotation levels. Furthermore, the data to which the annotations point to must not be altered, or else the pointers become invalid. This becomes problematic when the text itself or (some of) the elements would need to be corrected, either manually or semi-automatically. The TEI format is thus primarily suitable for manually

annotated reference corpora, where it is crucial to have the tags and the corpus format as thoroughly checked as possible.

Most corpora mentioned in this monograph are TEI compliant, but the use of its recommendations is complex. What is more, it is now over two decades since the compilation of some Slovene corpora, and the TEI guidelines have been since then modified a number of times. When adding new annotations, some past decisions may prove hard to generalise. Therefore, it would not only be useful to annotate existing corpora with new tools and models, but also to standardise their encoding, which could then serve as a reference for the corpora needed for a new dictionary of modern Slovene.

5 CONCLUSION

This chapter has provided an overview of the levels of automatic linguistic annotation that should be part of the annotation of corpora to be used as the basis for lexicographic analyses of modern Slovene, as well as for other purposes. We have given an overview of existing research in this field and then focused on a concrete set of open source and mainly language independent tools and their models for Slovene, and provided suggestions for their improvement. A short description of the proposed corpus encoding has also been provided.

Annotations in large corpora are always assigned automatically, which is why users of need to be aware of the fact that such tools will inevitably make errors, resulting in poorer performance with regard to extracting information that is relevant for lexicographical or similar purposes. Further improving the accuracy of these tools thus remains a priority.

Dictionary of Modern Slovene: lexicographical process

Polona Gantar, Iztok Kosem and Simon Krek

Abstract

This paper describes each phase in the compilation of a database that is to be used as a basis for an online dictionary of modern Slovene and in developing Slovenian language technologies. A proposal for archiving different versions of entries, as well as different versions of the entire database during the compilation process, is also presented. Furthermore, we describe how to include detecting lexical change (the continuous updating of headwords) and dictionary users in the process. This is an important issue in electronic lexicography, but one that still leaves many questions unanswered.

Keywords: lexicographical process, automatic data extraction, online dictionary, detecting lexical change, gradual dictionary compilation

1   INTRODUCTION

The compilation of dictionaries in the digital age is closely linked with modern way of life and access to different types of information via computers or various mobile devices. It is largely driven by the reliability, rapid and free access, and customizability of dictionary content, three characteristics also most valued by dictionary users (Müller-Spitzer et al. 2011). As a result, lexicographers and dictionary publishers are looking for solutions how to provide quality language descriptions with minimum investment of time and money, as well as keep them regularly updated. As leading lexicographers have been pointing out for some time now, it is clear that paper dictionaries, although still present, are becoming obsolete and will eventually no longer be compiled (cf. Krek 2011; Rundell 2014).[1] It is for this reason that planning the compilation of a dictionary is even more important for the language community, especially in Slovenia, where there is currently no corpus-based description of Slovene in existence, and the compilation of the new version of the *Dictionary of Slovene Literary Language* (DSLL2) [2] follows the principles of paper dictionary compilation.

The proposal for the compilation of a dictionary of modern Slovene Language (DMSL: Krek et al. 2013b: 52–60) presents a procedure of dictionary compilation in phases that allows gradual release of dictionary content according to the degree of lexicographic analysis and the amount of information in the entries. The proposal also describes the procedure for regular updating of dictionary entries (ibid.: 46) and the method of prioritising entry treatment (ibid. 45). This paper aims to provide a more detailed description of each phase of the proposed lexicographical process that will meet long-term lexicographic challenges and efficiently utilise all the ICT knowledge and language technologies available, both in terms of methods for extracting language data and ways of presenting dictionary information to users. As the lexicographical processes by which the compilation of digitally-born corpus-based dictionaries are still relatively poorly described,[3] this paper also addresses highly relevant topics such as the inclusion of the language community in the compilation of a dictionary. Furthermore, the problem of the continuous release of dictionary entries is also discussed in this work, including archiving of dictionary information and a developing database version control process.

---

1   The future of lexicography was also discussed at a round table titled *Will there still be dictionaries in 2020?*, held at the conference Electronic Lexicography in the 21st Century (eLex, Bled, 10–12 November 2011). A video of this is available at http://videolectures.net/elex2011_bled/.

2   The dictionary is already being compiled at the Fran Ramovš Institute for the Slovenian Language (FRISL). The dictionary is currently the only general monolingual dictionary receiving government funding, which is problematic because it is based on a concept that focuses on a paper format and thus static dictionary content, and disregards state-of-the-art lexicographic and language technology approaches.

3   It was for this very reason that the description and planning of the lexicographical process was the topic of one of the workshops of the European Network of e-Lexicography (ENeL) held in July 2014 in Bolzano. The related contributions can be accessed at: http://www.elexicography.eu/working-groups/working-group-3/wg3-meetings/wg3-bolzano-meeting/.

2   PHASES IN DICTIONARY COMPILATION

As a carefully planned process of dictionary compilation, the lexicographical process is one of the key organisational and logistic tasks that affect both the formation and organisation of a lexicographic team, as well as the project timeline and finances. As pointed out by Tiberius and Krek (2014), the existing literature mainly provides descriptions of lexicographical processes in relation to paper dictionary compilation (see Dubois 1990; Landau 1984; Zgusta 1971), consisting of three phases: planning, compilation and publication. Computers (especially the automatic processing of language data), the Internet, and the available quantities of linguistic and related data have undoubtedly affected the way dictionary content is compiled and published. According to Klosa (2013: 4), the lexicographical process in the compilation of non-static online dictionaries consists of six phases, which are not sequential, but can overlap or complement each other (Klosa 2013; Tiberius and Schoonheim 2015). These phases are: preparation, data acquisition, computerization, data processing, data analysis, and preparation for online release.

2.1   The lexicographical process in the proposals for a new dictionary of Slovene

Recently, Slovenian lexicographers have started discussing the need for a dictionary of modern Slovene, but it was not until the publication of the only publicly presented proposal for the compilation of DMSL (Krek et al. 2013b) that such discussions became more concrete. As the lexicographical process is heavily dependent on the dictionary content, methods and main medium for which the dictionary is designed, it represents a key element in the overall concept of a dictionary and how it will be realized.

Before focussing on individual phases in the compilation of DMSL, as proposed by the consortium lead by the Centre for Language Resources and Technologies at the University of Ljubljana,[4] we will present two other related projects: the *New Dictionary of Slovene Literary Language* (NDSLL), the compilation of which is expected to take at least 20 years,[5] and the *Monitor Dictionary of the Slovene Language* (MDSL), which is closely linked with NDSLL and could be seen as a form of dictionary under construction (Klosa 2013: 3),[6] a novelty in Slovenian lexicography.

---

4   http://www.cjvt.si/projekti/

5   See the responses in media to the Proposal of DMSL, e.g. http://www.24ur.com/novice/slovenija/na-nov-slovar-slovenskega-knjiznega-jezika-bomo-cakali-se-leta.html.

6   Perhaps a more suitable term would be "a never completed dictionary".

2.1.1 Monitor Dictionary of the Slovene Language

MDSL is described as a growing dictionary, and one that is only informative in nature during its compilation.[7] Its initial version contains words which are not found in existing dictionaries of Slovene. but can be found in corpora of Slovene. Also added to the headword list are words that have been unsuccessfully searched for by the users at the website http://bos.zrc-sazu.si, as well as words not found in existing corpora of Slovene, but attested in other, and especially electronic, resources.[8] In the introduction section of MDSL it is also stated that a similar approach will be used for updating the headword list in the future, and that new words will be added to the dictionary every six months. Although the "initial version" is mentioned, the authors do not give any details about how older versions of the dictionary will be archived, or even if is archiving envisaged. The relationship between MDSL and NDSLL is also unclear: "Only time will tell whether individual entries will end up in normative or explanatory dictionaries." (Introduction, MDSL). So despite the ambition indicated by its name, it can be concluded that methodologically, i.e. in terms of the gradual adding of dictionary content, the dictionary does not actually bring any novel lexicographic approach to Slovenian lexicography. The entries are compiled from scratch, and access to different versions is not provided.

2.1.2 The NDSLL concept

An overview of the compilation of NDSLL needs to be made, mainly because the editors claim that the procedure will include three important processes in the pre-editing phase that are nearly identical to the processes envisaged in the compilation of DMSL (Krek et al. 2013b). These are: (a) automatic extraction of data from the corpus, (b) development of a tool for detecting changes in meanings and grammar, and (c) upgrading of existing corpora of Slovene.

The compilation of NDSLL is divided into the pre-editing and editing phases. The pre-editing phase includes the preparation of the headword list, which will serve as a basis for the selection of dictionary entries. The dictionary authors anticipate that corpus data and data from existing dictionaries[9] and other language resources of the Fran Ramovš Institute for the Slovenian Language ZRC

---

7   The author and expert consultants claim that the words are selected and described purely from the informative perspective; no normative information is included.

8   The authors do not provide any details about these resources.

9   The authors claim that the information from existing dictionaries will be included as much as possible, which casts doubt on their stated intention of compiling a dictionary from scratch (NDSLL: 1, 2).

SAZU[10] will be automatically added to the entries in the dictionary database. Among the listed automatically extracted information are headword spelling, word class, frequency of the lemma and individual word forms, syntactic information, including collocations and examples, certain grammar labels, and certain information on language use, such as treatment of hyphenated words as one word or a multi-word unit. Further analysis of all this information will direct the dictionary treatment of individual headwords. The automatic extraction of the such information from the corpus demands exact decisions about the interpretation of data in terms of the relationships among corpus, lexicon and dictionary, as lemmatisation and morphosyntactic information are closely linked to corpus tagging, which means it is not possible to transfer this information directly into the dictionary entries. The experience from the SLD project shows that this process is by no means trivial. Considering that, at least to the best of our knowledge, these procedures have not yet been tested by the NDSLL team, any evaluation or a detailed presentation of the automatic extraction process cannot be expected. If the authors of NDSLL have decided to use the same methodology for the automatic extraction of data as was applied in the compilation of the SLD, it is then important to stress that the procedure in the SLD project followed very clear methodological guidelines and was adapted to the organisation of dictionary information, which differs in many aspects from the organisation of dictionary information as presented in the NDSLL concept.

The NDSLL concept also envisages the development of a tool that would detect semantic and grammar changes in language use (NDSLL: 78). According to the authors, this would shorten the time of dictionary compilation, as the editors would have the information prepared in advance in the dictionary-writing system (DWS), but in contrast to the purpose of automation (cf. Kosem et al. 2013a) it is anticipated that the editors will check all the information as if they had been designing the entries from scratch (NDSLL: 78).

One of the parts of dictionary compilation is making regular updates to the corpus, a far from a trivial task. There is however no detailed explanation provided on how existing corpora will be updated, how the taxonomy of corpus texts will be upgraded or adapted, how the permission for the new texts will be obtained, and so on, and thus an explanation similar to that provided by Logar (2015) on the updating of the Gigafida corpus is needed. At present, can only find the statement that all the changes made to previous versions of the dictionary will be documented and that, for reference purposes, the users will also have access to older versions of the entries (NDSLL: footnotes 4 and 32).

---

10 Based on this list of automatically extracted data we can assume that the team will use the same procedure as was used in the SLD project (Kosem et al. 2013; Kosem et al. 2013a). However, no references are provided in the outline of the NDSLL concept.

2.2 Lexicographic process in the compilation of DMSL

Separate phases in the compilation of DMSL have been first presented in the proposal (Krek et al. 2013b: 52), and in this paper we discuss them in more detail, also according to the experience gained during upgrading of the process of automatic extraction of corpus data and its evaluation (Kosem et al. 2016b). In addition, we consider the experience gained from projects involving the compilation of dictionaries conceived primarily for the online medium and for gradual release, meaning that the dictionary information is available to the users during its preparation. Furthermore, regular updates to completed dictionary entries are anticipated. These are therefore dictionaries which are constantly being compiled and are called *online dictionaries under construction* (Klosa 2013) in the lexicographic literature.

The compilation of DMSL is expected to include five phases (Figure 1), which are designed in a way that enables entry release during the compilation process, i.e. after the first phase. The advantages of this approach outweigh its complexity, shown in the fact that individual phases need to be specified in great detail and the tasks of lexicographers and other team members should be well-defined and coordinated. Multi-phase compilation of entries also enables a more efficient and economical division of work. Most of the work in the first phase is done by a computer, and human input starts in the subsequent phases where lexicographers are used for specific tasks which require lexicographic knowledge and experience. The division of headwords according into the difficulty levels also enables the training of less experienced lexicographers in sense division and definition writing. It is envisaged that certain routine tasks, such as identification and removal of incorrect or irrelevant data and the distribution of collocates and examples under relevant senses, will be left to crowdsourcing (see Fišer et al. 2015), thus reducing the cost of human resources and, most importantly, speeding up the compilation of entries.

During the multi-phase dictionary compilation process, where the availability of different versions of entries is planned, it is of utmost importance that the phase of releasing a dictionary entry is clearly signalled to users. For this reason, we plan to display a different symbol for each phase along with the date of entry release, which reflects the date of the last changes to the dictionary entry. On the one hand, this provides a reference for each entry that the users can use (this is particularly important for researchers and teachers), and on the other it gives some indication to the users as to what they can expect in terms of quantity, treatment and reliability of dictionary content.
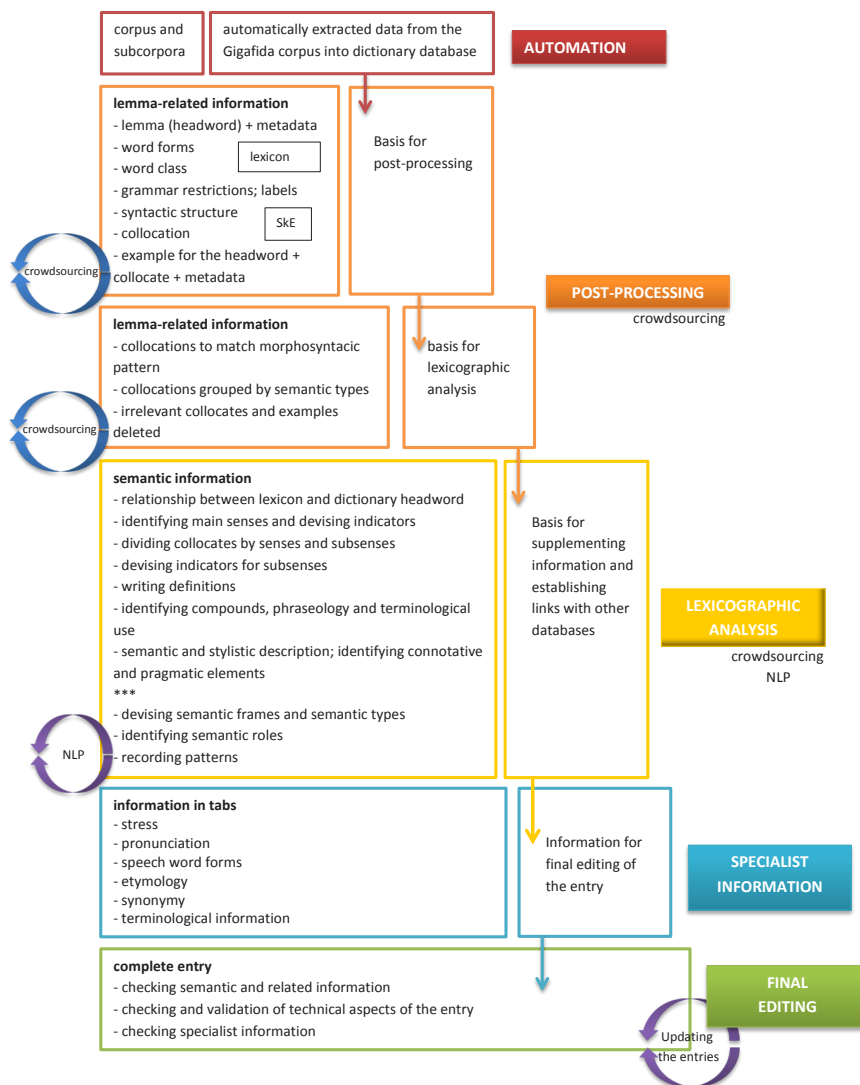
Figure 1: Phases of the DMSL lexicographic process

2.2.1 Phase 1: automatic extraction

Phase 1 in entry compilation consists of automatic extraction of lexical data from the Gigafida corpus (Logar Berginc et al. 2012), which will also be used when devising the headword list. In addition to using the frequency information from Gigafida, the headword list will, based on various statistical calculations, utilise

information from the Kres corpus, Gos corpus (Verdonik and Zwitter 2011) and other freely available corpora of Slovene. In order to obtain a more specialised vocabulary, the compilation of specialised corpora is envisaged, as well as updates to existing corpora, e.g. the compilation of a subcorpus of textbooks (see Logar 2015; Vintar and Logar 2015 for more). Furthermore, we plan to use thematic tagging of domain texts in the form of corpus metatags, which are then transferred into the dictionary database during automatic extraction (see Gantar and Kosem 2013; Kosem 2015).

Taking into account the structure of a dictionary entry in DMSL (see Klemenc et al. 2015), the automatically extracted data are as follows:

- **Lemma** in the basic form, as found in the Gigafida corpus and the Sloleks lexicon of Slovene word forms (Dobrovoljc et al. 2015), and all of its word forms (offered in a separate tab).

- Corpus or sub corpus **frequency** of the lemma.

- Word class, based on the word form tag in Gigafida and Sloleks.

- Certain **grammatical alerts** related to typical syntactic or contextual behaviour of the lemma in the corpus, such as frequent use with proper names, predominant use in third person or when citing (verbs). This information is extracted from the corpus using a combination of the directives CONSTRUCTION and UNARY in the Sketch Engine tool (Kilgarriff et al. 2004) and is presented in the dictionary database in the form of alerts, which can be later (in Phase 3) converted into dictionary labels such as *pogosto zanikano* ('often in negative'), *pogosto v 3. os. ednine* ('often in 3rd person singular') etc. (see Kosem 2015).

- **Syntactic structures,** identified during manual analysis of word sketches for the purposes of SLD and used as a basis for a new, improved version of sketch grammar for Slovene (Krek 2012a).

This new sketch grammar utilizes the directives *CONSTRUCTION, *COLLOC and *SEPARATEPAGE. The first enables the identification of grammatical relations without collocates, which is particularly useful for extraction of corpus examples containing all elements in verb patterns, such as "subject-predicate-indirect_object-direct_object", confirming the existence of the pattern for the particular verbal headword. The second directive is used to identify elements that are categorized as syntactic combinations in the lexical database, such as the statistically significant "preposition-noun-preposition" combinations. The third directive is intended for creating a separate word sketch page for relations with three elements (directive *TRINARY), which enables the introduction of relations with prepositions that can have more specific definitions: for example, they produce a separate

word sketch for each noun case (of the six cases in Slovene) in "noun-preposition-verb" or "noun-preposition-noun" patterns. This new sketch grammar for Slovene thus provides a very fine-grained overview of a word's collocational behaviour and is devised solely for automatic extraction of lexical data. The word sketches produced by such a sketch grammar are difficult to process by a human user, due to the high number of relations and their complex naming system.

- **Collocates** found with the lemma in a particular syntactic structure and forming potential collocations, syntactic combinations and compounds. The latter are identified and recorded under the relevant sense by lexicographers in Phase 3.

- **Corpus sentences** containing the lemma and collocate in a particular syntactic structure. Corpus sentences are extracted with the GDEX tool (Kilgarriff et al. 2008), with configurations based on the GDEX for Slovene (Kosem et al. 2011) but especially adapted for automatic extraction (Kosem et al. 2013b). The extracted sentences are candidates for inclusion in the dictionary (they may require minor modifications), and are thus potential dictionary examples.

The automatic extraction procedure has already been tested in the compilation of the SLD (Kosem idr. 2013, 2013a), where an API script using different parameters for each grammatical relation was used to automatically extract the above listed types of data, which were then imported into the dictionary database in the iLex DWS (Erlandsen 2004). We also conducted an evaluation of the procedure, comparing it with the manual entry compilation (Kosem et al. 2015), and later upgraded and improved the procedure for the use with the *Collocation Dictionary of the Slovene Language* (Gantar et al. 2015). Among the most notable upgrades are automatic removal of collocates that have all the same examples, and automatically converting the lemma and/or the collocation in the word form with appropriate case, gender and number according to the syntactic structure. In addition, the initial values used for extraction were improved considering the frequency and word class of the lemma (see Kosem et al. 2013a; 2013b for more), resulting in the development of several parameter configurations for each word class. In the improved procedure we also extracted collocates using salience and frequency order, respectively, and then combined both sets of data. This enables us to select the most relevant collocates for each lemma.

2.2.2 Phase 2: post-processing and clean-up

Phase 2 is intended for (a) post-processing, which includes (semi-)automatic removal of errors and irrelevant data, also by using crowdsourcing, and (b) adding of

metatags which enable the connecting of information in the dictionary database and establishing links with other dictionary databases (e.g. the initial dictionary database, collocations dictionary database, and synonym dictionary database). Automatically extracted data can be additionally improved with post-processing, e.g. by putting collocates in the appropriate gender and case according to the syntactic structure, and by forming collocation sets which include semantically related collocates. In order to facilitate combining the information from different databases, it is necessary to tag different elements within collocations (e.g. prepositions, conjunctions, and reflexive pronouns of verbs) and/or add relevant linking information in the tag attributes of the lemma or its collocates (e.g. ID from Sloleks).

We envisage the use of crowdsourcing to remove irrelevant collocations, which are the consequence of errors in lemmatisation or are simply corpus noise. Figure 2 shows an example of a task in which the crowdsourcers are asked to decide whether the combination in the automatically extracted sentence (coloured in blue and red in the example; *gre za franšizo*) reflects the identified syntactic structure:

Figure 2: The crowdsourcing task for removing irrelevant collocations and examples in the SLD.

The crowdsourcing tasks were conducted with the SlowCrowd tool (Tavčar et al. 2012), which has also proved useful in the improvement of the Slovene version of wordnet called SloWNet (Fišer 2009). The initial tests during the compilation of the SLD (Kosem et al. 2013b) have shown that the use of crowdsourcing for data clean-up is reliable, and can considerably reduce the time spent on this phase of the lexicographic process.

2.2.3 Phase 3: lexicographic analysis

The next phase involves the lexicographic analysis of data, which makes it most demanding in terms of expertise and logistics, and it also takes the most time. Tasks include sense division, definition writing, identification of grammatical, syntactic, normative and stylistic characteristics of words and their meanings.

In this phase, the lexicographers are presented with cleaned-up automatically extracted data for each lemma. Word class information is automatically attributed to the lemma, and is the same as the morphosyntactic tag in the Sloleks lexicon. Consequently, lexicographers' first task is to check the correspondence between the lexicon unit and the dictionary entry. This is by no means easy, and the efficiency and congruity of lexicographers' decisions relies on providing them with detailed instructions containing all possible situations and common solutions, especially in terms of homonymy and conversion, i.e. in accordance with the decisions outlined in the dictionary concept (see Gantar 2015 and Dobrovoljc 2015). In DMSL, there is a symmetric relationship between the entry in the lexicon and in the dictionary, while potential exceptions are signalled in the database using a predetermined set of machine-readable restrictions.

The main tasks of lexicographers in this phase are identifying senses and subsenses, and writing definitions for priority entries. The entry structure follows that of the entries in SLD, which means that the lexicographers also devise sense indicators that represent a constituent part of a sense menu, which offers an overview of the entry senses and subsenses, and, at certain noun and adjective entries, they must also devise semantic frames that contain a typical valency pattern of a particular (sub)sense. Also important in this phase is adding information intended for natural language processing, for example sentence patterns, semantic types (similar to the approach used in Corpus Pattern Analysis; Hanks 2004; Hanks and Pustejovsky 2005) and semantic roles. The lexicographers also identify and write definitions for compounds – marking those that require an input from the terminologists – and phraseological units.

Lexicographic work is organised according to the difficulty level of the entry and the availability of templates for semantically related entries. To achieve the optimal efficiency, the tasks are divided between (a) experienced lexicographers who perform sense division and write definitions, identify more complex grammatical and syntactic patterns, and record any stylistic and pragmatic information about the word's usage; (b) lexicographers specialised in phraseology, description grammatical and syntactic characteristics of individual (sub)senses, and normative information; and (c) relatively inexperienced lexicographers who conduct less demanding lexicographic tasks, such as checking whether collocates have been correctly assigned to

senses and syntactic combinations during crowdsourcing, forming collocation sets, and identifying context for compounds and phraseological units.

Lexicographic tasks that can be regarded as routine in nature and do not require considerable lexicographic knowledge will be left to crowdsourcing. One such task is assigning automatically extracted corpus sentences (and the collocation in a particular syntactic structure they attest) to one of the (sub)senses.[11] The secondary goal of the task is to get feedback on the suitability of the sense division and to identify unidentified or new senses.

At the end of Phase 3 the majority of relevant semantic information is already available to the users. The next step consists of adding information which will be presented in the online dictionary separately (e.g. in tabs), and this is done in Phase 4.

2.2.4 Phase 4: adding specialist language information

Phase 4 of the lexicographic process comprises of adding information from other databases and enriching existing entry information. This phase requires the involvement of experts from other fields, especially terminologists, and linguists with expertise in standardisation and norms. The following information is added to the entries at this point:

- information on **spelling**, based on the Sloleks lexicon and according to the connection between the lexicon entry and the dictionary entry. This includes detecting the overlap, or lack of it, between the pronunciation and declension paradigms of the headword, which is one of the criteria for detecting homonymous lemmas;

- information on **pronunciation** based on the Gos corpus and predetermined procedures, including marking the stress, pronunciation of word forms and providing pronunciation that cannot be deducted from the headword's spelling (Jurgec 2015);

- information on speech word forms and any special semantic characteristics observed in the speech corpus Gos (Verdonik 2015);

- information on **etymology**, more specifically on the origin of the word and its related word forms in different languages, and information about archaic forms of the word in and the time period in which they appeared;

- information on **synonyms**, obtained using the Sketch Diff feature in the Sketch Engine tool and information from SloWNet, and

---

11 This task is described in more detail in Fišer et al. (2015).

- **terminological analysis of data.** For this task, we will form a network of experts in different fields and develop an online platform that will facilitate the monitoring and coordination of work.

2.2.5 Phase 5: final editing

The final phase in the compilation of DMSL is intended for final editing of the entry and a consistency check of the information found in different tabs. Here the lexicographer's task is to check the consistency of information with the dictionary concept and real language use, and to check the validity of entry structure. The lexicographer has the option to edit, expand or even return the entry to one of the previous phases if they, for example, identify inconsistencies in sense division, or in compound or phraseology treatment, or find incomplete terminological information.

Another important step of this phase is automatic detection of semantic changes in a word's usage which can, if identified, return the entry to Phase 3 (sense division, multi-word unit identification, crowdsourcing, and so on), thus requiring another final editing at a later stage. So, even though Phase 5 represents the end of lexicographic process, entry compilation again involves automatic extraction, in this case of relevant new data, found in updated corpora.

3 UPDATING THE DICTIONARY DATABASE

The dictionary database plays a very important role in the process of entry compilation and their eventual presentation to users, representing the source of all dictionary information on one hand, and the archive of all the decisions made during the five phases of the lexicographic process on the other. Considering that the phases of dictionary compilation are clearly delineated, the dictionary database needs to include a workflow that can at any time provide the editors and lexicographers with the information on the phase status of each entry. Another level of complexity in planning the dictionary database is introduced by two connected decisions: regular updating of the dictionary and the option of uploading the entries after each phase is completed.

By regular updating of the dictionary we mean updates to the existing entries in the database that have already gone through all the phases of the lexicographic process, as well as the compilation of completely new entries, especially priority ones (e.g. neologisms). The latter need to have in the dictionary database a special

warning about their importance, distinguishing them from other entries, which in turn provides the editors with the option to alert users to such entries once the dictionary is updated. This is also true for updates to already completed entries, which can be made in the form of adding new information (e.g. senses, collocations, and phraseological units) based on the analysis of new data (e.g. monitor corpus or new version of the reference corpus) or in the form of modifying existing information (e.g. correcting errors). In this case, more important for the users is the temporal aspect, i.e. when was the new information added (and based on which resource).

A special case in regular updating of the dictionary is replacing old with new information, but only in terms of presentation to the users. For example, at some point the decision could be made to replace existing examples with new ones (cf. Klein and Geyken 2010; Lemnitzer et al. 2015). To enable this, examples (and other microstructural elements) in the database need to include the information on whether they are part of the online dictionary entry or not. This makes it possible to keep all the examples in the database while showing the users only those that are most relevant.

Releasing entries after each phase does not require any additional information, except that related to the workflow; existing dictionary entries that are being updated with new information should be excluded from this procedure, as the combination of analysed and non-analysed data could confuse users. More relevant for such a procedure are visualisation solutions which also require certain types of information (e.g. date of release and version).

It is therefore essential to prepare a procedure that gives the lexicographers a clear idea about the phase of the entry, date of its inclusion in the online dictionary, and the date of addition of new information (the completion date of the new version[12]). We believe that such a procedure can ensure an efficient and transparent lexicographical process, and facilitate clear and understandable presentation of dictionary content to users.

4 TREATMENT OF DIFFERENT VERSIONS AND PRESENTATION

This section is dedicated to three questions relevant for updating the dictionary using the proposed lexicographic process: How often should the updates be

---

12 It is important to distinguish between entry versions denoting larger changes (e.g. after the completion of each phase or after updating existing dictionary entries with new information), and entry versions in the DWS. Namely, the DWS records every single change made to the entry, thus enabling the comparison of two database "versions", reviewing and restoring of deleted data, etc. It is thus recommended to consistently use terminology that distinguishes between the two processes.

made? How to clearly distinguish between incomplete entries or incomplete entry information from the completed ones? How to handle different versions of dictionary entries and different versions of the dictionary?

Dictionaries of other languages use two different approaches to updating online dictionaries: updates in regular intervals (usually every few months) or continuous updates, as soon as new entries are compiled. The former approach is used by dictionaries such as the *Oxford English Dictionary* (OED),[13] which releases updates every four months and also has a special webpage dedicated to promoting the updated entries. A similar approach is used by the *Macmillan English Dictionary* (MED),[14] although instead of separately providing the information on the date of the updates,[15] the MED alerts the users to selected new entries in the New Words section on the front page.

The latter approach, i.e. immediate release of completed entries, is used by the *Comprehensive Dictionary of Polish*[16] (Žmigrodzki 2014) and the *Dictionary of Contemporary Dutch*[17] (Tiberius and Schoonheim 2015). It is noteworthy that these two dictionaries are being made from scratch, and are thus real dictionaries under construction. Consequently, the motivation for continuous release of dictionary content is much higher, in terms of satisfying both the users and funders, than at dictionaries that merely add new words or update existing entries. In view of this, the proposal in the NDSLL concept to update the dictionary annually (NDSLL: 3) is not ambitious enough, and fails to fully consider needs of the users. As such, we would expect that the users, who have been waiting for a new description of Slovene for over 25 years, will be offered the results of lexicographic work as soon as possible.

The approach of immediate release is also part of the lexicographic process in the proposed DMSL, where it is envisaged that entries will be released after each phase of their compilation. There are already dictionaries in Slovenia using this approach, such as iSlovar,[18] a dictionary of computer terms, which distinguishes between four phases of entry compilation: "predlog" ('proposal'; proposed by the editor or user), "pregledano" ('reviewed'; reviewed by the editor), "strokovno pregledano" ('reviewed by expert'; reviewed and edited by experts) and "urejeno" ('edited'; reviewed by the dictionary team; this is the final editing). It should be pointed out that the updates are not made every day or even every hour, but are made as packages (i.e. several entries at the same time) in frequent intervals, resulting in greater transparency for both lexicographers and users.

---

13 http://www.oed.com/
14 http://www.macmillandictionary.com/
15 The website with FAQ (http://www.macmillandictionary.com/faq.html provides the information that the dictionary is updated several times a year.
16 Wielki słownik języka polskiego: http://www.wsjp.pl/
17 Algemeen Nederlands Woordenboek: http://anw.inl.nl/show?page=search1.
18 http://www.islovar.org

The alerts about the entry status serve to distinguish between incomplete and complete entries. The proposal for the compilation of DMSL (Krek et al. 2013b: 52–60) suggests that the entry status is indicated with coloured dots (from red to green) and the date of last update (Krek et al. 2013b: 27). The date can be used to distinguish between different versions of the entry at the same phase (e.g. when updating the completed dictionary entry). The final visualisation solution may end up being different from the one in the proposal, but will need to include at least these two types of information. In addition, there will be, like on the OED website, a webpage dedicated to the updates, offering a list of new and updated entries and their status, as well as information on any major changes.

One of the important decisions related to the immediate release of entries concerns the treatment of previous versions. This is somewhat less problematic as far as releasing entries at different phases is concerned, as the version most relevant for the users is the one that contains the largest amount of information. This was in fact one of the hot topics at the OED Symposium in 2014, as a few participants complained about that after updates they could no longer see the previous versions of entries. Their argument was that by updating the definitions we lose information on how a certain sense or usage of the word was perceived by lexicographers working on the entry at a particular point in the past. While this argument is perfectly legitimate, it is worth pointing out that the OED is a historical dictionary in which a diachronic view of language use is of vital importance.

The decision on whether to include the option of comparing different versions of the dictionary entry in DMSL will be based on the findings of surveys among dictionary users. Nonetheless, access to previous versions of entries is already envisaged for the researchers working in the fields of linguistics, machine learning and natural language processing. Namely, we plan to make freely available new versions of the dictionary database, which will be released simultaneously with updates to the online dictionary, except in cases when the changes made will be relevant only to the dictionary database. An important part of this process will be detailed documentation, which will include not only a description of changes to the dictionary entries, but also a description of all the content and technical changes to the dictionary database, e.g. new types of database labels, changes in DTD (*Document Type Definition*) and so on.

5   CONCLUSION

The lexicographic process of dictionary compilation that envisages the continuous release of entries at different phases of compilation, regular updates and access to different versions of the entries, is a complex procedure, demanding a

well-designed and detailed strategy which affects the organisation of lexicographic work. The lexicographic process proposed in this paper is based on the automatic extraction of lexical data, which are in the subsequent phases first cleaned of incorrect and irrelevant data, and then analysed and supplemented with additional information. A distinction should be made between the information in the dictionary database which is part of the regular work on the entries and is not presented to users, and the database information in different versions of the entries offered to users. It is also vital that lexicographical process is clearly defined and recorded, enabling the lexicographers and editors to carry out continuous and consistent dictionary work. This also makes it possible to provide the users with information on the status of the entry, the entry information that is available at each phase, and with different versions of the entries, the latter being particularly relevant for the purposes of research, further processing or teaching. It is important to note that the proposed lexicographic process is devised for an online dictionary, using a specific entry structure and internal organisation of information, and containing different types of lexico-grammatical information linked both within the entry (e.g. sense menu, collocations, syntactic structures, patterns, examples under senses, collocations, compounds and phraseological units) and with the information in other tabs (word forms, speech, norm, synonyms etc.).

Dictionary examples

Iztok Kosem

Abstract

In this paper, the role of examples in dictionary entries is presented, and an overview provided of relevant studies into the use and usefulness of examples. We put forward the different ways of presenting examples in general monolingual dictionaries, list the characteristics of a good dictionary example, and discuss the different methods of finding good examples. The focus then turns to the role and characteristics of examples in the proposal for a dictionary of modern Slovene, the methods for their extraction, and the procedures to be followed for saving examples to the dictionary database and archiving them, before concluding with the different visualisation options for the (online) dictionary.

Keywords: dictionary examples, good examples, automatic extraction, visualisation, dictionary database

1   INTRODUCTION

Examples are one of the most important parts of a dictionary entry, as they are used for exemplifying the use of words, collocations, compounds, phraseology and so on in context, i.e. in real language. Putting the words back into context is vital for a dictionary, since the majority of dictionary content is decontextualized.

The paper first describes the role of examples in a dictionary, and makes an overview of research into the usefulness of dictionary examples. This is followed by the presentation of different ways of example presentation in monolingual dictionaries of Slovene and other languages. Next, the characteristics of good dictionary examples are presented and different methods for finding them are described. The paper focuses on the role and characteristics of examples in the proposed *Dictionary of Modern Slovene Language* (DMSL), their acquisition from corpora and ways of recording them in the dictionary database. Visualization of examples in the dictionary is also briefly discussed. The conclusion summarizes the main points of the paper and considers the future role of examples in dictionaries and related resources.

2   THE ROLE OF EXAMPLES IN A DICTIONARY

The role of examples concerns two aspects of dictionary use: receptive and productive. The receptive aspect, which examples are primarily intended for, is to supplement definitions, which is why examples first and foremost need to contain information related to the meaning they attest. As argued by Atkins and Rundell (2008: 454), it is sometimes difficult for the user to understand the definition without reading the examples. Examples can also be useful when navigating through (long) entries, as the users can "identify the particular sense they are seeking by finding examples that are similar to the one they need or have in front of them" (Fox 1987: 137).

The productive role of dictionary examples is to attest the syntactic patterns, valency, collocations and other characteristics of the headword (Húmble 2001), which are supposed to help the users when writing or, less often, speaking. Examples intended for production are found mainly in dictionaries for L2 learners, e.g. advanced learners' dictionaries or dictionaries for younger native speakers, such as school dictionaries.

Studies into dictionary examples have mainly focused on their value for language production of non-native speakers. The most commonly used research method involves asking the subjects to use (unknown) words in a sentence, and consulting dictionaries or selected dictionary entries in the process. The subjects are grouped into those that are provided only with definitions, and those that are provided with definitions and examples; some studies (e.g. Frankenberg-Garcia 2012; 2014) also

include a group of subjects that are provided only with examples. The findings of the majority of studies (Summers 1988; Laufer 1993; Nesi 1996; Al-Ajmi 2008) are not very encouraging, as they show that examples do not have considerable added value for the encoding needs of the users. However, as Frankenberg-Garcia (2012) pointed out, the aforementioned studies have two key methodological shortcomings: firstly, despite studying the productive value of examples, the studies contain tasks in which the subjects need to first decipher the meaning of an unknown word and then use that word in a sentence. This means that the tasks include both receptive and productive dictionary use, which is a rare form of dictionary use. Secondly, using unknown words for testing productive use does not reflect actual dictionary use and language production in general, as people rarely use completely new words when writing (Laufer 1993: 138),

Frankenberg-Garcia (2012; 2014) improved the methodology of previous studies by clearly distinguishing between testing the receptive and productive roles of dictionary examples, and also by using examples for reception and examples for production, respectively. The subjects were divided into four groups: the control group (without a dictionary), the group that was provided only with definitions, the group that was provided with one corpus example, and the group that was provided with several corpus examples. Her findings were that several corpus examples are almost equally valuable as the definition when trying to understand the meaning of a word, and that for encoding use several corpus examples are much more useful than one example, while in general examples are much more useful than definitions.

There are very few studies that research how frequently the users consult examples. In a study that involved his students, Béjoint (1981) found that they consulted examples quite frequently. Similar are findings of Kosem's study among 620 students (449 native speakers and 171 non-native speakers of English) at Aston University; examples were the fourth most frequently consulted part of the dictionary entry (after definitions, pronunciation and synonyms), and, when considering only non-native speakers, examples were the second most frequently consulted part of the entry (after the definitions).

3 EXAMPLES IN GENERAL MONOLINGUAL DICTIONARIES

An analysis of the treatment and form of examples in general monolingual dictionaries[1] shows three different groups of dictionaries. The first includes those that offer

---

1 The analysis included only online dictionaries. The dictionaries can have paper versions or were originally published in the paper format, but the list also includes dictionaries that exist only online (e.g. the *Comprehensive Dictionary of Polish* and *Comprehensive Dictionary of Dutch*). A detailed analysis of the treatment of examples in dictionaries of Slovene is provided after the description of all three groups of dictionaries.

examples mainly in the form of partial sentences (or phrases) and occasionally also as whole sentences (e.g. the Spanish monolingual dictionary). Some dictionaries limit the use of examples only to certain (sub)senses or phrases. The information on the source of the example is rarely provided (there are exceptions, such as the *Explanatory Dictionary of Estonian*). Such treatment of examples is often found in dictionaries that are not corpus-based, and were originally conceived for print and later transferred to the online format. A few recently published dictionaries have also adopted such treatment, mainly those that were conceptualised according to the lexicographic approaches of the 20th century. The group includes dictionaries such as the *Dictionary of Literary Czech*[2] (DLC; Slovník spisovného jazyka českého, 1989), *Royal Spanish Academy Dictionary of Spanish*[3] (RSADS; Diccionario de la lengua Española de la Real Academia Española, 2014), *Explanatory Dictionary of Estonian*[4] (EDE; Eesti keele seletav sõnaraamat, 2007) and the *Croatian Encyclopaedic Dictionary*[5] (CED; Hrvatski enciklopedijski rječnik, 2003).

In the second group are dictionaries that offer mainly whole-sentence (corpus) examples, examples in the form of partial sentences are rare or not used at all. This treatment of examples can be found in English dictionaries published by Oxford (Oxford Dictionaries;[6] ODE), Macmillan (*Macmillan English Dictionary*; MED[7]) and Merriam-Webster (*The Merriam-Webster Online Dictionary*; MWOD[8]), the *Dictionary of Contemporary Danish* (Den Danske Ordbog;[9] DDO), the *Comprehensive Dictionary of Polish*[10] (CDP; Wielki Słownik języka Polskiego) and the *Comprehensive Dictionary of Dutch*[11] (CDD; Algemeen Nederlands Woordenboek). Nonetheless, dictionaries differ in the manner they present the examples. In MED and DCD, whole-sentence examples are presented within the entry, under each sense, subsense, phrase and so on. MWOD and ODE use both whole-sentence examples and excerpts, but clearly distinguish between them in terms of their presentation in the entry. Excerpts are offered under senses and subsenses at the first level, so immediately upon opening the entry, whereas whole-sentence examples (for all senses) are provided together at the end of the entry (MWOD) or available under senses by clicking on "More example sentences" (ODE). A somewhat less prominent role is given to examples by CDP and CDD, where these are not shown upon opening the entry and are only available on a click (CDD) or in a separate tab (CDP). These two dictionaries also provide the information on the source of the example.

---

2   http://ssjc.ujc.cas.cz (the online version available since 2011).

3   http://lema.rae.es/drae

4   http://en.eki.ee/dict/ekss

5   Accessible through the Croatian Dictionary Portal http://hjp.novi-liber.hr.

6   http://www.oxforddictionaries.com/

7   http://www.macmillandictionary.com/

8   http://www.merriam-webster.com

9   http://ordnet.dk/ddo

10 http://wsjp.pl

11 http://anw.inl.nl

The third group includes portals such as German DWDS[12] (Das Digitale Wörterbuch der deutschen Sprache) that offer on one page the information from dictionaries, corpora and other relevant resources.[13] The most important characteristic of this group is the link between dictionaries and corpora, with corpora being a source of an abundant number of examples, especially considering Frankenberg-Garcia's findings about the benefits of multiple examples for dictionary users. A shortcoming of such portals is in the large amount of information they provide, which often makes it difficult for the users to interpret and correctly use them.[14]

As far as dictionaries of Slovene are concerned, the *Dictionary of Slovene Literary Language* (DSLL) and its successor DSLL2 belong to the first group of dictionaries, offering examples as excerpts. The excerpts were taken from texts or were in some cases invented.[15] At least for DSLL, this finding is not surprising, given that the dictionary was made before the corpus lexicography era. However, DSLL contains a considerable quantity of examples, much more than comparable dictionaries of other languages, including recently published ones (e.g. RSADS). Examples were one of the most heavily affected parts of dictionary entries during the preparation of DSLL2, as the examples from DSLL were modified or replaced due to social changes, or completely new examples were added. As noted by Krek (2014: 146), however, changes in existing examples are often not appropriate or necessary, or completely new examples do not bring any added-value to the user's understanding of the meaning of the word. Moreover, replacing or changing existing examples in the preparation of DSLL2 seems unnecessary, considering that the authors are presenting the dictionary as a resource that reflects 150 years of the Slovenian language.[16] This is confirmed by Krek (ibid.: 147), concluding that this approach erased a great deal of evidence on the usage of words before 1991.

The *Dictionary of New Words of the Slovenian Language* (2012; DNWSL) was published even before DSLL2, and its authors to some extent used state-of-the-art lexicographic methods and included (whole-sentence) corpus examples, in addition to excerpts. As stated in the Introduction (DNWSL: 9), the main resource in the compilation of the dictionary was Nova beseda, a 318-million-word corpus of Slovene: [17]

> Based on authentic usage, attested in the 300-million-word Nova beseda corpus, 5,384 dictionary entries consist of 6,512 senses and subsenses of newer words and multi-word units, coming from different domains.

---

12 http://www.dwds.de/

13 Other dictionaries, e.g. DCD, offer access to a corpus on their website, however they do not offer a simultaneous search in all the resources and aggregated display of hits.

14 DWDS does offer the option of limiting the hits to only selected sources.

15 As written in the Introduction to DSLL (1991: XXII), "[w]henever the texts didn't contain enough information, the excerpts were either taken from other resources or invented".

16 Marko Snoj 2nd November 2013 for STA: http://www.rtvslo.si/kultura/knjige/akademska-vojna-okrog-novega-slovarja/321592.

17 http://bos.zrc-sazu.si/s_beseda3.html

A close examination of the examples in DNWSL reveals that the absence of (good) examples in Nova beseda sometimes forced the lexicographers to obtain them from other corpora, especially from 1.2-billion-word Gigafida corpus. Although this may not be problematic, it does bring into question the above cited methodology of headword list compilation, especially at entries such as *bandži skok* ('bungee jump'):

> **bándži skòk --** skôka in skóka m (ô, ȍ ó; ọ)
>
> skok v globino, pri katerem je skakalec pripet z dolgo elastično vrvjo; skok z elastiko: Obnaša se kot frkolin, ki se pred tovarišijo postavi z bandži skokom, ko se privezan na elastično vrv vrže z mostu v globel **E ↑bungee (jumping) in (↑)skòk**

The example provided above is a (slightly) modified sentence from the Gigafida corpus. It is noteworthy that the Nova beseda corpus does not contain a single hit for *bandži skok* (even Gigafida has only five). The example is thus attesting the use of a word for which we do not even know how it got into the dictionary. In addition, the dictionary's focus on newer words, which tend to have lower frequency in corpora, means that examples are used merely for attestation purposes, as they do not bring any added value to the understanding of the meaning.

A more systematic and corpus-driven approach has been used in the compilation of the Slovene Lexical Database (Gantar et al. 2012; SLD). The SLD contains 2,500 entries with 152,996 examples, so on average over 61 examples per entry. All the examples are whole sentences and were taken from the Gigafida corpus (Logar Berginc et al. 2012). The examples in the SLD have not been modified in any way, as the selection of examples for a lexical database differs from the selection of examples for the dictionary. Namely, the examples in the SLD also have the potential to become good dictionary examples, with only a few modifications needed. The SLD is particularly important for Slovenian and international lexicography because of the methodology used in its compilation. Namely, several methods combining lexicographic work with automatic extraction of data (including examples) have been developed and tested, and represent a basis for the compilation of the *Dictionary of Modern Slovene Language* (DMSL) and its database (see Section 5).

4 CHARACTERISTICS OF A GOOD DICTIONARY EXAMPLE

The most frequently mentioned characteristics of good dictionary examples are naturalness or authenticity, typicality, informativeness, and intelligibility. Naturalness means that the example appears natural, i.e. like the one you would expect

to encounter in actual language use. It is for this reason that the naturalness of dictionary examples is often associated with authenticity, which is ensured by obtaining examples from corpora, collections of authentic texts, something that has become a standard practice in modern lexicography. It should be pointed out that dictionaries compiled before the corpus lexicography era already contained examples from authentic texts (e.g. the *Oxford English Dictionary*), or at least excerpts based on authentic texts (e.g. DSLL). However, many of those dictionaries adopted a practice of formulating or inventing examples based on lexicographers' intuition. Overreliance on one's intuition has been brought into question by the findings of corpus studies (e.g. Sinclair 1991; Hunston and Laviosa 2001), which is particularly relevant when selecting examples for general monolingual dictionaries.[18]

Similar to the principle of naturalness is the principle of typicality – examples must show typical usage of the word in terms of context, syntax, phraseology and the like. State-of-the-art corpus tools can already significantly help lexicographers with this task, as they can be used to identify common, and typical, grammatical relations, collocations, and even colligations of the word (e.g. predominant number of the word in a particular collocation).

An informative example brings added value to the entry, predominantly in terms of offering additional help to the user in understanding the definition. In addition, examples attest the information in the definition, and contextualise the use of the word in a particular sense or subsense. The informativeness of an example is also affected by the number of examples in the entry. Electronic media offer the possibility of including a high quantity of examples, although lexicographers should always be concerned with whether each additional example offers anything new to the entry. On the other hand, as Frankenberg-Garcia (2012; 2014) pointed out, several corpus examples can sometimes be even more useful than the definition.

Intelligibility of a dictionary example is achieved by selecting examples that do not contain complex syntax or rare or specialised vocabulary. Examples should also not be too long. All this will help users focus on the word and the relevant surrounding information, and reduce the amount of mental effort needed to process it all. Still, certain features are often difficult to avoid; for example, rare and "more demanding" words are often used together with other rare words, which means the lexicographer needs to select such examples to fulfil the criteria of naturalness and typicality. While examples should not be too long, they must also not be too short, especially if a dictionary is to be used for encoding purposes where the users require as much contextual information as possible.

---

18   This is less true of dictionaries for non-native speakers, as, according to Atkins and Rundell (2008: 456), many pre-corpus English dictionaries for non-native speakers contained many good dictionary examples, which looked authentic but were not.

Form has become an important characteristic of a dictionary example; whole sentences are found in more and more dictionaries, even in general monolingual dictionaries for native speakers, which until a few decades ago used only excerpts or (very) short examples. There are two main reasons for this development: firstly, studies have shown that excerpts and similar short examples, taken out of sentences, seem abstract and unnatural (see e.g. Williams 1996). Secondly, in printed dictionaries, shorter examples are preferred due to spatial constraints, and the rise of digital media, especially the online medium, has done away with this limitation.

A separate and very important topic in example selection is ideological perspective, as examples reflect the ideology of the dictionary, i.e. reality as seen by lexicographers. Lexicographers use examples to convey information that could not be included in the definition because it is either too complex or ideologically too explicit (cf. Meschonnic 1991; Béjoint 2000; Epple 2000; Schutz 2002; Gorjanc 2004; 2005; 2012). Consequently, examples are an element of dictionary microstructure which offers the clearest reflection of social values, and relatedly, the values of the dictionary team (Gorjanc 2014). Analysing vocabulary related to homosexuals in DSLL, Gorjanc (2014) shows how social stereotypes can be presented in a dictionary as acceptable or part of the norm. Problems with ideological changes can also be observed in examples in DSLL2 (Krek 2014a: 145-147). It is therefore vital that lexicographers selecting examples are aware of their non-neutral role, and thus sensitive to social values and socially responsible (Béjoint 2000: 124).

Finding an example that meets all the above mentioned criteria is far from easy. Although nowadays lexicographers have very large corpora and consequently many potential dictionary examples at their disposal, it is often the case that they find sentences that meet two criteria, even three, but very rarely those that meet all the criteria of a good dictionary example. In fact, candidates for dictionary examples could be grouped on a scale from bad, more bad than good, reasonably or potentially good, and good; good candidate examples are those that can be used in a dictionary without any modifications. But, as mentioned above, such examples are less common, and there are more potentially good examples, i.e. examples that need minor modifications. However, if the decision is made to include modified examples in the dictionary, what about the principle of authenticity? Will the dictionary, or dictionary examples, still be considered corpus-based? As argued by Atkins and Rundell (2008: 458), the choice between invented and authentic examples is often misleading, because it does not reflect actual lexicographic practice. Even corpus-driven dictionaries like COBUILD include modified examples, although it should be stressed that the COBUILD lexicographic team tried to avoid modifying the examples as much as possible (Fox 1987).

Most common forms of example modification are shortening or omission of irrelevant parts, such as relative clauses or interjected clauses, simplification of complex syntax, and replacement of rare words or phrases with more common ones, or marked vocabulary with less marked vocabulary. Shortening is probably the least contentious practice, and is generally needed to meet the criterion of informativeness, as sentences often contain parts that can be deemed redundant or irrelevant, if not provided with more context. This is the case with *na primer* ('for example') in the sentence for the headword *anonimnost* ('anonimity'):

> *Jane Austen, na primer, je živela v popolni* **anonimnosti**.

> Jane Austen, for example, lived in total **anonimity**.

On the other hand, the simplification of complex syntax and replacing certain words can significantly affect the naturalness or typicality of an example. There are cases when replacing words cannot be avoided, for example proper names need to be replaced with pronouns or generic names to avoid offending individuals (e.g. Janez Novak; 'John Smith') or words that may offend particular social groups need to be replaced with more neutral ones. However, even this is not always straightforward, especially if the person in question is a public person or the name is closely related to the context of the word that is exemplified. The corpus example for *mojstrsko* (masterfully) would not have had the same informative value, and would also not appear natural, if *Christiano Ronaldo* had been replaced with a generic name such as Janez Novak ('John Smith'):

> *Izid polčasa in tudi končni izid je z* **mojstrsko** *izvedenim prostim strelom postavil Cristiano Ronaldo.*

> The half-time score, and also the final result, was decided by Cristiano Ronaldo's **masterfully** taken free kick.

The frequency and extent of example modification also depend on the target users of the dictionary. Examples for a dictionary for non-native speakers, or a dictionary for younger native speakers who are still developing their language proficiency and possess a smaller vocabulary, will be subjected to modification much more often than examples intended for dictionaries targeted at adult native speakers.[19] The decision about modification of examples should also be driven by expected use. For example, if the dictionary is supposed to help the users with both decoding and encoding, then the examples must remain as natural and typical as possible.

A special form of example modification is language correction. If we find a good corpus sentence with a missing comma, can we insert a comma and include the

---

19 Even Atkins and Rundell (2008) limit their approval of example modification almost solely to dictionaries for non-native speakers.

example in a dictionary? And if a sentence contains a misspelled word, a word in the wrong case, or an incorrect word order? Correcting spelling and some other minor mistakes may seem trivial, but the line between a minor and a major mistake can be very subjective. Some lexicographers may consider the replacement of a longer phrase or wording as completely acceptable, even though this is very close to inventing the examples altogether. To sum up, it is good to adhere to the principle of giving priority to finding good corpus sentences that do not need any modification, and only when such sentences cannot be found do we look for sentences with (minor) language errors that can become good dictionary example if these are corrected.

5 METHODS OF IDENTIFYING GOOD DICTIONARY EXAMPLES

Identifying (good) dictionary examples is a very laborious and potentially time-consuming, and thus expensive, process. One reason is that finding an example that meets all the criteria is very difficult. Moreover, corpora are getting larger and larger, which in most cases means a bigger selection of example candidates for the lexicographer, but also a greater number of examples to analyse. Thirdly, examples are a key microstructural element, and can be found under many different parts of the entry, such as senses, subsenses, compounds, phrases, collocates etc. All this means that the lexicographer needs to search for a lot of examples in a large amount of data for each dictionary entry.

There are two methods of identifying good dictionary examples: manual and semi-automatic. When using the manual method, the lexicographer can use the sort, filter and other functions of the corpus tools. Additional help is provided by the division of examples according to collocations and grammar relations. In the Sketch Engine[20] corpus tool (Kilgarriff et al. 2004) this option is offered by the Word sketches feature.

In the semi-automatic method, a tool for identifying good dictionary examples, such as GDEX (Good Dictionary Examples; Kilgarriff et al. 2008), offers a selection of candidate sentences to the lexicographer who then selects the most appropriate ones. GDEX (see also Section 5.1) ranks corpus sentences according to characteristics such as length, whole-sentence form, sentence complexity, presence or absence of rare words, email addresses or web addresses, and so on. Many of these characteristics are indirectly related to characteristics of a good dictionary example, such as typicality, informativeness and understandability. The characteristics can be divided into mandatory and less/more desired. The former are those

---

20 http://www.sketchengine.co.uk/

that the example must include; if only one characteristic is scored as negative, the example is ranked to the bottom of the list of candidate sentences. For the less/more desired characteristics we set the points added or deducted for each characteristic, and the example ranking is determined by a total sum of points from all the characteristics.

The main difference between the two methods is how much time they take, as the semi-automatic method is much quicker than the manual one, without being any less reliable (Kosem et al. 2012b; 2013b). In modern corpus lexicography, the semi-automatic method is thus replacing the manual method, especially in projects that involve the compilation of dictionary databases which include more examples than the dictionaries based on them (e.g. CDD).

6 EXAMPLES IN THE DICTIONARY OF MODERN SLOVENE LANGUAGE

This section discusses the treatment of examples in the proposed DMSL, including their identification and way of recording them in the dictionary database, and the differences between the examples in the dictionary database and the dictionary. The section concludes with a discussion on the different options of example presentation offered by digital media.

6.1 Identifying and recording dictionary examples

Identification of examples with the GDEX tool is part of a semi-automatic method called the automatic extraction of lexical data (AELD; Kosem et al. 2012b). This includes the automatic extraction of data (grammatical relations, collocates and examples, as well as certain information on the headword and also suggestions for labels), via Word sketches in the Sketch Engine, using an API script, taken directly from the corpus and put into a dictionary-writing system (DWS). In DWS, the data is then examined, selected and edited by the lexicographers.[21] This still provides the lexicographers with enough information for a thorough analysis and entry compilation. Experience on the SLD project has shown that a lexicographer using this method inspects a similar amount of examples, often even more of them, than by using a combination of semi-automatic and manual methods with the corpus tool (i.e. analysing word sketches). One of the advantages of AELD is that it dispenses with a lot of tedious copying and pasting of

---

21 A similar method has already been envisaged by Rundell and Kilgarriff (2011).

data between the corpus tool and the DWS. Another advantage is the quicker, more dispersed and consequently more reliable analysis.

The key task of AELD in terms of dictionary examples is the preparation of GDEX configuration(s). The GDEX configuration developed in 2011 during the compilation of the SLD (Kosem et al. 2011) was quite successful in identifying good dictionary examples, as on average three out of 10 examples offered were considered good. However, the requirements of AELD are different; only the first X examples (usually three to five) are extracted and all are expected to be (potentially) good. In addition, the analysis of the initial configuration for Slovene has pointed to significant differences between the quality of examples across word classes. This is why the decision was made to devise a different configuration for each word class. The procedure was done in two steps: first, the initial configuration for each word class was devised by analysing (good) examples in the SLD. New versions were then developed by adjusting the values of different classifiers and evaluating the results by comparing them with those of the previous configuration. The procedure was repeated until the GDEX configurations that provided the most satisfactory results were obtained. Importantly, the procedure also enabled us to devise several new classifiers that were not part of the original GDEX configuration. The classifiers used in AELD are:

- Whole sentence. Whole-sentence examples are given priority.

- No tokens with a frequency of less than three. This classifier seeks to exclude[22] examples with rare words, rare misspellings or corpus noise.

- Sentence must be longer than seven tokens. We seek to avoid examples that are too short, as they are often lacking context. The principle is that it is easier to shorten longer examples than search for new ones.

- Sentence must be shorter than 60 tokens. Only very long sentences are excluded, longer sentences can always be shortened.

- Lemma must not occur more than once. This important classifier excludes examples with a repeated headword, as such examples are normally less understandable and informative.

- Sentence must not contain web or email addresses.

- Optimal length (between X and Y tokens). While the classifiers for minimal and maximum length exclude sentences that are either too short or too long, this classifier awards points to sentences with the length in a given range. The most frequently used range is 15-40 tokens, but it depends on word class. The analysis of good examples in

---

22 The word exclude is used here because the algorithm ranks such examples so low that the lexicographers in most cases do not see them.

LBS, which was part of the preparation of the first version of GDEX for Slovene (Kosem 2012), has shown that the average length of examples for adjective entries is 28.64 tokens, for nouns 27.03 tokens and for adverbs 27.39 tokens.

- Rare lemmas. The classifier penalises sentences for each rare lemma. The frequency limit determining what is rare is determined considering the size of the corpus.

- Tokens longer than 12 characters. The classifier penalises sentences for each token that meets this criterion. This is because analysis has shown that tokens longer than 12 characters are in most cases non-words or corpus noise.

- Number of punctuation marks (commas excluded). The classifier penalises sentences that contain more punctuation marks than a set value. Commas are not included in the count as they are addressed by a separate classifier.

- Number of commas in the sentence. The classifier penalises sentences with more than three commas, as analysis has shown that such sentences are often more complex and thus less likely candidates for good examples.

- Tokens beginning with a capital letter. The classifier penalises sentences containing tokens with capital letters, and the main purpose is to complement the classifier penalising proper names.

- Tokens with mixed symbols (e.g. letters and numbers). Another classifier that helps identify, and penalise, non-words and corpus noise in the sentences.

- Proper names. The classifier penalises sentences containing tokens that are tagged as proper names. The penalty is awarded for each proper name in the sentence.

- Pronouns. The classifier penalises every pronoun in the sentence. The classifier is particularly important for sentences with several pronouns, as these often require a lot of additional context and are thus less understandable.

- Position of lemma in the sentence. The classifier penalises sentences in which the headword occurs outside of a given range in the sentence. For example, for the verb headwords it was determined that much better example candidates are sentences in which the headword does not occur at the beginning of the sentence (in the first 40% of tokens in the sentence).

- Stop list of sentence initial words. The evaluation of various configurations revealed that certain words at the beginning of a sentence are a good indicator of a bad candidate sentence. During the evaluation, a list of such words was devised. The list includes words such as *sledi* ('following'), *tovrsten* ('such'), *oboji* ('both') and so on, indicating that the sentence requires additional (preceding) context. The classifier penalises sentences beginning with one of the words on the list.

- Stop list of sentence initial multi-word units. The classifier is similar to the one above, penalising sentences beginning with a multi-word unit found on the previously devised list.

- Second collocate. One of the most important classifiers, awarding points to sentences containing the most typical collocates of a given collocation, indirectly detects colligational typicality. For example, the sentences containing the collocation *klavrn + podoba* ('miserable state') are awarded points if they also contain the verb *kazati* ('show'), which is a statistically important collocate of this collocation. Further analysis has also shown that such sentences also contain a longer syntactic pattern *kazati klavrno podobo česa* ('indicate miserable state of sth')

- Levenshtein distance. An algorithm[23] that measures similarity between strings, in our case sentences. If the classifier finds two similar or even the same sentences, it sends one of them (the one with the lower score) to the bottom of the list of candidate sentences.

Most of the differences between configurations for different word classes can be observed between the settings of individual classifiers, although differences in classifiers can also be observed (e.g. an additional classifier for the position of the lemma in the sentence is found only in the configurations for verbs). Each sentence receives a score between 0 and 1, indicating a total of all classifier values (as mentioned above, classifiers are attributed weights according to their significance in comparison with other classifiers). The GDEX tool then ranks the candidate sentences from the highest to the lowest score, and this determines which top X examples are exported with the AELD method.

Each example should include metadata about the text, such as year, source, author, title and so on. This ensures example traceability and offers different possibilities of example visualisation in the dictionary. It is never good to consider only the needs of a particular dictionary, as searching the corpus for missing information is a long and time-consuming process (this is true not only for examples but also for the other parts of the dictionary). A good indication of the benefits of example metadata is seen in the updating of the dictionary:

---

23 http://en.wikipedia.org/wiki/Levenshtein_distance. This measure was recently replaced by the Jaccard similarity coefficient (https://en.wikipedia.org/wiki/Jaccard_index).

if one wants to replace older examples with newer ones, it is possible to use the information on the year in which the text was produced to identify all the examples that were produced before a certain time. Example metadata can also be useful in the detection of ideological examples in the entry. Thus, when extracting with GDEX, we should pay particular attention to cases when most of the examples in the entry or under a particular sense are from a single source or only a few sources (cf. the analysis of examples for *pederastija* in Gorjanc 2014).

6.2   Examples in a dictionary database vs examples in a dictionary

The discussion on identifying and recording examples also needs to consider the relationship between a dictionary and its database, and also the role of the dictionary archive (Figure 1). The procedures described in this paper are especially relevant for DMSL, but since the compilation of this dictionary involves undertaking Slovene language description from scratch, a large proportion of the data obtained with corpus analysis (including examples) could be used in the compilation of other dictionaries. A particular example could thus be used in different dictionaries; in dictionaries for adult native speakers it can be used without any modifications, while in dictionaries for younger speakers the example can be slightly modified, e.g. by shortening or replacing rare words with more frequent ones that these users are likely to be more familiar with. Due to such potential multi-purpose nature of dictionary examples, all extracted corpus sentences and their metadata need to be archived in their original form, as found in the corpus.

An archive of extracted corpus sentences also makes possible analysing the number and type of modifications made to these sentences when turning them into dictionary examples. The findings of such analyses can then be used to improve the configurations used in their extraction. Even bad or irrelevant sentences that are part of automatic data extraction and need to be excluded from the database should be archived, as analysing their characteristics can also help improve the configurations for extraction. A similar approach was already used when developing the first version of GDEX for Slovene (Kosem et al. 2011); the parameters of the classifiers in the test configuration were improved with an analysis of the examples that were selected (good) or not selected (bad) during the evaluation. In addition, the role of dictionary data in the development of language technologies for Slovene should not be forgotten. In short, the planning stage of a dictionary project should devote a considerable amount of time to considering the various types of data in the dictionary database and the ways they will be recorded. From this perspective, any dictionary, even a

general dictionary, is merely one of several products that can be derived based on the database.
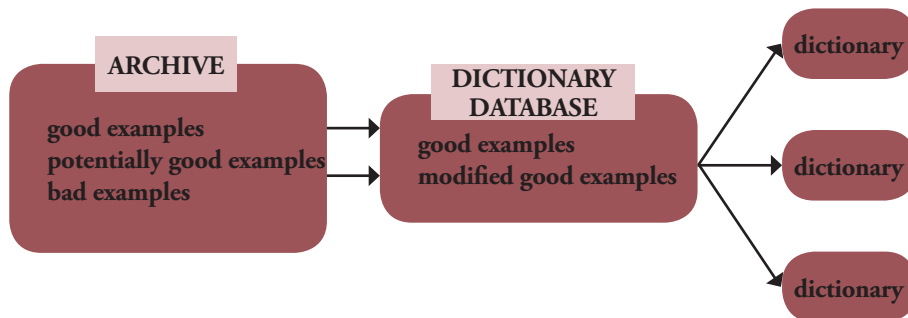
Figure 1: Examples in an archive, dictionary database and dictionaries.

A dictionary database contains (much) more information than any dictionary based on it, which means that lexicographers can spent a great deal of their time recording information that might not end up in the dictionary. As such, planning of the dictionary database should follow two principles: a) automating as many (routine) lexicographic procedures as possible, and b) ensuring that every single lexicographic decision is recorded and utilized. Consequently, the use of methods such as AELD is more or less mandatory, as without them it is difficult to imagine the successful compilation of the database (and a dictionary based on it) in a time frame that would satisfy funders as well as dictionary users. Let us consider the benefits of using automation on a very basic task, namely typing a headword and its word class in the dictionary entry in the database. Assuming it takes us on average five seconds to type these two types of information, we spend on this task 500,000 seconds for 100,000 entries, or little less than 139 hours. AELD writes these two types of information automatically, saving us nearly one person/month on the project. Much the same is true of lexicographic decisions: using manual analysis or analysis in a corpus tool, even if using a tool like GDEX, lexicographers must still examine many corpus sentences and decide whether each is a good dictionary example or not. But since the lexicographers only copy good or potentially good examples in the dictionary database, only such decisions can be archived. The AELD method makes it possible to record or track every single decision: the identification of a good example (a corpus sentence remains unchanged in the dictionary database, so it is the same as the final dictionary example), a potentially good example (the corpus sentence has been modified slightly when turning it into a dictionary example), and bad examples (the corpus sentence has been deleted from the database).

To additionally assist lexicographers with the identification of good examples, other methods such as crowdsourcing can be utilized. However, as good dictionary examples have to meet a combination of different criteria, it is difficult to imagine how such a task could be trusted to non-lexicographers. The answer is that it can be, if we are aware of the characteristics and limitations of crowdsourcing (see Čibej et al. 2015; Fišer and Čibej 2015). First and foremost, the tasks should be simple, mainly in the form of multiple-choice questions with options 'Yes', 'No' and 'I don't know'. In addition, the tasks should not focus on determining something abstract (e.g. the characteristics of a good dictionary example) or level of degree; questions such as *Is this a good dictionary example?* and *How good is this example?* are thus not suitable. Tests with crowdsourcing on examples from the SLD have shown that examples are very useful in tasks aimed at identifying incorrect information (e.g. when the use of the headword and its collocate in the example does not match the identified grammatical relation) or at assigning collocates and their examples to different senses and subsenses.

6.3   Visualizing examples

Lexicographic work with examples does not, or should not, conclude with the recording of good examples in the database or/and the dictionary. This is because presentation is very important if examples are to achieve their purpose. Research studies in the visualisation aspects of (electronic) dictionaries, although still rare in lexicography, indicate that visualisation plays a key role in the readability and retention of the dictionary information (Nesi 2011). Considering that the examples occupy a fairly large, if not the largest, share of text in any dictionary, suitable visualisation and presentation of them is obviously vital.

One of the techniques used to assist users in reading dictionary examples is highlighting the headword. Especially in modern dictionaries that often contain (longer) whole-sentence examples, it is useful to direct users' attention to the headword, i.e. the part of the entry with information more relevant to their needs. In most cases such highlighting is found in the form of bold text, while in electronic dictionaries a different colour is also used (Figure 2). Italics are rarely used for highlighting, mainly because in most dictionaries examples are already offered in italics, and so this option seems less effective (see Figure 3). Highlighting is also used to point to typical collocations, compounds, multiword units and phrases (Figure 4). However, it is definitely recommended to test any visualisation and presentation solution on the target users, preferably before publishing the dictionary.

fach

Prawie 60 lat temu zaczął się uczyć fryzjerskiego fachu i nadal pracuje w zawodzie.

źródło: NKJP: Katarzyna Skrzypek: Cyrkiel za uszami, Dziennik Zachodni, 2005-03-31

Miał dobry fach - przez kilkanaście lat pracował w dużej warszawskiej fabryce jako spawacz, był też ślusarzem, szlifierzem i monterem.

źródło: NKJP: Monika Mikołajczuk: Między Kantem a Wolterem, Polityka, 2001-07-14

Figure 2: Red headword highlighted in examples (CDP).

Examples of CLICK

He *clicked* his heels together and saluted the officer.

Her heels *clicked* on the marble floor.

Press the door until you hear the latch *click*.

To open the program, point at the icon and *click* the left mouse button.

Click here to check spelling in the document.

I know him fairly well, but we've never really *clicked*.

Figure 3: Highlighting the headword in examples using italics (MWOD).

---

3 **SURE ABOUT SOMETHING** feeling certain that you know or understand something [↪ clearly]
clear about/on
🔊 *Are you all clear now about what you have to do?*
clear whether/what/how etc
🔊 *I'm still not really clear how this machine works.*
🔊 *Let me **get** this **clear** - you hadn't seen her in three days?*
🔊 *a clearer understanding of the issues*
4 **THINKING** able to think sensibly and quickly [↪ clarity, clearly]:
🔊 *She felt that her thinking was clearer now.*
🔊 *In the morning, with a **clear head**, she'd tackle the problem.*

Figure 4: Highlighted collocations and phrases in examples
(*Longman Dictionary*) [24]

---

24 http://www.ldoceonline.com/

As already mentioned, it is useful to have as much metadata as possible on each example in the dictionary database. Although such metadata can be shown to the users, dictionaries rarely include it – one exception is the *Comprehensive Diction-ary of Polish* (Figure 2) – for a simple reason: metadata such as source, author(s) or title of the text from which the example comes from are referential and suggest/require direct copying from the source, taking away from the lexicographers the option of making any modifications. Another reason against showing example metadata is its non-essential nature; it would take up precious space on the screen and can distract the users' attention from the main purpose of the examples, namely showing the use of the headword in a particular sense.

The principle of informativeness also limits the lexicographers in the number of examples they can provide under each element of the entry. Even with that in mind, one can quickly end up with several examples per sense, subsense, syntactic structure or collocation, which can cause problems with visualisation/presenta-tion. A good solution is to show only a certain number of examples, offering ad-ditional examples on a click (Figures 5 and 6). More and more online dictionaries have also started to offer links to corpus hits, undoubtedly a very useful feature for (advanced) users.

> 1.1 **Pursuing** a commercial activity on a significant scale:
> *'many large investors are likely to take a different view'*
>
> MORE EXAMPLE SENTENCES
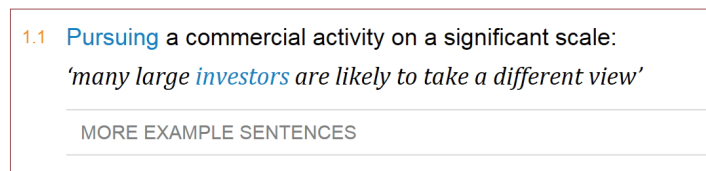
Figure 5: A link to show additional examples (more example sentences in Ox-ford Dictionaries).

> 1.1 Pursuing a commercial activity on a significant scale:
> *'many large investors are likely to take a different view'*
>
> MORE EXAMPLE SENTENCES ⌃
>
> *'The basic cause of the changed activities of large businesses is a matter of debate.'*
>
> *'The fate of rival bids for NatWest rest in the hands of the faceless large investors.'*
>
> *'Being a large economy, the euro zone is much less open than individual member states.'*
>
> GET MORE EXAMPLES

Figure 6: Additional examples revealed (Oxford Dictionaries).

7   CONCLUSION

Examples require a great deal attention when planning a dictionary. The instructions given to lexicographers thus need to clearly delineate the characteristics of good examples, including concrete cases of good and bad practice, and the role of ideology. It is also paramount to use or develop tools that facilitate consistency in adhering to these characteristics. In addition, examples of allowed modifications should be prepared, as well as a suitable system for archiving sentences in the form they are extracted from the corpora. DMSL will be an important resource for the development of language technologies for Slovenian, which means that the database should include as many examples (and their metadata) as possible.

The aim to include as many examples as possible necessitates the use of semi-automatic methods of example extraction from the corpus. Not using such methods can prolong the compilation of the dictionary to such extent that the examples need to be replaced before the work is even completed. This is the rationale behind using the AELD method that we propose for identifying and recording examples in DMSL, and which represents a new approach to lexicographic analysis. Based on the experience gained during the SLD project, a similar method has already been used in the compilation of a collocations dictionary for non-native speakers of Estonian (Kallas et al. 2015).

An important task for the lexicographic community is to keep conducting studies on how, when and in what ways dictionary users consult examples and what kind of examples are most useful to them. The findings of such studies will enable further improvements to the procedures used for example selection, and the techniques used to present them in dictionaries.

How specialised should a general dictionary be?

Špela Vintar

Abstract

The article discusses theoretical and methodological issues related to specialised vocabulary in the *Dictionary of Modern Slovene Language*. We address key questions such as the role of terminology in a general dictionary, user requirements and needs, the complexity of distinction between general and specialised terms, and finally corpus composition and corpus representativity. We propose a model where lexical items are categorised into three levels of termhood, and each level of specialisation requires a different strategy of lexicographical description. By illustrating possible relations between the proposed categories and the corpus-based methodology of candidate extraction we establish a working methodology for handling specialised units in a general dictionary.

Keywords: specialised vocabulary, general dictionary, terminology extraction, user requirements, specialised vs. general

1   INTRODUCTION

We all use specialised lexical items in our everyday lives, for the simple reason that nearly everyone engages in activities or fields which are not shared by all speakers of the language, and which involve the communication of specialised skills or know-how. It seems that as native speakers of a language we are equipped with an intuitive gauge of termhood by distinguishing between highly and less specialised items, and we often justify such intuitions with statements such as "This is sailors' jargon" or "I can't understand this medical gibberish". But would we expect to find such items in a general dictionary, or would we consult a dictionary at all when encountering them?

In this paper we present a series of reflections on the role of specialised vocabulary in a general dictionary, specifically the *Dictionary of Modern Slovene Language* (DMSL), considering various aspects from the established traditions and practices in Slovene lexicography, user requirements and profiles, corpus composition and representativeness, to lexicographic description and data presentation for different target groups. The aim of our discussion is to establish a methodological framework which would provide guidelines on the treatment of specialised vocabulary through all stages of dictionary creation, and which would be sustainable both in terms of adaptability and scalability to different target groups and in terms of labour intensity by employing (semi-)automatic techniques of data acquisition.

Clearly the above goal is not an easy one, and perhaps one would expect that such fundamental methodological questions have been extensively dealt with by lexicographers in related dictionary projects worldwide, and that their findings could easily be transferred into the Slovene language community. Surprisingly though, the body of literature with in-depth descriptions of methodological decisions regarding specialised lexis in general dictionaries is relatively lean, especially in comparison to the many studies dealing with specialised dictionaries, terminology or so-called LSP (language for special purposes); it is therefore necessary to draw conclusions from general dictionaries themselves, or occasionally their introductions. Moreover, the experiences and methodologies from related dictionary projects elsewhere are not directly replicable in the Slovene situation, firstly because of the strong influence of the specific lexicographic history in Slovenia, and secondly because of the currently prevailing social norms reflected in the official language policy. Both of these factors will inevitably influence the expectations of potential dictionary users and, as a consequence, the range of functions that the new dictionary should fulfil. The proposed methodological framework therefore relies on existing practices only as the point of departure from which an iterative cycle of improvements should evolve.

2 THE ROLE OF SPECIALISED VOCABULARY IN A GENERAL DICTIONARY

The tendency to include technical terms into general mono- and multilingual dictionaries has been on the increase since the 19th century (Boulanger 1996: 141), partly because the impact of science and technology on everyday life has been growing since the Enlightenment, but also due to the rising level of education and the inclusion of the so-called "technolects" into vernacular language use. The second part of the 19th century was a crucial period for the formation of basic terminology in Slovene, both in natural and human sciences, driven largely by numerous translations of scientific and reference works from German and other languages into Slovene (Prunč 2009).

From the 20th century onwards general language dictionaries gradually diminished their normative character and increasingly started to consider the expectations of users, which entailed the demand for a broad coverage of specialised items in a comprehensive dictionary. Landau (2001) even claims that contemporary comprehensive dictionaries seem as if multiple LSP dictionaries have been added to the traditional general language dictionary, mostly because emerging disciplines continually produce more new lexical items than general language. The reasons for the growing ratio of terms in general dictionaries are summarized by Josselin-Leray (2005) as follows:

a) An almost two centuries long tradition in lexicography of increasingly including specialised lexis in general dictionaries.

b) The growing trend of despecialisation (determinologisation), the process through which specialised terms move into everyday language and typically modify or broaden their meaning.

c) The didactic role of general dictionaries, which through working to meet the needs of EFL learners revolutionised English lexicography and brought profound changes to the dictionary-making process worldwide.

d) Striving for comprehensiveness, whereby a single reference work aims to satisfy the needs of the broadest possible target audience.

e) The expectations and requirements of users, who are today better informed and more interested in science and technology than in the past.

All of these reasons apply to the Slovene language community, and thus build a case for a strong representation of terminology in the new contemporary dictionary of Slovene.

The only existing comprehensive dictionary of Slovene, the *Dictionary of Slovene Literary Language* (DSLL), gives terminology an important role – indeed its authors explain their rationale in the Introduction to DSLL (DSLL, Introduction: XVI-XVIII), as follows:

> Terminology is included in the approximate scope of secondary school education, in particular if [term use] is supported by evidence from journalist or popular scientific publications. The terminological entries were created partly by copying from popular science books, secondary school textbooks and specialised dictionaries, and partly by contributions from over one hundred domain specialists. Of the entire term inventory collected, only terms used in the present day were retained in the dictionary.

In the proposal for the *New Dictionary of Slovene Literary Language* (NDSLL; Gliha Komac et al. 2015: 49–51), published in March 2015, the methodological considerations concerning terminology are largely retained from the old DSLL, with some amendments. The authors of the proposal distinguish between fully and partially despecialised lexical items, whereby the former are lexicographically treated in the same manner as general words with no domain labels or counselling from experts, while the latter are to be described with simplified but scientifically correct definitions formulated by domain experts. The proposal remains vague about the distinguishing criteria between the first and second groups. The authors refer to the level of despecialisation and the familiarity of the term to general users, with both criteria to be determined from corpus data. No further details are provided about this, in our view crucial, methodological procedure.

Returning to the reasons for including terminology into general dictionaries, the despecialisation process can be frequently observed in Slovene, especially in fields such as information technology, finance, environment or sports, meaning that originally specialised terms work their way into everyday language and possibly modify their semantic and expressive scope. A general dictionary should reflect such use and define despecialised items accordingly. As for the corpus-based techniques facilitating the discovery of despecialised terms, a stratified cross-comparison of frequencies in subcorpora should provide useful clues. For example, if a term such as *infarkt* 'infarction' is found in a subcorpus of medical abstracts, but at the same time appears in general newspaper articles and user-generated contents with a different network of collocates and modifiers (*prometni infarkt, vremenski infarkt, svetovni infarkt, dolžniški infarkt*), this is a strong indicator of despecialisation and the broadening of meaning.

3 THE RATIO OF SPECIALISED VOCABULARY IN A GENERAL DICTIONARY

Assuming that users value the presence of specialised items in a general dictionary, the question that inevitably follows is how many terms should be included, or what the ideal proportion between terms and non-terms should be.

Several authors have addressed these questions, including Landau (1974), who analysed *Webster's Third New International Dictionary* and found it contained around 40 percent of terminology, with selected dictionary pages having up to 89 percent. Béjoint (1988: 360) discusses the importance of terminology, but is reluctant to specify portions or ratios because the distinction between terms and non-terms is anything but straightforward. A similar view is adopted by Boulanger and L'Homme (1991: 25), however a later study by Boulanger (1996: 147) identified between 40 and 50 percent of specialised items in English and French monolingual dictionaries. More recently, Vrbinc and Vrbinc (2013) explored the differences in the treatment of terminology in 3rd, 4th and 8th editions of *Oxford Advanced Learners Dictionary* (OALD). The authors specifically focused on the use of subject-field labels and the improvements thereof, but also found that the proportion of scientific and technical vocabulary continually increased from one edition to the next.

The approach advocated to general dictionary creation here seeks to be pragmatic in the sense that the proposed methodology for the inclusion and treatment of specialised vocabulary should be feasible in terms of time and funding, while fulfilling all the necessary criteria regarding efficiency and sustainability. An online dictionary can follow the user-centred approach in selecting which information to display to which type of user, meaning that specialised vocabulary – if properly labelled as such – need not be restrained beforehand in terms of its volume or specificity. State-of-the-art computational methods facilitate the process of extracting terms, definitions, collocations and examples from corpora, moreover they provide reliable clues about their use, frequency or variants across registers and text types. The severe space constraints which governed data presentation in the times of printed dictionaries no longer apply, and today limitations are posed not by data storage capacities or bandwidth, but by the information processing capacities of the human user.

Still, even with the best information extraction techniques and language processing tools, automatically obtained data still requires a substantial amount of validation, editing and completion by lexicographers before it can be presented to the user. Returning to the issue of the volume of specialised items to be included into general dictionaries, one should therefore note that the decision to include a

large number of terms inevitably means a large investment of human labour into this task, possibly including the involvement of domain experts.

Within the context of corpus-based lexicography, which essentially describes language use, it would therefore have been futile to predefine the amount of specialised lexical items to be included, and lexicographers might instead adhere to the relatively loose principle that the dictionary should be as specialised as necessary in order to fulfil the broadest possible range of information needs from diverse user groups, while retaining the character of a general language dictionary.

The last question we briefly touch upon is the balanced representation of domains, in other words, should a general dictionary contain equal portions of terms from all the specialised domains it includes. A review of lexicographic traditions (Josselin-Leray 2005: 146ff) shows that balance between domains was generally considered irrelevant or impossible to achieve. In many cases the reasons for a detailed representation of a particular domain in a dictionary were purely anecdotal (Béjoint 1988: 361): "One of the editors of the OED happened to be an amateur mineralogist, and consequently the [SOED] is particularly rich in words of mineralogy." Specialised domains differ in the type and number of terms they use, and some domains are certainly of more interest to the general public than others, a fact likely to be directly observable in a well-balanced reference corpus. Ahmad et al. (1995) claimed that a general dictionary should focus on the domains and terms where the layperson's and the expert's interests overlap. In the following sections we propose some methodological guidelines to help us measure this overlap and classify specialised terms accordingly, but because our approach is corpus-based we first discuss notions of representativeness and balance in the context of specialised vocabulary.

4   COMPILING A CORPUS TO EXTRACT SPECIALISED VOCABULARY: SOME CONSIDERATIONS

Since the beginnings of corpus linguistics, digital data collections have provided invaluable empirical evidence for all kinds of lexicography. For general language dictionaries the most widely used type is the reference corpus, which is often understood as a common denominator representing the broad range of language varieties and text types occurring in a language community. While the compilers of early reference corpora devoted a great deal of critical reflection to the notions of balance and representativeness, in recent years we have witnessed a trend towards compiling very large web-crawled corpora which at best represent the language of the Internet, but cannot be taken as representative nor balanced samples of the language as a whole.

For Slovene, the first reference corpus FIDA (Erjavec et al. 1998) was built much in line with the principles of the Czech or British national corpora. The currently largest reference corpus is Gigafida, containing over a billion tokens from a broad range of genres and varieties, while not claiming to be balanced. Its subset KRES, with 100 million tokens, is "artificially" balanced in that it contains equal portions of the five main text genres.

If (absolute or relative) corpus frequency is considered one of the essential criteria for the lexicographic treatment of general language, it is certainly not always reliable for specialised lexical items. While some specialised terms, for instance those pertaining to economy, finance or sports, might be well-represented in a general language corpus and thus bear witness to the above-mentioned overlap of experts' and lay people's interest, others, such as terms from the domains of math, physics or chemistry at the level of high school textbooks, will not appear as frequently.

For instance, three sibling terms (see below) occurring in a high school math textbook and designating polygons, *trapez* (trapezium), *paralelogram* (parallelogram) and *deltoid* (deltoid), are found to occur in extremely different frequency ranges in Gigafida and Kres.

	GF	Kres
paralelogram	107	18
trapez	923	126
deltoid	23	3

A general dictionary striving to include terminology up to the level of secondary education should probably contain all three, but it remains a challenge how to discern their termhood from corpus data alone. Of course, situations such as the above can easily be explained: *trapez* is a highly polysemous word which is found to have at least five other (semi)specialised meanings in Gigafida, including the domains of basketball, gymnastics, medicine, information technology and sailing. The lexicographer will very likely need to review all of these specialised meanings manually in order to decide whether they are to be included or not. In addition to polysemy, skewed frequencies may be the result of an imbalanced corpus. More specifically, Gigafida contains several years' editions of *Monitor*, a leading Slovenian IT magazine, which regularly publishes tests of IT equipment. *Korekcija trapeza* is a term frequently used in relation to screen calibration, so that the frequency of *trapez* is substantially increased due to the inclusion of *Monitor* in the data.

The above considerations bring us to the conclusion that if the dictionary has no claim to a balanced representation of various domains, but on the other hand

cannot rely solely on frequency data from a reference corpus as to which terms to include, it is necessary to identify priority domains which should be represented, and for these to provide sufficient material to allow us to exploit automatic methods of data extraction and processing. By sufficient material we mean subcorpora, of which some may already exist while others still need to be compiled.

The selection of priority domains along with the target genres and text varieties is closely related to the overall dictionary concept, and should reflect the needs and requirements of target users. However, since a thorough and extensive analysis of the Slovenian users' needs with regard to terminology has not been performed yet, these decisions will inevitably be subjective and intuitive. One attempt to overcome this issue is the analysis of Termania queries described below.

For the identification of priority domains we propose three guiding principles. The first is related to the above-mentioned intersection of expert and lay interest, meaning that the dictionary should focus on the domains which are frequently discussed in general public discourse and are well represented in the media. A quick scan through the 1,000 most frequent noun lemmata from the Gigafida corpus reveals the following list of topics: politics, sports, law, economy, finance, media, environment, administration, health, culture, IT, traffic, and tourism. It should be noted that such a scan might point us to the domains of the so-called public interest, but the lexical items occurring within the top 1,000 nouns could hardly be considered terms. Their use in despecialised contexts inevitably broadens their semantic spectrum and loosens their membership in a specialised domain.

The second principle addresses the target group of potential dictionary users in education, and at the same time aims to achieve the terminology coverage of a high school graduate. This implies that one of the essential subcorpora should consist of high school textbooks covering all subjects taught at the level of secondary education in Slovenia. Since the current version of Gigafida is imbalanced in this respect, one of the future tasks entails a systematic revision and extension of the textbook subcorpus.

The third guiding principle acknowledges the fact that even the largest and most carefully designed corpora cannot represent the entire vocabulary of a language. Leaving the huge landscape and variety of spoken discourse aside, certain areas of written communication are underrepresented in existing corpora. We have identified one such gap and labelled it broadly as life events, by which we mean a range of lexical items referring to various administrative, legal, social, religious, medical and other procedures people regularly encounter. Such vocabulary may be found in banking, insurance or administrative forms, identity cards and other types of documents associated with individual life events.

Another aspect of the third principle, which could also be referred to as the awareness of the noncomprehensiveness of corpora, is neology. New terms and expressions are coined almost exclusively in specialised domains, where a high term formation rate can be observed for the naming of new technologies, scientific findings, devices and services. Often these innovations are recieved with a wave of attention from the general public and the media. While it is difficult to follow the evolution of new terms through corpora their inclusion in the dictionary is important, especially in those cases where the newly coined term is not merely the result of a journalist's or translator's creativity, but represents the result of a term planning and harmonisation process.

One example of such a process is the quest for the Slovenian equivalent of crowdsourcing. At first the term was directly borrowed from English, and can be found in the original English spelling four times in Gigafida (crowdsourcing), then several possible translations started appearing: *moč množic, množicanje, množgančkanje, množičenje,* and *množično zunanje izvajanje.* Lively discussions about the most appropriate equivalent began and leading lexicographic institutions contributed their views, but still none of these equivalents can be found in today's Gigafida. For the new dictionary we must therefore systematically devise strategies to follow the evolution of terms and make well-founded decisions about their inclusion in the dictionary.

5 EXTRACTING SPECIALISED VOCABULARY AND OTHER LEXICOGRAPHICALLY RELEVANT DATA FROM CORPORA

Before the computer era lexicographers spent the bulk of their time building inventories of words in a language to be included in a dictionary. State-of-the-art language technologies substantially reduce this task and allow lexicographers to focus on the validation, revision and completion of automatically extracted information.

5.1 Adapting term extraction tools to the task at hand

Several tools exist for the automatic extraction of terminology from text for many languages, including Slovene (Vintar 2010). The LUIZ Term Extractor was originally been developed for bilingual extraction of terminology from English-Slovene parallel and comparable corpora, but may equally well be used for Slovene

alone. The underlying assumption of many term extraction methods is the notion of keyness (Scott 1997), which compares the frequency of a selected lexical item in a specialised corpus with its frequency in a general language corpus, and assumes that if the item is relevant or "key" to the specialised domain it will occur in the specialised corpus with a higher relative frequency than in the reference corpus. In LUIZ, the "keyness" of a term candidate is combined with part of speech patterns and a ranking heuristic to provide a list of candidate single- and multiword terms as output.

LUIZ has been tested for various domains (Vintar and Erjavec 2008; Vintar in Fišer 2009; Vintar 2010; Logar et al. 2012; Pollak 2014), but has never been used as a tool to extract terms for a general language dictionary. Several extraction parameters need to be adjusted to the task at hand, such as:

- The lenght of extracted terms. While for specialised terminography target units may contain three, four or more words, for a general dictionary it is better to focus on less specialised, therefore shorter terms containing one or two, and exceptionally three, words.

- Part-of-speech patterns. For specialised dictionaries the typical morphosyntactic term patterns may vary, but usually we attempt to achieve maximum recall by specifying all potentially productive patterns for a given language. Here, the great majority of specialised items is expected to be either single nouns or two-word combinations of adjective+noun or noun+noun, because in previous experiments these two patterns have proved to be most productive in term formation.

- Ranking heuristics. In a specialised terminography task it is not uncommon to extract units occurring only a few times, while for a general dictionary we tend to avoid items which are too specialised or rare.

- The definition of subcorpora for the computation of keyness. Keyness works if a clearly delimited specialised corpus is compared to a much larger reference corpus. In the context of our dictionary project, subcorpora will need to be defined for each individual extraction task. The textbook subcorpus is for instance entirely unsuitable for term extraction, because it contains a number of domains which will level each other out.

The result of multiple rounds of term extraction and cross-comparisons between subcorpora is lists of candidate terms for selected domains requiring thorough validation and supplementation. This will involve the identification of new terms which may have been overlooked by the term extractor due to low frequency, but also the ordering of terms into concept networks which will help identify gaps, near-synonyms and missing hyper- and hyponyms.

This stage of specialised vocabulary compilation already requires the involvement of domain experts to help identify term variants, obsolete or non-standard terms and synonyms, which will facilitate the lexicographers' decisions regarding classification into various groups of lexicographic treatment.

5.2 PILOT EXPERIMENT: EXTRACTING TERMS FROM TWO SUBCORPORA

In order to get a clearer insight into issues related to the representativeness of subcorpora and the necessary adjustments of the term extraction tool, we performed a pilot experiment. The LUIZ term extraction tool was used on two subcorpora, one containing a selection of texts pertaining to physics, biology and chemistry from the ccGigafida corpus (Logar Berginc et al. 2012: 77–97), and the other a more homogeneous specialised corpus of textbooks on music theory.

The subcorpus of natural science (Nature) is composed entirely of texts already included in ccGigafida and contains 13 primary or secondary school textbooks on natural science, biology, physics or chemistry, and 16 other popular scientific books from various publishers on the topics of astronomy, botany and gardening. We also included texts from related magazines including educational and popular-scientific periodic publications (*Gea* [geography], *National Geographic* [geography], *Kmetovalec* [agriculture], *Moj lepi vrt* [gardening], *Mrgolazen* [biology], *Ribič* [fishing], etc.).

The music subcorpus (Music) contains 10 contemporary textbooks used for musical education at the level of primary and secondary education. The corpus was compiled as part of a PhD study by Jelena Grazio at the Department of Musicology at the University of Ljubljana, and the textbooks deal with diverse areas of music theory (harmony, solfeggio, counterpoint, music forms). It is important to note that none of these textbooks had been part of the Gigafida or ccGigafida. Table 1 lists basic information about the two subcorpora.

Table 1: Basic information about the Music and Nature subcorpora

	Music	Nature
Tokens	280,060	1,053,897
Types	12,121	59,788
Number of documents	10	388
Text variety	textbooks	textbooks, popular-scientific books, magazines

For the term extraction experiment we used LUIZ with modified settings, more specifically we limited the extraction patterns to single nouns and adjective+noun phrases. Termhood is computed using the LUIZ heuristics, and is based on measuring keyness against the entire Gigafida corpus. Table 2 demonstrates the differences between subcorpora in size and the number of candidates extracted, with 9.3% single-word term candidates in Music and 13% in Nature, while two-word candidates seem to be more common in Music (6.8%) than Nature (2.2%).

Table 2: Number of term candidates extracted from both subcorpora

	Music	Nature
Noun	1,137	7,853
Adjective + Noun	828	1,309

The higher percentage of two-word terms in Music already highlights one important difference between the subcorpora – the level of specialisation, and another difference becomes apparent when we inspect lists of term candidates and observe a high level of domain homogeneity in Music, while the Nature corpus is much more diverse with regard to domains and topics. We analysed the top 150 term candidates from both single- and two-word lists and from both subcorpora and arrived at the following conclusions:

- The lists of terms from Nature reveal an imbalance between different text domains and sources, so that terms from certain domains seem overrepresented (such as fishing and gardening). For the purposes of term extraction subcorpora should ideally be as homogeneous and balanced as possible.

- The term candidates from the Music subcorpus contain a relatively low proportion (about 30% single-word nouns and 15% two-word candidates) of entirely despecialised terms which can be considered part of general vocabulary and require no special lexicographic treatment (e.g. *takt, glas, nota, melodija, skladba, harmonija etc.*; *notno črtovje, klasična glasba,* and *klavirska spremljava*). All the other candidates are specialised terms clearly belonging to the musical domain and not necessarily understood by lay persons (e.g. *modulacija, fuga, kvintakord*; *tritonusna kvinta, eolska septima,* and *napolitanski sekstakord).*

- With the Nature subcorpus, the situation is reversed: a large majority of single-word nouns refer to general concepts with a very vague affiliation to a specialised domain (*rastlina, voda, list, seme, plod, poganjek, temperatura, svetloba*; *okrasna rastlina, soška postrv, organski odpadek,* and *listna uš),* and the list of two-word candidates contains only about 15% of terms which might require a more technical definition (e.g. *ogljikov*

hidrat, celična membrana, magnetno polje, maščobna kislina, potencialna energija, and *vrtilni moment).*

It would appear that a more specialised and homogeneous corpus is more appropriate for the term extraction task, but on the other hand a general language dictionary need not contain highly specialised terms unless they fulfil the inclusion criteria discussed above. Textbooks and magazines certainly represent valuable sources of terminology, but their lexicographic description will depend on their degree of specificity which we discuss in more detail in the following section.

6   DEGREES OF SPECIFICITY AND LEXICOGRAPHIC DESCRIPTION

As illustrated by the examples above, terms vary in their degree of specificity. The goal of the new dictionary is to satisfy at least three diverse groups of users: learners at all levels of education, language professionals and those having language – including its peculiarities and specialised expressions – as a hobby, and as a consequence lexicographic descriptions should be tailored both to the needs of potential dictionary users and to the properties of the lexical item itself.

In line with these considerations we propose the categorisation of specialised items into three groups or baskets, each requiring a different approach to lexicographic description.

The first is the general basket, which contains the least specialised items. As such these items may exhibit a vague relation to a specific specialised domain, however knowledge of the domain is not a prerequisite for understanding or use. Such lexemes typically occur in the reference corpus with a relatively high frequency (> 3,000), and are used in general texts entirely devoid of domain-specific references. The lexicographic description may be generic, without domain labels or technical definitions, and theinput of experts is not required. Examples of such items from our Music subcorpus might include *tempo, koncert, dirigent, nota,* and *harmonija.*

The second is the so-called school basket, and contains terms less frequently encountered in general texts, although they still designate basic concepts of the domain and may already occur in textbooks at the level of primary education. Their membership in the specialised domain is clearly identifiable, however they do occur in the reference corpus (> 300). A lexicographic description may contain a domain label within the gloss, especially in cases where the despecialised meaning deviates from the domain-specific one. Examples include *sonata, akord, dur, mol, oboa,* and *trozvok.*

The third is the so-called technical basket, containing truly specialised lexical items less familiar to the general public and requiring some knowledge of the domain. They seldom occur in general texts and are not used with a despecialised meaning, although they may be considered for inclusion into the general dictionary for various reasons: either they occur in secondary school textbooks, and thus represent the vocabulary of an average high school graduate, or they have been found to belong to one of the priority domains and occur in the reference corpus with a minimum frequency. The lexicographic description will reflect the degree of specificity and should contain a domain label, the definition should be a (possibly simplified) terminological one, formulated or validated by a domain expert. Examples of such terms from the musical domain include *septima, alikvotni ton,* and *sekstakord.*

If the lexical item does not fulfil any of the above criteria for inclusion, meaning that it does not pass the threshold frequency in the reference corpus, is not part of a priority domain and does not occur in a textbook, nor does it represent an indispensable part of the conceptual network of another – more frequent – term, we may exclude it from further treatment and assume its degree of specificity to be too high for a general language dictionary.

The most challenging part of the proposed classification is the efficient and reliable exploitation of corpus data, since – as illustrated above – sheer corpus frequency cannot be seen as a reliable indicator of termhood. A common reason for skewed frequencies is ambiguity, where lexical items may have specialised meanings in several domains. An example is the musical expression *sinkopa*, which occurs 269 times in the Gigafida corpus, although the majority of occurrences pertain to the medical meaning of the term (a temporary loss of conscience). A number of occurrences include *sinkopa* in names of companies, musical groups and products, and only a small number represent the meaning in the musical domain. Since reliable word sense disambiguation tools still have not been developed for Slovene, ambiguity represents the largest obstacle to exploiting raw corpus data. This problem is largely resolved by using domain-specific subcorpora, and by cross-comparisons between them we may arrive at more realistic conclusions about the frequencies of individual meanings.

7 ANALYSING TERMANIA SEARCH QUERY LOGS TO ASSESS POTENTIAL USERS' TERMINOLOGY NEEDS

The new dictionary is ambitious in that it attempts to satisfy a broad range of target users whose communicative actions are increasingly embedded in the digital world. While a number of studies deal with user scenarios and their expectations

for online dictionaries, few studies focus specifically on users' needs with regard to terminology. An important contribution has been made by Amelie Josselin's PhD thesis (Josselin-Leray 2005), which investigates the role of terminology in general dictionaries from various aspects, including conducting surveys to identify what dictionary users want and need as far as terms are concerned. The empirical results of these surveys are summarized in Josselin-Leray and Roberts (2007), and among other findings reveal that users place a high value on the exhaustiveness of a dictionary, that they generally expect monolingual dictionaries to contain more terminology that bilingual ones, but also that French users on average place a significantly higher value on the presence of terminology in dictionaries than English users.

To date, no similar study has been performed in Slovenia, but in order to shed some light on user information needs regarding specialised vocabulary we performed an analysis of search query logs for the Termania.net dictionary portal. Termania.net is a free dictionary aggregation portal created by Amebis d.o.o. in 2010, and represents one of the largest online resources for terminology from various domains. The portal provides unified access to 44 dictionaries, of which approximately half are categorised as general. These include access to the DSLL, the Slovene lexical database, various dictionaries of rhymes, abbreviations, dialects and several bilingual dictionaries. Specialised resources include small, medium and large dictionaries from numerous domains, including education, medicine, biology, IT, and tourism.

Since the Termania.net site is very well-known among translators, students and the general public, its search query logs might provide useful insights into the information needs of Slovene users. For the purpose of this analysis, Amebis d.o.o. provided logs for the past two and a half years ordered by frequency. The number of all queries was over 6.5 million, with 433,692 different ones, of which 287,283 occur only once. Unfortunately the logs do not tell us whether the query was successful nor do they reveal any information about the users.

Our main interest was to see whether Termania users search for terminology, but also to inspect query types (single- or multi-word, use of wildcards, etc.) and gain insights into their information needs. Since it would have been impossible to manually inspect all of these we automatically compared the list to the following resources:

- the Gigafida reference corpus of Slovene,
- the EN1010 web-crawled corpus of English,
- the first edition of the *Dictionary of Slovene Literary Language* (DSLL1),
- the second edition of the *Dictionary of Slovene Literary Language* (DSLL2).

Table 3: Checking the presence of Termania search queries in other resources

	Število poizvedb	Odstotek poizvedb
Termania	**433.692**	**100%**
Termania - 1x	287.283	66.3%
is in GF	177.687	40.1%
is in DSLL1	61.393	14%
is in DSLL2	64.348	15%
is in EN1010	114.044	26%

Table 3 presents the results of these comparisons. Our first observation was that over a quarter of all queries can be found in the English corpus, which leads to believe that Termania is frequently referred to as a bilingual resource. About 40 percent of queries can be found in Gigafida, but since the comparison was performed with the list of lemmata, this number might be slightly higher because users occasionally search for inflected word forms. Only 14 percent of different queries can be found in the DSLL, although the sum of these queries amounts to 2.6 million, which is around 40% of all queries. As many as 94,891 queries are multi-word, of which the majority are terms.

The queries not occurring in any of the compared resources can be categorised as follows:

- Slovene words in inflected forms, rarely terminological: *podatki, iščem, priljubljena, spremljaj,* and *zadržano*

- English words in inflected forms: *prospecting, sieving, levying, inventorying,* and *garnishing*

- Slovene specialised terms, especially borrowings: *hipersoničen, ekhimoza, acianotičen, transudat, distenzija, mezotelij,* and *hipersplenizem*

- Queries containing an asterisk: *an*, k*, turist*, pos*, spoln*,* and *hidroksi**

- Searching for word suffixes using a hyphen (not supported by the Termania search engine and therefore always unsuccessful): *-olg, -okate, -njiva,* and *-njice*

- Abbreviations, acronyms: *crh, ToR, ZAZV, Tfc,* and *accn*

- Words from languages other than Slovene or English: *einrichtung, belegen, verteilen, stellung, ausgleich, Spannungsversorgung, knikken, gebruiken, plutajuči,* and *το θηρζον*

- Other: proper names, expressions in brackets or quotation marks, misspellings, numbers, symbols, and nonsensical character strings.

The analysis shows that terminology constitutes an important part of all queries on Termania, and thus that users frequently refer to an online portal to resolve their information needs. Cross-comparisons between different resources provide important clues for the new dictionary, for example by exploring queries not present in any of the Slovene dictionaries, but coming up in Gigafida with several hundred occurrences.

8   CONCLUSIONS

The purpose of this chapter was to explore the role of terminology in the new *Dictionary of Modern Slovene Language* from as many aspects as possible, first by positioning it with regard to existing lexicographic traditions on the one hand, and the aims of the new dictionary on the other, then by setting up a methodological framework of corpus-based terminology acquisition for the purposes of a general language dictionary, and finally by reflecting on the categorisation and lexicographic treatment of terms, in light of the target users' needs.

Clearly, to create an exhaustive dictionary with a high coverage of specialised terminology is an ambitious goal in itself, and to achieve it the methodological considerations presented above will need to be continually refined, improved or modified, at all stages of the process. A concern for Slovene specialised language is a recurring theme in the national language policy, and has found its way into Slovene legislation and higher education. We hope that the *Dictionary of Modern Slovene Language*, with its comprehensive and non-exclusive coverage of terminology, will primarily help users communicate knowledge.

The potential of crowdsourcing in modern lexicography

Darja Fišer and Jaka Čibej

Abstract

Due to increasing volumes of linguistic data and time constraints, the nature of lexicographic work has changed significantly in the past two decades. A number of steps in dictionary production have already been automated, but the developed algorithms are still far from perfect. Dictionary construction thus still involves a number of routine but time-consuming and expensive manual procedures for which experienced lexicographers are overqualified. This is why contemporary lexicography has started to explore options such as crowdsourcing, which can save both time and financial resources without negative effects on the quality of the results, provided that key principles of microtask design and campaign management, which are presented in this paper, are taken into account.

Keywords: crowdsourcing, microtask design, crowd motivation, quality control, legal and ethical aspects of crowdsourcing

1   INTRODUCTION

Over the past decade, the growing presence of the Internet and ever greater digitisation of work have led to numerous forms of online collaboration in which users contribute toward a common project. Aside from open-source projects (e.g. *Linux*) and collaborative initiatives (e.g. *Wikipedia*), these new forms of work also include crowdsourcing, a process in which a group of people (the crowd) contributes toward a specific goal by dividing the work load. Each individual takes on a small and manageable task that does not require much effort or time to complete, while the combined results represent a significant achievement (Howe, 2008). It is important to note that members of the crowd are typically amateurs, not field experts. Nevertheless, a number of crowdsourcing projects have shown that, with adequate support and task design, even non-experts are capable of solving tasks that were once the domain of experts. Modern technology and the wide availability of the Internet have made harnessing the potential of crowdsourcing increasingly simple and efficient.

The term *crowdsourcing* was first introduced by Jeff Howe in 2006, and has since been used to describe a wide range of work practices. In order to separate crowdsourcing from other forms of collaborative work, such as *co-creation* and *user innovation,* Estellés-Arolas and González-Ladrón-de-Guevara (2012) propose the following definition:

> "Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken." (ibid.: 9–10)

An essential element of a crowdsourcing project is its initiator, a company, organisation or individual who designs and manages the campaign as well as recruits crowdsourcers to perform the specified tasks. The crowdsourcers' contribution benefits both the initiator, who obtains valuable data, as well as the participants, who receive either monetary or some other type of compensation in return for their services.

In modern lexicography, the most widespread form of online user contribution is collaborative lexicography, which involves users providing new dictionary entries or suggesting updates of existing ones (Abel and Meyer 2013). The most

famous examples of collaborative dictionary projects are *Wiktionary*[1] and *Urban Dictionary*.[2] For Slovene, the best-known among such projects is *Razvezani jezik* (The Tongue Unleashed), but many smaller ones also exist, focusing mainly on collecting dialectal vocabularies.[3]

As the increasing automation of lexicographic work has been introduced to tackle stricter time constraints and increasing quantities of data, certain phases of dictionary construction have become routine tasks for which the lexicographers are overqualified. In this context, crowdsourcing has a lot of potential and can save valuable time, not as the main phase of dictionary construction, but as a way of post-processing, cleaning-up and validating automatically extracted data. It is therefore surprising that crowdsourcing has not been embraced by publishers and incorporated in the workflow of lexicographic projects. This is why the goal of this paper is to demonstrate successful implementations of crowdsourcing in related fields, as well as to outline the key principles of crowdsourcing task design and project management in lexicography.

2 CROWDSOURCING LANGUAGE DATA

In this section, we present an overview of related projects from various fields of natural language processing that have successfully implemented crowdsourcing.

2.1 Language resources

Klubička and Ljubešić (2014) used crowdsourcing to build an MSD-tagged and lemmatised corpus of Croatian to be used as a dataset. The evaluation of the crowdsourcing process showed that the accuracy of an individual crowdsourcer amounts to 90% on average, while the average accuracy of the majority vote of three crowdsourcers was approximately 97%.

Through their online application *Wordrobe*[4], Venhuizen et al. (2013) presented crowdsourcers with a series of tasks to annotate the *Groningen Meaning Bank*.[5] These tasks included homograph disambiguation, proper noun annotation and sense assignment to polysemous words. Compared to the gold standard, the results proved to be very reliable, even with only a small number of answers being collected.

---

1 http://sl.wiktionary.org/
2 http://www.urbandictionary.com/
3 http://razvezanijezik.org/
4 http://wordrobe.housing.rug.nl/
5 http://gmb.let.rug.nl/

Rumshisky (2011) and Rumshisky et al. (2012) used the *Amazon Mechanical Turk* crowdsourcing platform to build a semantically annotated corpus and semantic lexicon of English, both annotated by non-expert native speakers. The results showed that even non-experts can build a product with the same quality as one developed by experts.

Fossati et al. (2013) used the *CrowdFlower* crowdsourcing platform to annotate semantic roles in English texts. The comparison of crowdsourcing to standard annotation methods showed that the former, which divides the annotation process into several less complex steps, is faster as well as more accurate.

Fišer et al. (2014) used *sloWCrowd* (Tavčar et al. 2012), a custom-designed crowdsourcing tool, to clean up errors in the automatically constructed semantic lexicon sloWNet. With an average inter-annotator agreement of 80%, which is high for complex semantic tasks, the crowdsourcers showed a high degree of consensus, with very few ambiguous solutions.

Crowdsourcing was also used by Kosem et al. (2013) to validate and classify automatically extracted collocations and examples of use from the Slovene corpus *Gigafida*. The results of their experiment showed that a well-formulated, one-dimensional and objective question is crucial to achieving reliable crowdsourcing results.

Last but not least, the online game *Igra besed* (A Game of Words)[6] was designed to collect collocations by asking the player to suggest three typical adjective or noun collocates for a random noun or adjective. The suggestions are then scored according to the ranked list of collocations automatically extracted from the *Gigafida* corpus. The game offers a single-player mode as well as two two-player modes, either with a selected or a random player. The game collects the players' answers as well as their metadata (e.g. usernames, time of playing, co-player, points collected). The goal of the campaign is to identify the collocability measure that best represents the speakers' language intuition.

2.2   Language technologies

Crowdsourcing in language technologies was first embraced by machine translation researchers. Zaidan and Callison-Burch (2011) recruited crowdsourcers to vote for the best machine translation suggestion, and showed that the quality of the crowdsourced results rivals the work of professional translators. Crowdsourcing has also been successfully used to evaluate machine translation systems (Bentivogli et al. 2011; Denkowski and Lavie 2010), align texts in parallel corpora

---

6   http://www.igra-besed.si

(Gao and Vogel 2010) and collect datasets for statistical machine translation (Negri et al. 2011).

Chamberlain et al. (2008) use the online game *Phrase Detectives*[7] to crowdsource data for anaphora resolution. To attract more players, they have made their game available as a Facebook app.

Snow et al. (2008) conducted crowdsourcing campaigns on *Amazon Mechanical Turk* for a series of tasks, e.g. sentiment analysis in English newspaper titles. The evaluation of the data annotated by non-experts showed that as few as four annotations per task are required to achieve expert quality levels.

3   MICROTASK DESIGN

The basic concept of crowdsourcing is to divide a complex and large-scale problem into smaller, simpler and more manageable parts. The overall collection of activities that aim to provide a solution for the problem at hand is called the *crowdsourcing campaign*. The individual parts that are solved by crowdsourcers are called *microtasks*. Microtask design is a key phase in every crowdsourcing project. In this section, we provide an overview of the principles that need to be taken into account when designing microtasks in order to attain quality and useful crowdsourcing results. In addition, we provide several examples of well-designed microtasks.

3.1   Key principles of microtask design

Simplicity – Because microtasks are often undertaken by non-experts, it is important to keep the tasks as cognitively simple as possible, with the goal of collecting as many answers as possible in the shortest time period (cf. Rumshisky 2011; Snow et al. 2008; Lease and Alonso 2014).

Adequate questions – Microtasks should not include questions that cannot provide accurate and reliable results, such as opaque or ambiguous questions or overly subjective and unreliable estimates (cf. Kosem et al. 2013b). The questions posed need to be one-dimensional, which is why complex, multi-dimensional problems need to be split into several less complex steps (cf. Biemann and Nygaard 2010).

Adaptation to the target group – Different problems require varying levels of expertise. This needs to be taken into account when recruiting crowdsourcers

---

7   http://anawiki.essex.ac.uk/phrasedetectives/

(e.g. non-experts, students, or experts). Crowdsourcers with insufficient knowledge require more training (which is time-consuming and often destimulating) and will provide less reliable results. On the other hand, it is both demotivating and expensive to hire experts for trivial tasks.

Technical simplicity and user-friendly interface – User registration, login and task solving need to be logistically straightforward. They should not involve too much clicking or movement across the screen, and should avoid unnecessary manual data entry. This is already incorporated in most popular crowdsourcing platforms, but needs to be taken into account if using a custom-built one. Von Ahn and Dabbish (2008) stress the importance of bite-sized batches of micro-tasks that can easily be solved in a single sitting, while Jurgens and Navigli (2014) emphasise the importance of a user-friendly interface which does not rely on linguistic terminology.

Short instructions – The instructions for solving microtasks must be concise and should include illustrative examples.

Feedback – It is recommended to provide crowdsourcers with feedback for their answers. This allows them to improve and at the same time motivates them to continue their work.

Challenge, randomness and time restriction – Von Ahn and Dabbish (2008) make a number of recommendations for successful GWAP design that are relevant for other types of crowdsourcing campaigns. For example, the more the task is entertaining to solve, the more effective it is. It therefore needs to be designed in such a way that it presents a challenge to the player, e.g. by introducing scores, time restrictions, halls of fame, and so on. The task should also include an element of randomness, e.g. in assigning partners or questions. This prevents the tasks from being too predictable, while also eliminating the possibility of cheating.

3.2   Microtask examples

In this section, we present examples of various types of crowdsourcing tasks that have proved successful in related works.

3.2.1 Classical microtasks

Figure 1 shows an example of a microtask for semantic role annotation (Fossati et al. 2013). This consists of a brief instruction followed by a sentence in which the

crowdsourcer is asked to annotate the *agent* and *body part*. In this case, the correct answers are *he* and *none*.

Figure 1: Microtask for semantic role annotation.

Figure 2 shows an example of microtasks used for removing noise in the automatically generated sloWNet (Fišer et al. 2014b) in sloWCrowd (Tavčar et al. 2012). The crowdsourcer needs to confirm or reject an automatically assigned literal candidate (word or word phrase) to a wordnet synset using the English equivalents and definition provided. In this case, the correct answer is *no*.

Figure 2: Microtask for synset validation.

3.2.2 Microtasks through games with a purpose

Figure 3 shows the interface of *Puzzle Racer* (Jurgens and Navigli 2014), a GWAP in which the player drives a race car in order to gather coins and other rewards that lead to the collection of points. Before the start of the game, the player is presented with a hint in the form of three images (Figure 4). The player then needs to find out what the images have in common in order to solve the puzzle presented during the racing phase (Figure 5). In this case, the correct answer is *money*.

Figure 3: *Puzzle Racer*, **a game with a purpose.**
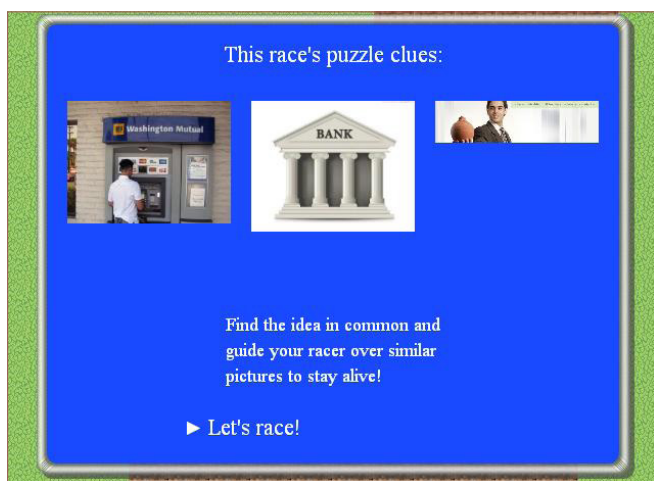
Figure 4: Hint in the *Puzzle Racer* **game.**
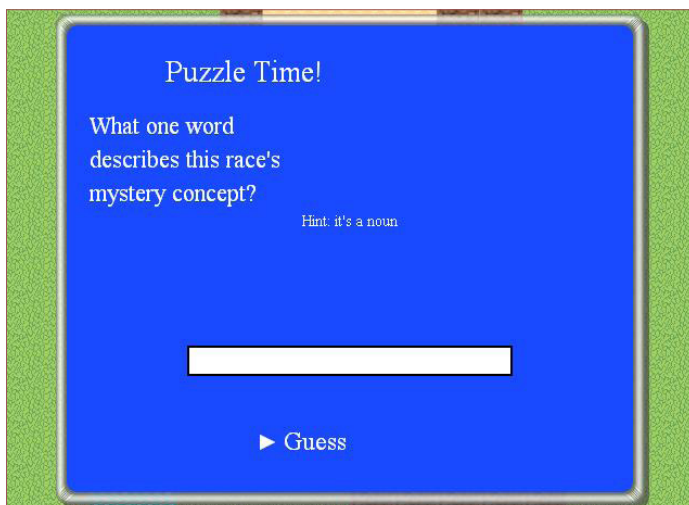
Figure 5: Riddle in the *Puzzle Racer* game.

Figure 6 shows the interface of *Igra besed* (A Game of Words). The player is presented with a word (in this case the adjective *gasilska*) and is required to provide three suggestions of typical collocations within 30 seconds. The game also has a two-player mode in which a player can compete against a chosen or random opponent. The player's answer yields points based on the word's ranking in the collocation lists from the Gigafida corpus of Slovene.
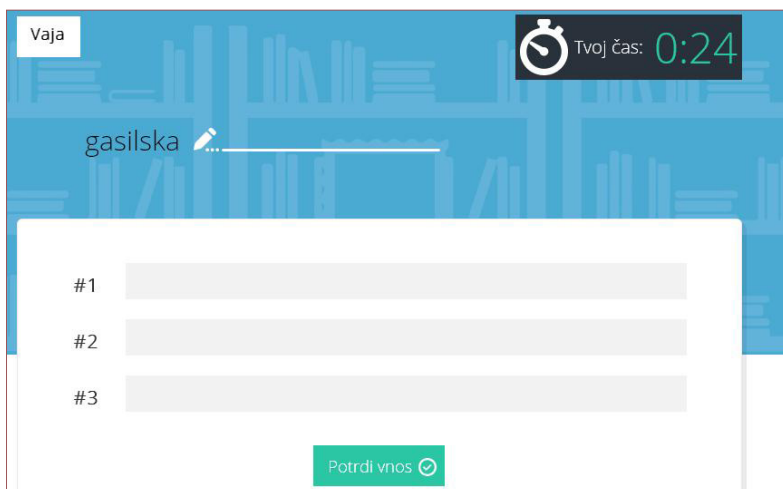
Figure 6: The interface for *Igra besed*.

3.2.3 Microtasks on social media

Games with a purpose can also be implemented on social media, where they are readily accessible to a large pool of users. Figure 7 shows a screenshot of the Facebook version of *Phrase Detectives* (Chamberlain et al. 2008). The player is presented with an example with two highlighted phrases, one of which refers to the second, and is asked to validate the correct annotations. Points are awarded according to inter-annotator agreement among the players who answered the same question. In this case, the correct answer is *agree*.

Figure 7: The Facebook version of *Phrase Detectives*.

4   MOTIVATIONAL ASPECTS OF CROWDSOURCING

Motivation is a crucial aspect of crowdsourcing projects, especially for languages with a limited pool of crowdsourcers. Crowdsourcing motivation can be either material or non-material, but it should always fulfil one or more of the crowdsourcers' needs, e.g. financial compensation, social recognition, confidence boost, or skill development. When discussing the motivation behind user-generated content on the web, Lew (2013) considers three categories of motivation: psychological, social, and economic. These are discussed in terms of crowdsourcing in this section.

4.1   Social motivation

The social aspect of motivation is based on the individual's need to connect and collaborate with other individuals sharing similar interests. This type of

collaboration enables all participants to gain new knowledge or skills and to improve their reputations in the community.

4.1.1 Community motivation

When motivating crowdsourcers, enthusiasm is much more important than the size of the community. It is important for the members to identify with the community and have the desire to contribute to its success, development, or recognition, and this has been the driving force of most collaborative lexicographic projects. For example, in the 10 years since the launch of *Razvezani jezik*, the Slovene collaborative dictionary of creative language use, approximately 1,600 anonymous authors have contributed more than 3,700 entry words and 2,300 dictionary entries (Dolar 2014). There is also substantial user involvement in language-related user groups on Facebook, which suggests that Slovenes could also be motivated to participate in the construction of Slovene language resources.

4.1.2 Educational motivation

Educational motivation is found in tasks that enable crowdsourcers to gain new knowledge or skills. This approach has been implemented by *Duolingo*[8] (von Ahn 2013), a website offering free language courses that consist of various tasks, e.g. translating sentences into foreign languages, which are then used to train models for machine translation of web content.

4.1.3 Acknowledgments and titles

Another type of social motivation are the acknowledgments that crowdsourcers receive as a reward for their work in the community. These range from certificates to prestigious titles (e.g. Wikipedia editor), or a mention in the community's hall of fame.

4.2 Psychological motivation

Many users find creating user-generated content psychologically fulfilling, either because they enjoy sharing their knowledge with others, because this allows them

---

8   https://www.duolingo.com/

to fulfil a need for self-expression, or simply because they find it entertaining. One of the examples that best utilise entertainment to obtain tangible results are games with a purpose (GWAP), an increasingly popular form of collaborative work/crowdsourcing. These rely on people with an internet connection and an interest in video games to improve search engine performance, filter content, and collect linguistic data through gameplay, tasks still too complex to be performed by computers.

Two early GWAP for collecting linguistic data were the *ESP Game* (von Ahn and Dabbish 2004) and *Peekaboom* (von Ahn et al. 2006b). In the *ESP Game*, a pair of random players are presented with an image. The goal of the game is to guess the word the second player will use to describe the image. The results data has proved useful in a number of ways, such as improving search engines and developing software for the visually impaired. The second game, *Peekaboom*, employs a similar concept: the players are presented with an image and asked to determine the position of an object. The collected data is then used to train computer vision models.

Other successful games with a purpose include *JeuxDeMots* (Joubert and Lafourcade 2012), a game designed to build a French lexical network; the aforementioned *Phrase Detectives* (Chamberlain et al. 2008); *Puzzle Racer* (Jurgens and Navigli 2014), a game for semantic image annotation; and *Verbosity* (von Ahn et al. 2006a), which uses questions or sentence completion tasks to collect the general knowledge data (e.g. statements such as *milk is white*) needed to build ontologies and improve the intelligence of computer systems.

4.3   Economic motivation

Economic motivation is based on monetary remuneration or other material rewards for completing tasks.

4.3.1 Micropayments

Monetary remuneration is commonly employed in large-scale (especially commercial) projects that expect crowdsourcers to complete a significant amount of work over a longer period of time. It is usually implemented in the form of *micropayments*, paid out for each task or a batch of tasks solved. Micropayments are supported by most popular crowdsourcing platforms, such as *Amazon Mechanical Turk*,[9] *CrowdFlower*,[10] and *Clickworker*.[11]

---

9   https://www.mturk.com/
10   http://www.crowdflower.com/
11   http://www.clickworker.com/en/

The crowdsourcing workflow involving micropayments is structured as follows: the crowdsourcing initiator uploads their project, consisting of a set of micro-tasks, on a crowdsourcing platform and transfers a certain amount of funds to the platform owner. The amount depends on project size, task complexity, the number of different tasks, and so on. The platform owner is entitled to a share for hosting the project, while the rest of the amount is distributed to crowdsourcers according to the number of tasks they complete.

Micropayments have been used in a number of linguistic crowdsourcing projects (Akkaya et al. 2010; Rumshisky 2011; Rumshisky et al. 2012; Fossati et al. 2013). However, it should be noted that certain platforms, e.g. Amazon Mechanical Turk, have their own pool of crowdsourcers (registered users allowed to solve tasks) that mainly consists of English speakers (or speakers of other large languages), and therefore cannot be used by researchers interested in smaller languages. In addition, local financing and tax legislation may pose additional restrictions for project financing and micropayment management.

4.3.2 Other rewards

Economic motivation may be achieved through other types of rewards, such as vouchers, tickets, software licences, or other tangible benefits (T-shirts, pins, etc.). It is commonly employed by small-scale projects (El-Haj et al. 2014; Fišer et al. 2015) with limited funding, often in combination with social and/or psychological motivation.

5   QUALITY CONTROL AND ASSURANCE

In this section we present some of the mechanisms used for direct or indirect quality assurance of crowdsourcing results and noise elimination, in order to overcome issues related to unclear instructions or unreliable crowdsourcers.

5.1   Gold standard

The most common method of quality control is the gold standard, a manually annotated dataset of microtasks that have been solved correctly by an expert. The gold standard microtasks are randomly added to the batches of genuine microtasks in order to measure the reliability of crowdsourcers and exclude the responses of unreliable individuals from the final results.

When forming the gold standard, it is necessary to take into consideration that it should be representative of the entirety of microtasks, both in scale and difficulty. If the gold standard is too simple, it cannot effectively separate reliable and unreliable crowdsourcers. On the other hand, a too complex gold standard will exclude too many crowdsourcers. In addition, when solving microtasks an appropriate balance of gold standard and genuine microtasks is necessary. Too few gold standard microtasks cannot reliably track a crowdsourcer's accuracy, while too many will be uneconomic, as this would mean that the crowdsourcers are provided with tasks that have already been solved.

5.2   Inter-annotator agreement

The second means of assuring the quality of crowdsourced data is via inter-annotator agreement. By presenting multiple crowdsourcers with the same microtask, several answers are available for every task. The distribution of the answers can then be used to calculate a confidence score for every individual microtask or crowdsourcer (Oyama et al. 2013). In this scenario the majority vote can be taken into account when making the final decision, which means that we accept the answer provided by the majority of the crowdsourcers.

Low inter-annotator agreement might indicate that the microtasks were not properly designed, assigned to crowdsourcers with insufficient knowledge, or presented with unclear annotation guidelines. The optimal number of times the same question is given to multiple crowdsourcers is very important in this context, as each further repetition increases the costs but does not provide new answers. For most tasks three annotations are required, while more complex tasks call for 5 repetitions (Krek et al. 2013b).

5.3   Refereeing

Refereeing is a process in which problematic examples that the crowdsourcers were unable to solve reliably are resolved by an expert referee. When microtasks and annotation guidelines are well-prepared, then referees are left only with a small portion of difficult tasks to resolve, while the bulk of the work is still crowdsourced. With regard to annotating a corpus of Croatian, Klubička and Ljubešić (2014) report that this process nearly halved the amount of work done by experts.

5.4 Crowdsourcer consistency

The final method of quality control is based on the concept of consistency, also called *intra-annotator agreement* (Gut and Bayerl 2004), which tracks the consistency of a crowdsourcer's answers when presented with the same microtask multiple times throughout the annotation session. If the answers to the same task differ to a great extent, then they are excluded from the final dataset, as the crowdsourcer is either not confident or knowledgeable enough, or is selecting random answers in order to increase their productivity.

6 LEGAL, FINANCIAL AND ETHICAL ASPECTS OF CROWDSOURCING

In this section, we present the legal, financial and ethical factors to be taken into account when setting up a crowdsourcing project. The related restrictions depend on local legislation, and although they do not directly influence project quality or content, they often present an obstacle to using crowdsourcing in research, especially in lexicography, where many researchers are not familiar with legal restrictions and rarely have access to expert legal advice. Moreover, as crowdsourcing is a relatively new form of work, it is not explicitly covered by the current legislation, which is why several issues remain unresolved.

6.1 Crowdsourcer payment

Sabou et al. (2014) appeal for ethical micropayments so that crowdsourcers' earnings correspond to the local average salary or hourly wage that is standard for the services provided. Fort et al. (2014) warn that the concept of crowdsourcing as a new form of work is still absent in work-related legislation, which puts crowdsourcers in a precarious position when it comes to payment, occupational safety, worker's rights, and so on. This remains the case, even though as many as 20% of workers on Amazon Mechanical Turk are said to earn their living exclusively through crowdsourcing. It is thus imperative that crowdsourcers are guaranteed fair and prompt payment, which is not always the case in many projects (Silberman et al. 2010).

Sabou et al. (2014) recommend a pilot task-solving campaign be executed before the actual crowdsourcing project in order to determine how long the project will take to complete. Certain tasks are more difficult and complex than others, and

as such require more input and time from the crowdsourcers. The micropayments should thus be higher with such tasks, in order to achieve a comparable hourly wage. This was taken into account by e.g. Krek et al. (2013), proposing a micropayment of 0.02€ for simple crowdsourcing tasks (with approximately 350 decisions per hour, this amounts to 7€) and 0.05€ for more difficult tasks (the hourly wage remains the same, as the number of decisions per hour is somewhat smaller). When paying crowdsourcers the existing payment methods provisioned by local legislation and potential restrictions in tax legislation need to be taken into account.

Considering the ethical aspects of crowdsourcing is particularly important if the collected data will be used for commercial purposes and will be of direct financial benefit to the project initiator. In such cases it is controversial to offer crowdsourcers no or very low payment.

6.2   Recruitment restrictions

When selecting crowdsourcers for the project, potential legal restrictions need to be taken into account. This is especially important if minors are involved, and in most countries parental consent needs to be obtained before they can participate in a crowdsourcing project.

6.3   Authorship recognition

As crowdsourcers often do a significant amount of work on a project, it is necessary to determine how and where they will be credited. Although there are no clear guidelines when it comes to authorship recognition in crowdsourcing, several volunteer projects (e.g. *FoldIt*,[12] *Phylo*[13]) have listed crowdsourcers as project co-authors.

6.4   Intended use and distribution licence

Before starting work on a project, it is common for crowdsourcers to sign a consent form that informs them of the intended use of the collected data. The consent form needs to make clear whether their contribution will be used for research

---

12   https://fold.it/portal/
13   http://phylo.cs.mcgill.ca/

purposes only or also for commercial ends, and whether it will be used in-house only or made available to third parties. If the crowdsourced data will be made publicly available, an adequate licence needs to be selected in accordance with local legislation on copyright and personal data protection.

7 CONCLUSION

Several projects related to the development of language resources and technologies have successfully implemented crowdsourcing, which indicates that this method could also prove useful in lexicography. However, all relevant aspects of this process need to be considered: from data preparation, microtask design and crowdsourcer recruitment, to ensuring crowdsourcer motivation and taking into account the legal, financial, and ethical restrictions of the project.

It is anticipated that crowdsourcing will soon be seen as a useful tool for lexicographers, one that will speed up the lexicographic process in a period of a growing demand for a the rapid processing of ever increasing amounts of linguistic data, as well as reduce the lexicographers' workload with regard to routine tasks, allowing them to focus on expert tasks. In addition, crowdsourced datasets will be useful for other purposes besides the construction of dictionaries, such as to help improve NLP tools through machine learning, with crowdsourced data as a high-quality training set. If adequately implemented, crowdsourcing could have a lasting impact on the workflow of future lexicographic projects, as well as on the use and life-cycle of lexicographic products.

Crowdsourcing workflows in lexicography

Darja Fišer and Jaka Čibej

Abstract

The success of a crowdsourcing campaign depends on a number of factors, e.g. an effective workflow, the funding available, the technological framework for crowdsourcing, the type of crowdsourcer motivation, and the type and volume of the data to be processed. Before embarking on a project it is therefore imperative to analyse its needs and plan the implementation of crowdsourcing that best fits the specific circumstances, to ensure the feasibility of the campaign and good results. In this paper we propose a general crowdsourcing workflow for lexicographic projects that can then be tailored to various scenarios. We also provide an overview of the most popular crowdsourcing platforms and discuss the criteria to be taken into account when selecting the one used for a specific lexicographic project.

Keywords: crowdsourcing, workflow, dictionary construction, crowdsourcing platforms

1   INTRODUCTION

Crowdsourcing has a lot of potential in contemporary lexicographic projects, especially as a means to post-process automatically extracted data and facilitate the work of lexicographers. Although crowdsourcing has not yet been thoroughly tested on large-scale lexicographic projects, a number of related projects have proved that it can be both sufficiently accurate and effective (cf. Klubička and Ljubešić 2014; Fišer et al. 2015; Kosem et al. 2013b). These encouraging results indicate that the power of the crowd could also be harnessed in the field of lexicography. However, each crowdsourcing campaign needs to take into account a number of external factors such as the budget and time available, the amount and type of data that needs to be processed, as well as the pool and type of crowdsourcers that can be recruited. In this paper we propose a general workflow for lexicographic projects, each step of which can be tailored to specific project circumstances. We then describe a set of crowdsourcing scenarios for the most common lexicographic project types, highlight the key principles that need to be taken into account and present the customized workflow for each of these. Finally, we give an overview of the most popular crowdsourcing platforms and present the criteria for choosing the best one for the lexicographic project at hand.

2   CROWDSOURCING WORKFLOW IN LEXICOGRAPHIC PROJECTS

In this section, we propose a general crowdsourcing workflow that can be used in various phases of corpus-based lexicographic projects. Our approach is modular and can therefore be adapted according to the needs of the project at hand. The order of the stages can be changed, some can be done in parallel or even left out, but it is important to at least consider the stages we recommend and address the issues each of them raise, as crowdsourcing is a complex, time-consuming and potentially costly procedure that cannot yield useful results without careful planning and task design.

Before deciding on a crowdsourcing campaign, an estimate of the required investment should be made with respect to the time, money and personnel required, as the campaign should not take up more time and financial and/ or human resources than conventional annotation methods. An important advantage of including crowdsourcing from the very beginning of dictionary project planning is the fact that the initial input in the preparation of an appropriate crowdsourcing environment pays off in the long run: crowdsourcing can be used in numerous phases of dictionary construction, microtasks can be

designed according to the same principles, and data can be annotated and processed using the same platform.

We describe the individual stages of the crowdsourcing workflow in the following sections.
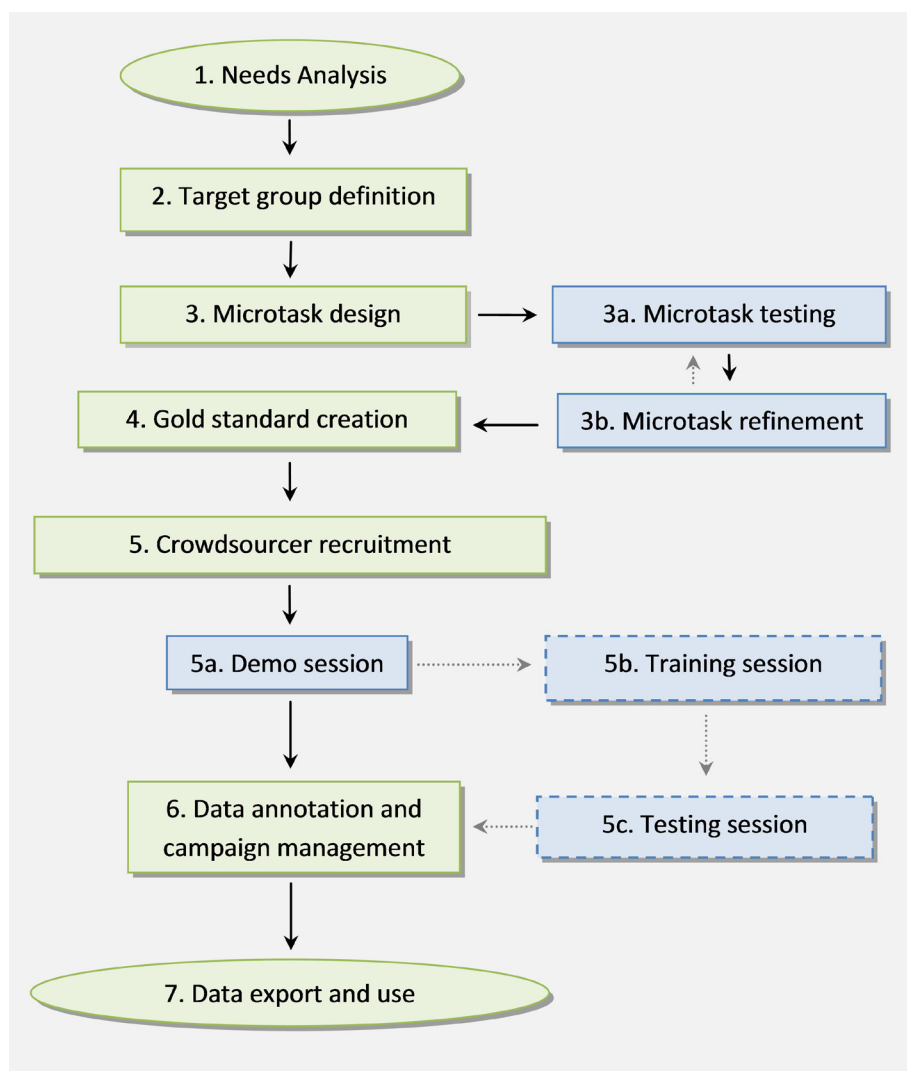
Figure 1: Crowdsourcing workflow for lexicography. Green-coloured boxes represent the main stages, while blue-coloured ones are the subphases. Dashed boxes and arrows represent optional stages which can be omitted in small-scale, low-budget campaigns.

2.1   Needs analysis

The first step of each crowdsourcing campaign requires a thorough needs analysis. Several things need to be determined: the goals and expectations of the campaign, the quantity of the data to be processed, the purpose for which it is to be used, as well as its format and availability. With dictionary projects, in which crowdsourcing can be used in different phases, it is recommended to analyse the needs of each phase and design the workflow, platform and timeline of the crowdsourcing campaign in such a way that it ensures compatibility of input data and software throughout the entire project.

2.2   Crowdsourcer profile

Once the needs have been analysed it is necessary to determine the required crowdsourcer profile, as tasks can vary in complexity and require different skills. The problem at hand may be suitable for the general public without any specialized linguistic or lexicographic knowledge, or it may require a certain degree of expertise and can only be solved effectively by language students or expert lexicographers.

3.3   Microtask design, testing and refinement

The most important and difficult part of crowdsourcing is microtask design. Microtasks should be one-dimensional questions with concise instructions and suited to the target crowdsourcer profile.

For instance, tasks aimed at the general public should not contain terminology or complex structures, which should be replaced with practical examples (e.g. the question "*Which meaning best corresponds to the use of the word in the phrase contained in the example?*" can be simplified to "*What is the meaning of the underlined word in the sentence below?*"). It is very important not to design microtasks in such a way that they yield unreliable results. This is especially problematic with multi-dimensional questions, as crowdsourcers will not be able to answer them accurately (e.g. the question "*Is the collocation below suitable to be included in a dictionary?*" can be divided into two parts: 1. "*Is the collocation below correctly extracted from the corpus?*" and 2. "*Does the collocation below fit into a learning dictionary?*").

The designed microtasks need to be tested in a pilot study in order to check their effectiveness, determine potential incongruences or mistakes, and eliminate all identified shortcomings. If a microtask turns out to be too complex for the chosen crowdsourcer profile, it needs to be adapted or reassigned to crowdsourcers with more expertise in the field.

2.4   Gold standard creation

A certain number of microtasks needs to be annotated by experts to create a gold standard that is later used to ensure the quality of the crowdsourced results. The dataset should be as representative of the entire set of microtasks as possible, both in terms of size and complexity.

2.5   Crowdsourcer recruitment and training

After microtasks have been designed and the gold standard created, it is time for crowdsourcer recruitment and training. The crowdsourcing initiator usually holds a **demo session**, either live or, most often, as a presentation or demo video that is made available on the project website. The demo session introduces crowdsourcers to the annotation process. This is followed by a **training session**, during which crowdsourcers solve tasks under the supervision of an expert who provides advice or further explanation. Alternatively, the training session can be held online with automatic feedback for every solved task. The final recruitment step is the **testing session**, during which crowdsourcers solve tasks independently and are recruited if they pass a given accuracy threshold. With low-budget projects, training and testing sessions are often combined with the main part of the campaign, while the unreliable results/crowdsourcers are excluded.

2.6   Data annotation and campaign management

This is the main stage of every crowdsourcing campaign, during which the recruited crowdsourcers solve the microtasks provided by the initiator. The initiator needs to monitor the campaign and decide whether any additional fine-tuning is necessary, e.g. whether the set of microtasks needs to be expanded, the crowdsourcers are motivated enough to provide a consistent flow of answers, and so on.

2.7   Data export and use

The final stage involves exporting the crowdsourced data into an appropriate format for further use in the project (e.g. for algorithm training or inclusion in a dictionary). The crowdsourcing platform should allow the data to be exported at any point of the crowdsourcing campaign, as checking whether interim results are meeting the expectations of the project is crucial for good campaign management.

3   TYPES OF LEXICOGRAPHIC PROJECTS

In this section, we present potential scenarios of implementing crowdsourcing into various types of lexicographic projects. As already emphasised, the flow of the crowdsourcing campaign depends a great deal on funding. Funding is directly related to the range and timeframe of the crowdsourcing campaign, the project phases in which crowdsourcing will be used, the number of microtask types designed, the complexity of the crowdsourcing workflow, the number of recruited crowdsourcers, and the type of motivation used. The more financial resources there are available, the more specialised the applications that can be developed, tested, optimised and finally presented to a wide circle of crowdsourcers. Low-budget projects, on the other hand, require more input when it comes to recruiting and motivating crowdsourcers. However, the social motivation of crowdsourcers can (and should) be used in all scenarios.
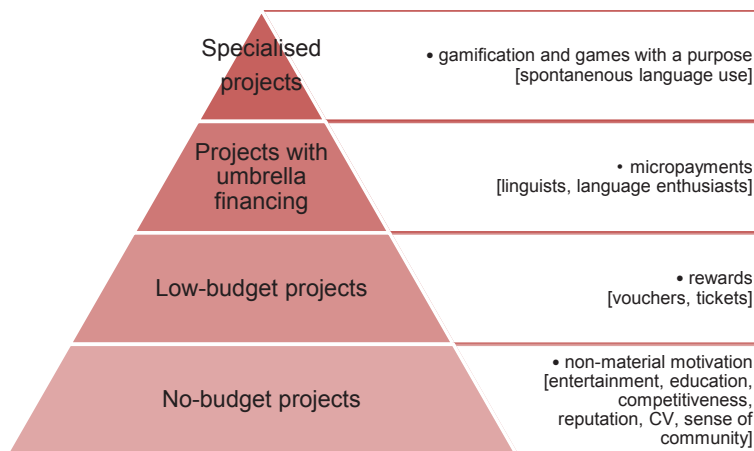
Figure 2: An overview of crowdsourcing scenarios for various types of lexicographic projects.

3.1 Specialised projects

Most specialised projects with full financing can afford tailor-made crowdsourcing applications, most notably games with a purpose (GWAPs). Their entertaining and competitive elements ensure three long-term interest of a wide group of players and spontaneous language use. GWAPs have proved highly successful in a number of related projects (cf. Jurgens and Navigli 2014; Joubert and Lafourcade 2012; Chamberlain et al. 2008). Jurgens and Navigli (2014) found that *Puzzle Racer*, a game that involves players annotating corpus data, achieves the same level of quality as conventional data annotation by experts, while lowering the overall costs by 73% compared to a classical crowdsourcing campaign involving microtasks. A specialised GWAP can be used to collect large quantities of data, can be adapted for different devices and platforms, and allows for the inclusion of different tasks for different profiles and phases of the lexicographic project.

3.2 Projects with umbrella financing

Many contemporary lexicographic projects have no direct funding and are instead realised as one of the non-primary activities of a wider research project or programme. In this scenario, it is recommended to use existing resources and technologies to plan a crowdsourcing campaign in such a way that the results can be directly applicable, not only in the context of the lexicographic project, but also to the main project and any future projects that arise. The resources to develop customized applications are most likely not available, but any of the popular crowdsourcing platforms can be used. The campaign should attract lexicographers, language editors, translators and language enthusiasts who can be paid through micropayments. The number of microtasks, the quantity of the crowdsourced data, and the number of crowdsourcers should correspond to the available financial resources. If necessary, optional phases (e.g. cyclic microtask editing, crowdsourcer training and testing) can be left out of the workflow (see Figure 1).

3.3 Low-budget projects

In low-budget projects it is recommended to invest the majority of the financial resources available in automating data preparation as much as possible, while also significantly simplifying the crowdsourcing workflow. In this scenario, the default

quality control parameters can be used and the majority vote without an expert referee can be used to make final annotation decisions. The crowdsourcers best suited for this scenario are students of linguistics or language enthusiasts, whose work can be rewarded with vouchers, tickets or other material rewards. This approach has already proved to be feasible (El-Haj et al. 2014; Fišer et al. 2015). However, it demands realistic expectations when it comes to crowdsourcer input in terms of time and effort. The crowdsourcers should not be presented with overly ambitious tasks, nor should they be expected to do a significant amount of work in a short period of time – a fact that needs to be taken into consideration when planning the project. Instead, a longer campaign should be foreseen compared to scenarios with funding.

3.4   No-budget projects

In cases when no financial resources are available, crowdsourcing can be implemented in a manner similar to that employed by numerous collaborative lexicographic projects that recruit and motivate crowdsourcers with non-material rewards based on social motivation. Aside from enthusiasts who enjoy contributing to the construction of new language resources, the wider public can also be motivated to join the campaign if offered entertaining tasks or organised competitions (Fišer et al. 2015). In addition, students and graduates can be recruited by offering awards or certificates for participating in the project, which they can use for extra-credit or as a reference to add to their CV.

As with the low-budget scenario, it is crucial to plan a no-budget crowdsourcing campaign as a long-term project. Crowdsourcers should only be given simple tasks, and the project should be relevant for their community. It is also necessary to take into consideration the fact that the crowdsourcers involved are participating out of enthusiasm for the project, which is why it is even more important to keep in touch with them regularly and build a well-connected community.

4   CROWDSOURCING PLATFORM SELECTION

A crowdsourcing platform is an application that allows the crowdsourcing initiator to upload a project containing microtasks that are then solved by the recruited crowdsourcers. In this section, we describe the criteria to follow when selecting an appropriate platform, as well as the process of choosing the platform that is to be used for crowdsourcing the *Dictionary of Modern Slovene Language*.

4.1   Selection criteria

The selection of a suitable platform is one of the first steps to undertake in a crowdsourcing campaign. Several criteria need to be taken into account.

Data format – The platform needs to support uploading different types of microtasks and exporting crowdsourcing results in formats suitable for the needs of the project.

Interface – It is important for the platform to offer a simple, user-friendly interface both for the campaign administrator and the crowdsourcers. The administrator should be able to use the platform to form different types of tasks of varying complexity, to monitor the statistics of data collection and crowdsourcer reliability, to expand the gold standard if necessary (without interrupting the crowdsourcing process), or update the set of microtasks and export preliminary results. The crowdsourcers, on the other hand, should be provided with a simple registration process (e.g. using a Gmail, Twitter or Facebook account), personal data protection, and a comfortable working environment that increases their motivation.

Quality control – It is important to make sure that the platform contains as many quality control measures as possible, e.g. a gold standard, inter-annotator agreement, crowdsourcer consistency, and majority vote. In addition, the platform should allow the administrator to fine-tune the settings that control the inclusion of gold standard microtasks into crowdsourcer tasks, repeating the same microtasks with multiple crowdsourcers, the time restriction for individual tasks, and so on.

Financial aspect – The platform needs to support micropayments if this type of economic motivation is to be used to motivate crowdsourcers. With commercial crowdsourcing platforms that offer campaign hosting, the amount the crowdsourcing initiator needs to transfer depends on the size and complexity of the campaign. The bulk of the resources are used for micropayments (their size is usually determined by the crowdsourcing initiator), while a certain percentage is taken by the platform manager.

Motivational mechanisms – It is advantageous if the platform already incorporates mechanisms for additional crowdsourcer motivation, e.g. a scoring system, hall of fame, automatic notifications when someone beats the current high score, and automatic reminders for crowdsourcers who have been inactive for longer periods of time.

4.2   Overview of crowdsourcing platforms

When selecting a platform for the construction of a dictionary of modern Slovene, we reviewed approximately 150 crowdsourcing platforms between October and November 2014. In the following sections, we list and describe those that are suitable for crowdsourcing linguistic data.

4.2.1 Commercial platforms

The most popular crowdsourcing platform is Amazon Mechanical Turk.[1] Its interface already incorporates mechanisms for quality control, campaign management and micropayment support. The platform also has a large existing pool of registered crowdsourcers. However, they are mostly speakers of larger languages.

Similar examples are Crowdflower[2] and Clickworker.[3] Both offer a number of applications for various fields of data processing (e.g. data categorisation and sentiment analysis). Microtasks can be uploaded in CML, CSS or Javascript, and crowdsourcers can be filtered according to their age, knowledge prerequisites and geolocation.

4.2.2 Open-source platforms

The most notable open-source platform is Crowdcrafting,[4] which recruits volunteer crowdsourcers to contribute to various research projects by solving tasks. The platform is based on PyBossa,[5] an open-access software for creating crowdsourcing projects that can be installed on a local server and is available under the Creative Commons BY-SA 4.0 licence.

Another open-source crowdsourcing tool is sloWCrowd[6] (Tavčar et al. 2012), which is PHP/MySQL-based and was developed for cleaning automatically generated sloWnet synsets but later extended to enable other types of crowdsourcing tasks (Fišer et al. 2015).

---

1   https://www.mturk.com
2   http://www.crowdflower.com
3   http://www.clickworker.com/
4   http://crowdcrafting.org/
5   http://pybossa.com/
6   http://nl.ijs.si/slowcrowd/

4.3   Platform selection for the construction of the *Dictionary of Modern Slovene*

After reviewing the existing crowdsourcing platforms, we decided to use PyBossa for the crowdsourcing of the *Dictionary of Modern Slovene*. The reasons for this are as follows:

Flexibility – Unlike commercial platforms, PyBossa can be installed on a local server, and its interface can be adapted to the needs and conditions of the project.

Support – As an open-source platform, PyBossa is well supported and constantly developed. It has already been successfully used in numerous projects, and many additional libraries are available to enable more mechanisms for monitoring the results of the crowdsourcing process, and other outcomes.

Financial independence – In case of insufficient funds, paying crowdsourcers with micropayments will not be possible. Commercial platforms do not offer other types of payment (rewards, tickets, etc.). In addition, using an open-source platform will save the commission that needs to be paid to professional crowdsourcing platforms for handling the micropayments.

Logistical reasons – There are a number of technical barriers when dealing with commercial platforms. For instance, Amazon Mechanical Turk requires the crowdsourcing initiator to have a bank account in the US. In addition, the platform would require registration and personal data from every Slovene crowdsourcer, which is very inconvenient. Difficulties would probably arise with micropayments as well, as the spending of public funds is strictly regulated in Slovenia.

Best practice – PyBossa has already been successfully used for crowdsourcing in numerous research projects. The platform's website[7] lists a number of users, e.g. the British Museum, the Swiss Research Institute CERN, and UNITAR.

5   LIMITATIONS OF CROWDSOURCING

Despite the great deal of attention crowdsourcing has recently received among lexicographers, misconceptions and prejudices about it are still common. We address these issues in this section.

To ensure the appropriate role of crowdsourcing in lexicographic projects, it is imperative to recognise its limitations as well as its potential. Crowdsourcing is

---

7   http://crowdcrafting.org/about

not effective for every type of data, every phase of lexicographic work or every lexicographic project. For instance, it cannot be implemented unless regular campaign management can be guaranteed (designing microtasks, controlling the collected answers, motivating and paying crowdsourcers). Crowdsourcing is also not suitable for open-ended questions or tasks that require subjective answers. It can only be useful when it saves time and/or financial resources for the lexicographic project, despite all the preparation and management that it warrants, while still providing reliable results.

5.1   Amateur lexicographers

Because certain authors are somewhat inconsistent when defining crowdsourcing (cf. Estellés-Arolas and González-Ladrón-de-Guevara 2012), it is often unjustifiably mistaken for – or even equated with – collaborative lexicography. Unlike numerous collaborative projects in which all the work is done by non-experts, a project initiator or manager is always heavily involved in crowdsourcing by preparing data, designing microtasks, controlling quality, ensuring crowdsourcer motivation, and so on. Although some collaborative projects have shown that users can also contribute to useful and widely used dictionary products (Meyer and Gurevych 2012), crowdsourcing as proposed in this paper primarily involves post-processing automatically extracted corpus and lexicon data before actual dictionary construction. Furthermore, user contributions are not immediately published as the content of the dictionary, and the organisation of lexicographic information is still controlled by the lexicographers.

Meyer and Gurevych (2012) pointed out that collaborative projects represent the sum of the opinions of numerous authors, who put considerable effort into improving dictionary entries until a consensus is reached on their structure and content. For this reason, collaborative lexicography in many respects gives results comparable to official lexicographic products. However, its biggest shortcoming is the lack of an effective mechanism to separate mature and high-quality dictionary entries from those that still require improvement. A similar observation was made by Lew (2013), who noted that in certain cases the order of definitions in Wiktionary can be somewhat random, with completely marginal meanings displayed at the top. A similar issue is found in Urban Dictionary, where users can vote to influence the order of definitions, and the most popular definition is not necessarily the most appropriate, but rather one that best reflects the users' ideology or the one they find most entertaining.

In contrast, we envisage crowdsourcing as one of the phases of dictionary construction. First, data is automatically extracted from corpora and other datasets.

The data is then post-processed by crowdsourcers through microtasks and finally used by lexicographers in manual lexicographic work. Crowdsourcing is thus an intermediate link between automatic data processing and manual expert data processing, as it significantly reduces the lexicographer's workload through automatic data extraction and crowdsourcing, while also including a manual approach in processes that still cannot be automated effectively. Although crowdsourcing has not yet been thoroughly tested on large-scale lexicographic work, the results of related projects have proved effective (Klubička and Ljubešić 2014; Fišer et al. 2015; Kosem et al. 2013b) and indicate that it can be successfully implemented in the field of lexicography.

5.2   Reliability of crowdsourced dictionaries

A common misconception about crowdsourced results is that they are unreliable, especially because the pool of crowdsourcers can include non-experts. We emphasise that microtasks should always be designed for a specific crowdsourcer profile and take into account their level of expertise. A well-designed crowdsourcing project will assign more complex tasks to crowdsourcers with more expertise in the field (e.g. students or graduates of linguistics), while only simple tasks will be left to non-experts.

Fišer and Čibej (2015) presented a number of quality control mechanisms, e.g. a gold standard, inter-annotator agreement, majority vote, consistency, and refereeing. These can be used to effectively eliminate those crowdsourcers that provide incorrect or unreliable answers. These quality control measures have already been tested by numerous authors (cf. Rumshisky 2011; Fišer et al. 2015; Klubička and Ljubešić 2014; Fossati et al. 2013), and found to ensure high accuracy of crowdsourcing results that achieve the same level of quality as if the work were done only by experts (Snow et al. 2008).

5.3   Impact of crowdsourcing on lexicography

As a new form of work not yet explicitly covered by legislation, crowdsourcing undoubtedly raises many ethical issues regarding payment, work conditions and authorship recognition. Crowdsourcing platform providers act as employment agencies, but it is the crowdsourcing initiators who determine payment conditions and the work load. Although crowdsourcers are not obligated to accept badly paid tasks, they are often forced to if they want to earn a living. Low

payments and unfair work practices in language resource crowdsourcing have been criticized by several authors (Sabou et al. 2014; Silberman et al. 2010; Lease and Alonso 2014; Felstiner 2011). For example, e.g. Snow et al. (2008) offered a total of 2$ for 7,000 non-expert annotations, and 1$ for 1,500 expert annotations via Amazon Mechanical Turk. It is thus the duty of the coordinators of every lexicographic project to treat crowdsourcers fairly and credit their contributions to the final product. Their pay needs to be taken into account at the very inception of the project, when the budget is determined.

In addition to issues stemming from payment and work conditions, crowdsourcing has also faced accusations that it degrades the profession of lexicographers and linguistics, redirecting their workload to an unqualified (and poorly paid) crowd. We wish to emphasise that the basic concept of crowdsourcing in this context is the rational use of resources: expert lexicographers are spared trivial and routine tasks, and crowdsourcers can contribute to language resource construction as best they can, while at the same time receiving different forms of motivation (monetary or material rewards, gaining experience and references, entertainment, etc.).

Because the misconceptions surrounding crowdsourcing campaigns are not only present among experts, but also in the general public, it is important for the crowdsourcing initiator to form an intelligent strategy for public relations. Communication with the potential crowdsourcers needs to be carried out with great care and respect, and the input expected from the crowd should reflect the type of motivation. For instance, if the crowd receives no monetary remuneration for their work then it should not be presented with overly ambitious tasks. It is also important for the crowdsourcing initiator to keep in touch with the crowdsourcers throughout the campaign, e.g. by informing them about the project workflow, inviting them to project presentations or similar events, and publicly thanking them for their contributions.

6   CONCLUSION

A well-planned crowdsourcing project that observes the key principles of microtask design and management can be of great help in lexicography, as it can handle the post-processing of automatically extracted noisy data in an economical and timely manner, with reliable results. This paper gave a comprehensive and detailed account of the organisational, technical, linguistic as well as financial aspects of successful crowdsourcing for dictionary creation, and proposed a general crowdsourcing workflow as well as several specialised scenarios that take into account various lexicographic project types and circumstances.

Crowdsourcing has already been embraced by language technologies and language resource creation. Recent successful small-scale specialized lexicographic projects have built a firm foundation for crowdsourcing to be included in more complex, large-scale lexicographic projects. The dictionary of modern Slovene is one of the first lexicographic projects that plans on implementing crowdsourcing in its entire workflow, and, as a pioneer project, it will pave the way for future dictionaries and language resources, both in Slovenia and abroad.

Bibliography
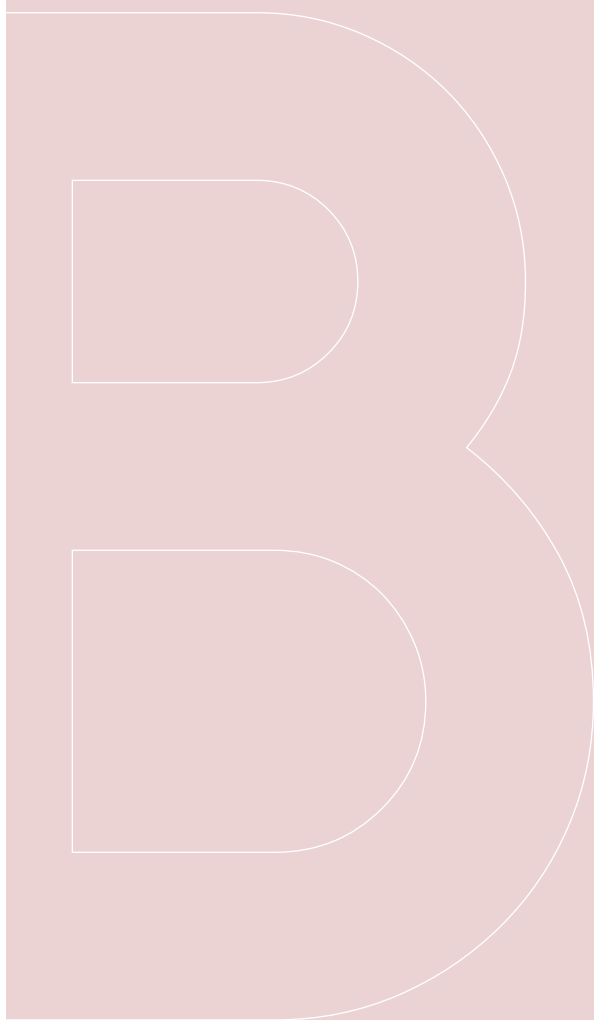
I

Abel, Andrea and Christian M. Meyer, 2013: The dynamics outside the paper: user contributions to online dictionaries. Kosem, Iztok, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets and Maria Tuulik (eds.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference.* Ljubljana: Trojina, Institute for Applied Slovene Studies in Tallinn: Eesti Keele Instituut. 179–194.

Ahn, Luis von and Laura Dabbish, 2004: Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems.* New York: ACM. 319–326.

Ahn, Luis von, 2006: Games with a Purpose. *Computer* 39/6. 92–94.

Ahn, Luis von, Mihir Kedia and Manuel Blum, 2006a: Verbosity: a game for collecting common-sense facts. Grinter, Rebecca, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries and Gary Olson (eds.): *Proceedings of the SIGCHI conference on Human Factors in computing systems.* New York: ACM. 75–78.

Ahn, Luis von, Ruoran Liu and Manuel Blum, 2006b: Peekaboom: A Game for Locating Objects in Images. Grinter, Rebecca, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries and Gary Olson (eds.): *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM. 55–64.

Ahn, Luis von and Laura Dabbish, 2008: Designing Games with a Purpose. *Communications of the ACM* 51/8. ACM. 58–67.

Ahn, Luis von, 2013. Duolingo: learn a language for free while helping to translate the web. *Proceedings of the international conference on Intelligent user interfaces.* New York: ACM. 1–2.

Akkaya, Cem, Alexander Conrad, Janyce Wiebe and Rada Mihalcea, 2010: Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation. *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk.* Los Angeles, California, ZDA. Association for Computational Linguistics. 195–203.

Al-Ajmi, Hashan, 2008: The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos* 18. 15–26.

Arhar Holdt, Špela and Vojko Gorjanc, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

Arhar, Špela and Peter Holozan, 2009: Leksikalna podatkovna zbirka ASES (Amebisov skupni elektronski slovar). Mikolič, Vesna (ed.): *Jezikovni korpusi v medkulturni komunikaciji.* Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales in Zgodovinsko društvo za južno primorsko. 30−51.

Arhar, Špela, 2009: Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54/3-4. 43–56.

Arhar Holdt, Špela, Gaja Červ, Polona Gantar, Iztok Kosem, Karmen Kosem, Irena Krapš Vodopivec, Simon Krek, Sara Može, Tadeja Rozman, Ana Marija Sobočan, Mojca Stritar Kučuk and Ana Zwitter Vitez, 2013a: *Pedagoški slovnični portal.* Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://slovnica.slovenščina.eu/.

Arhar Holdt, Špela and Tadeja Rozman, 2015: *Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov.* Smolej, Mojca (ed.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. 67–74.

Atkins, B. T. Sue and Antonio Zampolli (eds.), 1994: *Computational approaches to the lexicon*. Oxford: Oxford University Press.

Atkins, B. T. Sue and Michael Rundell, 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Aust, Ronald, Mary Jane Kelley and Warren Roby, 1993: The Use of Hyper-Reference and Conventional Dictionaries. *Educational Technology Research and Development* 41/4. 63–73.

Baroni, Marco and Silvia Bernardini, 2004: BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*. Lizbona. 1313–1316.

Baroni, Marco, Adam Kilgarriff and Jan Pomikálek, 2006: WebBootCaT: Instant Domain-Specific Corpora to Support Human Translators. *Proceedings of EAMT*. Oslo. 247–252.

Battenburg, John, 1989: *A Study of English Monolingual Learners' Dictionaries and their Users*. Ph.D. Dissertation. Purdue University, IN, ZDA.

Béjoint, Henri, 1981: The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* II/3. 207–222.

Béjoint, Henri, 1988: Scientific and technical words in general dictionaries. *International journal of lexicography* 1/4. 354–368.

Béjoint, Henri, 2000: *Modern Lexicography. An Introduction*. Oxford: Oxford University Press.

Bentivogli, Luisa, Marcello Federico, Giovanni Moretti and Michael Paul, 2011: Getting expert quality from the crowd for machine translation evaluation. *Proceedings of the 13th Machine Translation Summit*. Xiamen, Kitajska. 521–528.

Bergenholtz, Henning and Mia Johnsen, 2005: Log files as a tool for improving Internet dictionaries. *Hermes. Journal of Linguistics* 34. 117–141.

Biemann, Chris and Valerie Nygaard, 2010: Crowdsourcing WordNet. *Proceedings* of the *5th Global WordNet Conference*. Mumbaj, Indija.

Bishop, Jonathan, 2009: Enhancing the Understanding of Genres of Web-Based Communities: the Role of the Ecological Cognition Framework. *Int. J. Web Based Communities* 5/1. 4–17.

Bizjak Končar, Aleksandra, Helena Dobrovoljc, Kaja Dobrovoljc, Nataša Logar Berginc, Polonca Kocjančič, Simon Krek and Tadeja Rozman, 2011: Slogovni priročnik – projekt *»Sporazumevanje v slovenskem jeziku« – Kazalnik 17*. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik17/Kazalnik_17_Slogovni_prirocnik_SSJ.pdf.

Blei, David M., Andrew Y. Ng and Michael I. Jordan, 2003: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3. 993–1022.

Bogaards, Paul, 1998: What Type of Words do Language Learners Look Up? Atkins, B. T. Sue (ed.): *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag. 151–157.

Boguraev, Bran and Ted Briscoe, 1987: Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE. *Computational Linguistics* 13/3-4. 203–18.

Boguraev, Bran and Ted Briscoe (eds.), 1989: *Computational Lexicography for Natural Language Processing*. London and New York: Longman.

Boonmoh, Atipat, 2012: E-dictionary Use under the Spotlight. Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos* 22. 43–68.

Boulanger, Jean-Claude and Marie-Claude L'Homme, 1991: Les technolectes dans la pratique dictionnairique générale. Quelques fragments d'une culture. *Meta: Journal des traducteurs* 36/1. 23–40.

Boulanger, Jean-Claude, 1996: Les Dictionnaires généraux monolingues, une voie royale pour les technolectes. *TradTerm* 3. 137–151.

Brants, Thorsten, 2000: TnT: a statistical part-of-speech tagger. *Proceedings of the sixth conference on Applied natural language processing*. Seattle, Washington: Association for Computational Linguistics. 224–231.

Buzássyová, Klára, 2009: Slovar sodobnega slovaškega jezika (Z vidika zasnove in organizacije dela). Perdih, Andrej (ed.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 119–124.

Caluwe, Johan de and Johan Taeldeman, 2003: Morphology in dictionaries. Sterkenburg, Piet van (ed.)*: A practical guide to lexicography*. Amsterdam and Philadelphia: John Benjamins. 114–126.

Chamberlain, Jon, Massimo Poesio and Udo Kruschwitz, 2008: Phrase Detectives: A Web-based collaborative annotation game. Ghidini, Chiara, Axel-Cyrille Ngonga Ngomo, Stefanie Lindstaedt and Tassilo Pellegrini (eds.): *Proceedings of the 7th International Conference on Semantic* Systems, I-*SEMANTICS*. New York: ACM.

Chen, Yuzhen, 2010: Dictionary use and EFL learning. A contrastive study of pocket electronic dictionaries and paper dictionaries. *International Journal of Lexicography* 23/3. 275–306.

Corris, Miriam, Christopher Manning, Susan Poetsch and Jane Simpson, 2000: Bilingual Dictionaries for Australian Languages: User studies on the place of paper and electronic dictionaries. Heid, Ulrich, Stefan Evert, Egbert Lehmann and Christian Rohrer (eds.): *Proceedings of the Ninth EURALEX International Congress, Stuttgart, Germany, August 8th-12th 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung. 169–181.

Crowston, Kevin, 2010: Internet Genres. *Encyclopedia of Library and Information Sciences*. New York: CRC Press. http://crowston.syr.edu/sites/crowston.syr.edu/files/elis-chapter.pdf.

Crystal, David, 2001 (2006): *Language and the Internet*. Cambridge: Cambridge University Press.

Čebulj, Monika, 2013: *Raba slovarja v 1. in 2. triletju osnovne šole*. Diplomsko delo. Ljubljana: Pedagoška fakulteta UL.

Čibej, Jaka, Darja Fišer and Iztok Kosem, 2015: The role of crowdsourcing in lexicography. Kosem, Iztok, Miloš Jakubiček, Jelena Kallas and Simon Krek (eds.): *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of eLex 2015, 11–13 August 2015, Herstmonceux Castle, UK*. Ljubljana and Brighton: Trojina, Institute for Applied Slovene Studies in Lexical Computing Ltd. 70–83.

De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe and Sarah Hillewaert, 2006: Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. *Lexikos* 16/1. 67–83.

Denkowski, Michael and Alon Lavie, 2010: Exploring normalization techniques for human judgments of machine translation adequacy collected using Amazon Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10*. Stroudsburg, ZDA: Association for Computational Linguistics. 57–61.

Dobrovoljc, Helena and Simon Krek, 2011: Normativne zadrege – empirični pristop. Kranjc, Simona (ed.): *Meddisciplinarnost v slovenistiki. Obdobja* 30. Ljubljana: Znanstvena založba Filozofske fakultete UL. 89–97.

Dobrovoljc, Kaja, Simon Krek and Jan Rupnik, 2012: Skladenjski razčlenjevalnik za slovenščino. *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 42–47.

Dobrovoljc, Kaja, Simon Krek, Peter Holozan, Tomaž Erjavec and Miro Romih, 2013: Morphological lexicon Sloleks 1.2. *Slovenian language resource repository CLARIN.SI.* http://hdl.handle.net/11356/1039.

Dobrovoljc, Kaja and Simon Krek, 2013: Spletni portal Slogovni priročnik: luščenje in prikaz podatkov o jezikovni rabi. Žele, Andreja (ed.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 101–107.

Dobrovoljc, Kaja, 2014: Re-evaluating morphological dictionaries: the case of adverbs in Slovene. *International NooJ 2014 Conference*. Sassari, Italia.

Dobrovoljc, Kaja, Simon Krek and Tomaž Erjavec, 2015: Leksikon besednih oblik Sloleks in smernice njegovega razvoja. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 80–105.

Dobrovoljc, Kaja, 2015: Oblikoslovne informacije v sodobnih slovarskih priročnikih. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 64–79.

Dolar, Kaja, 2014: Kolaborativni slovar Razvezani jezik. *Slavistična revija* 62/2. 235–252.

Domingo, David and Ari Heinonen, 2008: Weblogs and Journalism: a Typology to Explore the Blurring Boundaries. *Nordicom Review* 29/1. 3–15.

Dubois, Claude, 1990: Considérations générales sur l'organisation du travail lexicographique. Hausmann, Franz J., Oskar Reichmann, Herbert Ernst Wiegand and Ladislav Zgusta (eds.): *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin and New York: de Gruyter. 1574–1588.

Dziemianko, Anna, 2010: Paper or electronic? The role of dictionary form in language reception, production and the retention of meaning and collocations. *International Journal of Lexicography* 23/3. 257–273.

El-Haj, Mahmoud, Udo Kruschwitz and Chris Fox, 2014: Creating Language Resources for Under-resourced Languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation*. Springer.

Epple, Barbara, 2000: Sexismus in Wörterbüchern. Heid, Ulrich, Stefan Evert, Egbert Lehmann and Christian Rohrer (eds.): *Proceedings of the 9th EURALEX International Congress*. Stuttgart: Universität Stuttgart. 739–749.

Erjavec, Tomaž, Nancy Ide, Vladimir Petkević and Jean Véronis, 1995: Multilingual Text Tools and Corpora for Central and Eastern European Languages. *Language resources for language technology: proceedings of the first European seminar*. Tihany, Madžarska.

Erjavec, Tomaž, Vojko Gorjanc and Marko Stabej, 1998: Korpus FIDA. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik konference Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 124–127.

Erjavec, Tomaž, 1998: Oznake korpusa FIDA. Štrukelj, Inka (ed.): *Jezik za danes in jutri.* Ljubljana: Društvo za uporabno jezikoslovje Slovenije in Inštitut za narodnostna vprašanja. 85–95.

Erjavec, Tomaž, Matija Ogrin and Jože Faganel, 2004: E-Slomšek: a TEI encoding of a critical edition of 19th century Slovenian rhetoric prose. *New Technologies and standards: Digitization of national heritage*, junij 3-5, 2004, Beograd. Pregled Nacionalnog centra za digitalizaciju 5. 31–41.

Erjavec, Tomaž and Sašo Džeroski, 2004: Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. *Applied artificial intelligence* 18. 17–41.

Erjavec, Tomaž, Camelia Ignat, Bruno Pouliquen and Ralf Steinberger, 2005: Massive Multi Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences* 15. 529–540.

Erjavec, Tomaž and Simon Krek, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Šeste konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 46–53.

Erjavec, Tomaž, Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman and Aleš Velušček, 2008: *Specifikacije za leksikon (besednih oblik) – projekt »Sporazumevanje v slovenskem jeziku« – kazalnik 3.* http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik3/SSJ_Kazalnik_3_Specifikacije-leksikon_v1.pdf.

Erjavec, Tomaž, 2009: Odprtost jezikovnih virov za slovenščino. Stabej, Marko (ed.): *Infrastruktura slovenščine in slovenistike. Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 115–121.

Erjavec, Tomaž, Darja Fišer, Simon Krek and Nina Ledinek, 2010: Jezikovni viri projekta JOS. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Sedme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 42–48.

Erjavec, Tomaž and Simon Krek, 2010: *Training corpus jos1M 1.1. Slovenian language resource repository CLARIN.SI.* http://hdl.handle.net/11356/1037.

Erjavec, Tomaž and Nikola Ljubešić, 2011: hrWac in slWaC: Compiling Web Corpora for Croatian and Slovene. Habernal, Ivan and Václav Matoušek (eds.): *Text, Speech and Dialogue. Proceedings of the 14th International Conference, TSD.* Pilsen: Springer Berlin Heidelberg. 395–402.

Erjavec, Tomaž, 2012: MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation* 46/1. 131–142.

Erjavec, Tomaž and Nataša Logar Berginc, 2012: Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 57–62.

Erjavec, Tomaž, 2013: Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0* 1/1. 24–49. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Erjavec, Tomaž, 2014: Odprt dostop do podatkovne baze slovarja. Grahek, Irena and Simona Bergoč (eds.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo.* Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/20-_Tomaz_Erjavec-SlovarPosvet.pdf.

Erjavec, Tomaž and Nikola Ljubešić, 2014: The slWaC 2.0 Corpus of the Slovene Web. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Devete konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 50–55.

Erjavec, Tomaž, 2015: The IMP historical Slovene language resources. *Language resources and evaluation* 49/3. 753–775.

Erjavec, Tomaž, Nikola Ljubešić and Nataša Logar, 2015: The slWaC Corpus of the Slovene Web. *Informatica* 39/1. 35–42.

Erlandsen, Jens, 2004: iLex – new DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004), Brno, 6.–7. September 2004.*

Estellés-Arolas, Enrique and Fernando González-Ladrón-de-Guevara, 2012: Towards an Integrated Crowdsourcing Definition. *Journal of Information Science* 38/2. 189–200.

Felstiner, Alek, 2011: Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley Journal of Employment and Labor Law*. 143–203.

Ferbežar, Ina, Mihaela Knez, Andreja Markovič, Nataša Pirih Svetina, Mojca Schlamberger Brezar, Marko Stabej, Hotimir Tivadar and Jana Zemljarič Miklavčič, 2004: *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete UL and Ministrstvo RS za šolstvo, znanost in šport.

Finkel, Jenny Rose, Trond Grenager and Christopher Manning, 2005: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. 363–370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf.

Finkel, Jenny Rose, Christopher Manning and Andrew Y. Ng, 2006: Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, July 22-23*. 618–626.

Fišer, Darja, 2009: sloWNET – slovenski semantični leksikon. Stabej, Marko (ed.): *Infrastruktura slovenščine in slovenistike. Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 145–149.

Fišer, Darja, Tomaž Erjavec, Ana Zwitter Vitez and Nikola Ljubešić, 2014a: JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Devete konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 56–61.

Fišer, Darja, Aleš Tavčar and Tomaž Erjavec, 2014b: sloWCrowd: A crowdsourcing tool for lexicographic tasks. *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14*. 4371–4375.

Fišer, Darja, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Damjan Popič and Tomaž Erjavec, 2015: Množičenje za slovar sodobnega slovenskega jezika. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 566–586.

Fišer, Darja and Jaka Čibej, 2015: Potencial množičenja v sodobni leksikografiji. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 565–586.

Fort, Karen, Gilles Adda, Benoît Sagot, Joseph Mariani and Alain Couillault, 2014: Crowdsourcing for Language Resource Development: Criticisms About Amazon Mechanical Turk Overpowering Use. *Human Language Technology Challenges for Computer Science and Linguistics*. Springer. 303–314.

Fossati, Marco, Claudio Giuliano and Sara Tonelli, 2013: Outsourcing FrameNet to the Crowd. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofija, Bolgarija. Association for Computational Linguistics. 742–747.

Fox, Gwyneth, 1987: The Case for Examples. Sinclair, John McH. (ed.): *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: Collins. 137–149.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet and Claudia Soria, 2006: Lexical Markup Framework (LMF). *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06*. Pariz: ELRA. 233–236.

Frankenberg-Garcia, Ana, 2012: Learners' Use of Corpus Examples. *International Journal of Lexicography* 25/3. 273–296.

Frankenberg-Garcia, Ana, 2014: The Use of Corpus Examples for Language Comprehension and Production. *ReCALL* 26. 128–146.

Gantar, Polona, 2009: Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo* 54/3-4. 69–94.

Gantar, Polona, 2010: K uporabniku usmerjeni slovnično-leksikalni opisi slovenskega jezika. Gorjanc, Vojko and Andreja Žele (eds.): *Izzivi sodobnega jezikoslovja*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 35–51.

Gantar, Polona, 2011: Leksikalna baza za slovenščino: komu, zakaj in kako (naprej)? *Jezikoslovni zapiski* 17/2. 77–92.

Gantar, Polona and Simon Krek, 2011: Slovene lexical database. Majchráková, Daniela and Radovan Garabík (eds.): *Natural language processing, multilinguality*. Brno: Tribun EU. 72–80.

Gantar, Polona, Simon Krek, Iztok Kosem, Mojca Šorli, Katja Grabnar, Olga Pobirk, Petra Zaranšek and Nina Drstvenšek, 2012: *Leksikalna baza za slovenščino*. Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://www.slovenscina.eu/spletni-slovar/leksikalna-baza, http://www.slovenscina.eu/spletni-slovar/prenos.

Gantar, Polona and Iztok Kosem, 2013: Beleženje in prikazovanje podatkov o jezikovni rabi: od leksikalne baze do spletnega slovarja. Žele, Andreja (ed.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 133–139.

Gantar, Polona, 2015: Homonimija in večpomenskost: od teorije do slovarja, 2015: Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 340–357.

Gantar, Polona, Iztok Kosem, Simon Krek and Vojko Gorjanc, 2015: Collocations dictionary of Slovene: challenge for automatization and crowdsourcing. Corpas Pastor, Gloria, Miriam Buendía Castro and Rut Gutierrez Florido (eds.): *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*. Europhras 2015, Malaga, 29 June to 1 July 2015.

Gao, Qin and Stephan Vogel, 2010: Consensus versus expertise: a case study of word alignment with Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10*. Stroudsburg, ZDA: Association for Computational Linguistics. 30–34.

Gliha Komac, Nataša, Nataša Jakop, Janoš Ježovnik, Simona Klemenčič, Domen Krvina, Nina Ledinek, Tanja Mirtič, Andrej Perdih, Špela Petric, Marko Snoj and Andreja Žele, 2015: *Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika.* Različica 1.1. Ljubljana: Inštitut za slovenski jezik Frana Ramovša; Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, 2015. http://www.fran.si/179/novi-slovar-slovenskega-knjiznega-jezika/datoteke/Koncept_NoviSSKJ.pdf.

Gorjanc, Vojko, 2004: Politična korektnost in slovarski opisi slovenščine – zgolj modna muha? Stabej, Marko (ed.): *Moderno v slovenskem jeziku, literaturi in kulturi. 40. seminar slovenskega jezika, literature in kulture.* Ljubljana: Filozofska fakulteta. 153–161.

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje.* Domžale: Izolit.

Gorjanc, Vojko, 2014: O heteronormativnosti slovarskega opisa slovenskega jezika: homoseksualnost, ekshibicionizem in druge perverznosti. *Narobe* 7/27-28. 12–15.

Górski, Rafał L. and Marek Łazinski, 2012: Typologia tekstów w NKJP. Przepiórkowski, Adam et al. (eds.): *Narodowy korpus języka polskiego.* Varšava: Wydawnictwo Naukowe PWN. 13–23.

Grčar, Miha, Simon Krek and Kaja Dobrovoljc, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 89–94.

Gut, Ulrike and Petra Saskia Bayerl, 2004: Measuring the Reliability of Manual Annotations of Speech Corpora. *The Proceedings of Speech Prosody 2004.* Nara, Japan. 565–568.

Haase, Peter, Jeen Broekstra, Andreas Eberhart and Raphael Volz, 2004: A Comparison of RDF Query Languages. McIlraith, Sheila A., Dimitris Plexousakis and Frank van Harmelen (eds.): *The Semantic Web – ISWC 2004.* Berlin and Heidelberg: Springer. 502–517.

Hajnšek-Holz, Milena, 1993: Leksikografski problemi prenosa knjižne oblike Slovarja slovenskega knjižnega jezika v računalniško. Štrukelj, Inka (ed.): *Jezik tako in drugače.* Ljubljana: Društvo za uporabno jezikoslovje Slovenije. 420–432.

Hanks, Patrick, 2004: Corpus Pattern Analysis. Williams, Geoffrey and Sandra Vessier (eds.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004.* Lorient: Universite de Bretagne-sud. 87–97.

Hanks, Patrick and James Pustejovsky, 2005: A Pattern Dictionary for Natural Language Processing. *Revue Française de Linguistique Appliquée* 10/2. 63–82.

Hartmann, Reinhard R. K., 1999: The Exeter University Survey of Dictionary Use. Hartmann, Reinhard R. K. (ed.): *Dictionaries in Language Learning: Recommendations, National Reports and Thematic Reports from the TNP Sub-Project 9: Dictionaries.* Berlin: Freie Universität. 36–52.

Harvey, Keith and Deborah Yuill, 1997: A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics* 18/3. 253–278.

Hatherall, Glyn, 1984: Studying dictionary use: some findings and proposals. Hartmann, Reinhard R. K. (ed.): *LEX'eter ,83 Proceedings: Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983.* Tübingen: Niemeyer Verlag. 183–189.

Heinonen, Tarja, 2014: Workflow in Kielitoimiston sanakirja. *Workflow of Corpus-based Lexicography*, COST ENeL WG3 meeting, Bolzano, 19 julij. http://www.elexicography.eu/wp-content/uploads/2014/07/Heinonen_2014_COST_Bolzano.pdf.

Herring, Sussan C., 2001: Computer-Mediated Discourse. Schiffrin, Deborah, Deborah Tannen and Heidi E. Hamilton (eds.): *The Handbook of Discourse Analysis.* Oxford: Blackwell Publishers. 612–634.

Herring, Susan C., Lois Ann Scheidt, Sabrina Bonus and Elijah Wright, 2004: Bridging the Gap: a Genre Analysis of Weblogs. *Proceedings of the 37th Hawaii International Conference on System Sciences.* IEEE – Institute of Electrical and Electronics Engineers. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.2930&rep=rep1&type=pdf.

Hinrichs, Erhard, Marie Hinrichs and Thomas Zastrow, 2010: WebLicht: Web-based LRT services for German. *Proceedings of the ACL 2010 System Demonstrations.* Association for Computational Linguistics. 25–29. http://www.aclweb.org/anthology/P10-4005.

Holozan, Peter, 2011: *Samodejno izdelovanje besedilnih logičnih nalog v slovenščini.* Magistrsko delo. Ljubljana: Fakulteta za računalništvo in informatiko UL.

Holozan, Peter, 2012: Kako dobro programi popravljajo vejice v slovenščini. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 101–106.

Holozan, Peter, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman and Aleš Velušček, 2008: *Specifikacije za učni korpus – projekt »Sporazumevanje v slovenskem jeziku« – Kazalnik 2.* http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik2/SSJ_Kazalnik_2_Specifikacije-ucni-korpus_v1.pdf.

Honselaar, Wim, 2003: Examples of design and production criteria for bilingual dictionaries. Sterkenburg, Piet van (ed.)*: A practical guide to lexicography.* Amsterdam and Philadelphia: John Benjamins. 323–332.

Horvat, Aleš and Jernej Vičič, 2012: Strojno prevajanje med slovenščino in španščino. Baldomir Zajc and Andrej Trost (eds.): *Zbornik enaindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2012.* Portorož, Slovenija. 101–104.

Howe, Jeff, 2008: *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business.* New York: Crown Publishing Group.

Humblé, Philippe, 2001: *Dictionaries and Language Learners.* Frankfurt am Main: Haag & Herchen.

Hunston, Susan and Sara Laviosa, 2001: *Corpus linguistics.* Birmingham: Centre for English Language Studies, The University of Birmingham.

Ide, Nancy and Jean Véronis, 1994: MULTEXT: Multilingual Text Tools and Corpora. *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94.* Kyoto, Japan. 588–592.

ISO 24611, 2012: *Language resource management – Morpho-syntactic annotation framework (MAF).*

Jackson, Howard, 1988: *Words and Their Meaning.* London: Longman.

Jackson, Howard, 2002: *Lexicography: an introduction.* Routeledge.

Jakopin, Primož and Aleksandra Bizjak Končar, 1997: O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija* 45/3-4. 513-532.

Javoršek, Jan Jona, 2015: *Razvoj korpusnega skladenjskega razčlenjevalnika.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Jesenovec, Mojca, 2004: Poučevanje, učenje in pomnjenje leksike drugega/tujega jezika. *Jezik in slovstvo* 49/3-4. 35–47.

Jewler, A. Jarome and Bonnie L. Drewniany, 2005: *Creative strategy in Advertising*. Belmont: Thomson and Wadsworth.

Josselin-Leray, Amelie, 2005: *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues*. Doktorska disertacija. Université Lumière Lyon II.

Josselin-Leray, Amelie and Roberts, Roda P., 2007: La définition des termes dans les dictionnaires généraux unilingues: analyse de quelques exemples du domaine de la volcanologie à la lumière d'un corpus de vulgarisation. L'Homme, Marie-Claude and Sylvie Vandaele (eds.): *Lexicographie et terminologie: compatibilité des modèles et des méthodes*. Ottawa: Presses de l'Université d'Ottawa. 141–188.

Joubert, Alain and Mathieu Lafourcade, 2012: A new dynamic approach for lexical networks evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC'12*. Istanbul, Turky. 3687–3691.

Jurgens, David and Roberto Navigli, 2014: It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics* 2. Association for Computational Linguistics. 449–463.

Juršič, Matjaž, Igor Mozetič, Tomaž Erjavec and Nada Lavrač, 2010: LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of universal computer science* 16/9. 1190–1214.

Kallas, Jelena, Maria Tuulik and Margit Langemets, 2014: The Basic Estonian Dictionary: the First Monolingual L2 Learner's Dictionary of Estonian. *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano. http://www.euralex. org/elx_proceedings/Euralex2014/euralex_2014_086_p_1109.pdf.

Kallas, Jelena, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik and Ülle Viks, 2015: Automatic generation of the Estonian Collocations Dictionary database. Kosem, Iztok, Miloš Jakubiček, Jelena Kallas and Simon Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of eLex 2015, 11–13 August 2015, Herstmonceux Castle, UK*. Ljubljana and Brighton: Trojina, Institute for Applied Slovene Studies and Lexical Computing Ltd. 1–20.

Kennedy, Graeme, 1999: *An Introduction to Corpus Linguistics*. London and New York: Longman.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell, 2004: The Sketch Engine. Williams, Geoffrey and Sandra Vessier (eds.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*. Lorient: Universite de Bretagne-sud. 105–116.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychly, 2008: GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Bernal, Elisenda and Janet DeCesaris (eds.): *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 425–432.

Klein, Wolfgang and Alexander Geyken, 2010: Das Digitale Wörterbuch der Deutschen Sprache (DWDS). Heid, Ulrich, Stefan Schierholz, Wolfgang Schweickard and Herbert Ernst Wiegand (eds.): *Lexikographica*. Berlin and New York: de Gruyter. 79–93.

Klemenc, Bojan, Marko Robnik-Šikonja, Luka Fürst, Ciril Bohak and Simon Krek, 2015: Tehnološka izvedba sodobnega digitalnega slovarja. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 52–63.

Klosa, Anette, 2013: The lexicographical process (with special focus on online dictionaries). Gouws, Rufus H., Ulrich Heid, Wolfgang Schweickard and Herbert Ernst Wiegand (eds.): *Dictionaries. An international Encyclopedia of Lexicography.* Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin and Boston: de Gruyter. 517–524.

Klubička, Filip and Nikola Ljubešić, 2014: Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian. *Jezikovne tehnologije. Zbornik 17. mednarodne multikonference Informacijska družba – IS 2014.* Ljubljana: Inštitut Jožef Stefan.

Kohlschütter, Christian, Peter Fankhauser and Wolfgang Nejdl, 2010: Boilerplate Detection Using Shallow Text Features. *WSDM 2010 The Third ACM International Conference on Web Search and Data Mining.* New York: ACM. 441–450.

Kola, Kjersti Wictorsen, 2012: A study of pupils' understanding of the morphological information in the Norwegian electronic dictionary Bokmålsordboka in Nynorskordboka. Vatvedt Fjeld, Ruth and Julie Matilde Torjusen (eds.): *Proceedings of the 15th EURALEX international Congress. EURALEX 2012*. Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier. 672–675.

Kosem, Iztok, 2006: Definicijski jezik v Slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel. *Jezik in slovstvo* 51/5. 25–45.

Kosem, Iztok, 2010: *Designing a model for a corpus-driven dictionary of academic English.* PhD dissertation. Aston University, UK.

Kosem, Iztok, Milos Husák and Diana McCarthy, 2011: GDEX for Slovene. Iztok Kosem and Karmen Kosem (eds.): *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, 10–12 November 2011, Bled, Slovenia.* Ljubljana: Trojina, Institute for Applied Slovene Studies. 151–159.

Kosem, Iztok, 2012: Using GDEX in (semi)-automatic creation of database entries. SKEW-3, 3rd International Sketch Engine workshop, 21–22. marec, Brno, Češka. https://www.sketchengine.co.uk/documentation/attachment/wiki/SKEW-3/Program/GDEX-automatic-entry-extraction-Iztok_Kosem.pdf?format=raw.

Kosem, Iztok, Mojca Stritar, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt and Tadeja Rozman, 2012a: *Analiza jezikovnih težav učencev: korpusni pristop.* Ljubljana: Trojina, zavod za uporabno slovenistiko.

Kosem, Iztok, Polona Gantar and Simon Krek, 2012b: Avtomatsko luščenje leksikalnih podatkov iz korpusa. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012.* Institut Jožef Stefan. 117–122.

Kosem, Iztok, Polona Gantar and Simon Krek, 2013a: Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0* 1/2. 139–164. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_07.pdf.

Kosem, Iztok, Polona Gantar and Simon Krek, 2013b: Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. Kosem, Iztok, Je-

lena Kallas, Polona Gantar, Simon Krek, Margit Langemets and Maria 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: Trojina, Institute for Applied Slovene Studies in Tallinn: Eesti Keele Instituut. 32–48.

Kosem, Iztok, 2015: Oznake: slovarska baza in slovar. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 482–494.

Krek, Simon and Tomaž Erjavec, 2009: Standardised encoding of morphological lexica for Slavic languages. Anatoliiovych Shyrokov, Volodymyr and Ludmila Dimitrova (eds.): *MONDILEX Second Open Workshop, Organization and development of digital lexical resources: proceedings*. Kijev: National Academy of Sciences of Ukraine. 24–29.

Krek, Simon, 2011: Language data for digital natives: old wine in a new bottle or...? Plenarno predavanje na konferenci: Electronic lexicography in the 21st century: new applications for new users (eLex2011), Bled, 10–12. november 2011. http://videolectures.net/elex2011_krek_language/?q=simon%20krek.

Krek, Simon, 2012a: New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3, tretja mednarodna delavnica orodja Sketch Engine*. Brno, Češka, 21.–22. marec 2012.

Krek, Simon, 2012b: *Slovenski jezik v digitalni dobi/The Slovene Language in the Digital Age*. http://www.meta-net.eu/whitepapers/e-book/slovene.pdf.

Krek, Simon, 2012c: Spletni portal Slogovni priročnik. Krakar Vogel, Boža (ed.): *Slavistika v regijah – Koper. Zbornik Slavističnega društva Slovenije* 23. Ljubljana: Zveza društev Slavistično društvo Slovenije and Znanstvena založba Filozofske Fakultete UL. 225–231.

Krek, Simon, 2013: *Sporazumevanje v slovenskem jeziku: vsebina in rezultati – 2008–2013*. http://videolectures.net/zakljucnakonferencassj2013_krek_vsebina/.

Krek, Simon and Iztok Kosem, 2013: *Odgovor na prispevek „SSKJ danes in jutri, potem pa … "*. http://www.sssj.si/datoteke/SSKJ_danes_in_jutri_odgovor.pdf.

Krek, Simon, 2014a. Prva in druga izdaja SSKJ. *Slovenščina 2.0* 2/2. 114–158.

Krek, Simon, 2014b: SSKJ v slovarski bazi. Grahek, Irena and Simona Bergoč (eds.): *Novi slovar za 21. stoletje*. Ljubljana, Ministrstvo za kulturo. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/5-_Simon_Krek_SSKJ_v_slovarski_bazi.pdf.

Krek, Simon, Helena Dobrovoljc, Kaja Dobrovoljc and Damjan Popič, 2013a: Online style guide for Slovene as a language resources hub. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference*. Ljubljana: Trojina, Institute for Applied Slovene Studies in Tallinn: Eesti Keele Instituut. 379–391. http://eki.ee/elex2013/proceedings/eLex2013_26_Krek+etal.pdf.

Krek, Simon, Iztok Kosem and Polona Gantar, 2013b: *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*. Verzija 1.1. http://sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf.

Krek, Simon, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek, Nanika Holz, 2013c: Training corpus ssj500k 1.3. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1029.

Krvina, Domen, 2014: Sprotni slovar slovenskega jezika. Gradivo: okrogla miza Slovensko slovaropisje, Pišece, 2. 10. 2014. *Slavia Centralis* 2. 90–92.

Landau, Sidney I., 1974: Of Matters Lexicographical. Scientific and Technical Entries in American Dictionaries. *American Speech* 49/3-4. 241–244.

Landau, Sidney, I., 1984 (2001): *Dictionaries. The Art and Craft of Lexicography*. New York: Scribners (Cambridge: Cambridge University Press).

Laufer, Batia, 1993: The Effects of Dictionary Definitions and Examples on the Comprehension of New L2 words. *Cahiers de Lexicologie* 63. 131–142.

Laufer, Batia, 2000: Electronic dictionaries and incidental vocabulary acquisition: does technology make a difference? Heid, Ulrich, Stefan Evert, Egbert Lehmann and Christian Rohrer (eds.): *Proceedings of the Ninth EURALEX International Congress, Stuttgart, Germany, August 8th-12th 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung. 849–854.

Laufer, Batia and Tami Levitzky-Aviad, 2006: Examining the effectiveness of 'Bilingual Dictionary Plus' – a dictionary for production in a foreign language. *International Journal of Lexicography* 19/2. 135–155.

Lease, Matthew and Omar Alonso, 2014: *Crowdsourcing and Human Computation. Introduction*. Springer.

Ledinek, Nina, 2014a: Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine. Ljubljana: Založba ZRC, ZRC SAZU.

Ledinek, Nina, 2014b: Terminologija v enojezičnem razlagalnem slovarju srednjega obsega. Grahek, Irena and Simona Bergoč (eds.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/10-_Nina_Ledinek_-_koncni_prispevek.pdf.

Leech, Geoffrey, 1992: Corpora and Theories of Linguistic Performance. Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Berlin and New York: de Gruyter. 105–122.

Leffa, Vilson, 1993: Using an Electronic Dictionary to Understand Foreign Language Texts. *Trabalhos Em Linguistica Aplicada* 21. 19–29.

Lew, Robert, 2013: User-generated content (UGC) in online English dictionaries. *OPAL - Online publizierte Arbeiten zur Linguistik*.

Ljubešić, Nikola and Tomaž Erjavec, 2011: hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. *Text, Speech and Dialogue: Lecture Notes in Computer Science* 6836. 395–402.

Ljubešić, Nikola, Marija Stupar and Terezija Jurić, 2013: Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0* 1/2. 35–57. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_03.pdf.

Ljubešić, Nikola, Tomaž Erjavec and Darja Fišer, 2014: Standardizing tweets with character-level machine translation. Gelbukh, Alexander (ed.): *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014*, *Kathmandu, Nepal*. Heidelberg: Springer. 164–175.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak and Iza Škrjanec, 2015: Predicting the level of standardness of text in user-generated content. *Proceedings of the Conference RANLP "Recent Advances in Natural Language Processing"*. Hissar, Bolgarija.

Logar Berginc, Nataša and Špela Vintar, 2008: Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izraza. *Jezik in slovstvo* 53/5. 3–17.

Logar Berginc, Nataša, 2009: Slovenski splošni in terminološki slovarji: za koga? Stabej, Marko (ed.): *Infrastruktura slovenščine in slovenistike. Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 225–231.

Logar Berginc, Nataša and Simon Šuster, 2009: Gradnja novega korpusa slovenščine. *Jezik in slovstvo* 54/3-4. 57–68.

Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt and Simon Krek, 2012: *Korpusi slovenskega jezika Gigafida, Kres, ccGigafida in ccK-RES: gradnja, vsebina, uporaba*, Ljubljana: Trojina, zavod za uporabno slovenistiko and Fakulteta za družbene vede.

Logar Berginc, Nataša and Nikola Ljubešić, 2013: Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0* 1/1. 78–110. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_05.pdf.

Logar Berginc, Nataša, Špela Vintar and Špela Arhar Holdt, 2013: Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka. *Slovenščina 2.0* 1/2. 113–138. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_06.pdf.

Logar, Nataša, Špela Vintar and Špela Arhar Holdt, 2012: Luščenje terminoloških kandidatov za slovar odnosov z javnostmi. Erjavec, Tomaž, Žganec Gros, Jerneja (eds.). *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 135–140.

Logar, Nataša, 2013: *Korpusna terminografija: primer odnosov z javnostmi*. Ljubljana: Trojina, zavod za uporabno slovenistiko and Fakulteta za družbene vede.

Logar, Nataša, 2014: Verodostojnost korpusa kot gradivnega vira za slovar. Grahek, Irena and Simona Bergoč (eds.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/7-_Natasa_Logar_-prispevek_-za_oddajo.pdf http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/20-_Tomaz_Erjavec-SlovarPosvet.pdf.

Logar, Nataša, 2015: Gradnja referenčnih korpusov na novo: nadgradnja Gigafide. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 218–241.

Logar, Nataša, Nikola Ljubešić and Tomaž Erjavec, 2015: KRES in Gigafida kot korpusna osnova za slovar: podobnosti in razlike. Smolej, Mojca (ed.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. V tisku.

Lorentzen, Henrik and Liisa Theilgaard, 2012: Online dictionaries – how do users find them and what do they do once they have? Vatvedt Fjeld, Ruth and Julie Matilde Torjusen (eds.)*: Proceedings of the 15th EURALEX International Congress. EURALEX 2012*. Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier. 654–660.

Lui, Marco and Timothy Baldwin, 2012: langid. py: an Off-the-Shelf Language Identification Tool. *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistic. 25–30. http://www.aclweb.org/anthology/P12-3005.

McCreary, Don R. and Fredric T. Dolezal, 1999: A Study of Dictionary Use by ESL Students in an American University. *International Journal of Lexicography* 12/2. 107–146.

McCreary, Don R., 2002: American Freshmen and English Dictionaries: ‚I Had Aspersions of Becoming an English Teacher‘. *International Journal of Lexicography* 15/3. 181–205.

McDonald, Ryan, Kevin Lerman and Fernando Pereira, 2006: Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. *Tenth Conference on Computational Natural Language Learning (CoNLL-X), NYC, USA.* Stroudsburg, ZDA: Association for Computational Linguistics. 216–220.

Meschonnic, Henri, 1991: *Des mots et des mondes: dictionnaires, encyclopâedies, grammaires, nomenclatures.* Paris: Hatier.

Meyer, Christian M. and Iryna Gurevych, 2012: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. *Electronic Lexicography.* 259–291.

Migla, Ilga and Ieva Zuicena, 2014: The Dictionary of Contemporary Latvian Language and its Lexicographical Process. *Workflow of Corpus-based Lexicography, COST ENeL WG3 meeting, Bolzano, 19. julij.* http://www.elexicography.eu/wp-content/uploads/2014/07/Migla_2014_COST_Bolzano.pdf.

Mikolič, Vesna, 2007: Modifikacija podstave in argumentacijska struktura besedilnih vrst. *Slavistična revija* 55/1-2. 341–355.

Mikolič, Vesna, 2013: Področni govor in terminologija na primeru jezika turizma. Žele, Andreja (ed.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 255–261.

Mikolič, Vesna and Maša Rolih, 2015: Besedilna zvrstnost v novih medijih kot slovarska vsebina. Smolej, Mojca (ed.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. V tisku.

Miller, George and Patricia Gildea, 1987: How Children Learn Words. *Scientific American* 257/3. 94–99.

Mitchell, Evelyn, 1983: *Search-Do Reading: Difficulties in Using a Dictionary.* Aberdeen: College of Education.

Müller, Jakob, 1996: Slovar slovenskega knjižnega jezika in kritika z bibliografijo (1960–1992). *Razprave SAZU. Razred za filološke in literarne vede* 15. Ljubljana: SAZU. 187–234.

Müller, Jakob, 2009: Kritične misli in zamisli o SSKJ. Perdih, Andrej (ed.): *Strokovni posvet o slovarju slovenskega jezika.* Ljubljana: Založba ZRC, ZRC SAZU. 17–21, 25.

Müller-Spitzer, Carolin, Alexander Koplenig and Antje Töpel, 2011: What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. Kosem, Karmen and Iztok Kosem (ur): *Proceedings of eLex 2011, Bled, 10–12 November 2011: Electronic Lexicography in the 21st Century - New Applications for New Users.* Ljubljana: Trojina, Institute for Applied Slovene Studies. 203–208.

Müller-Spitzer, Carolin, Sascha Wolfer and Alexander Koplenig, 2015: Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography* 28/1. 1–26.

Negri, Matteo, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo and Alessandro Marchetti, 2011: Divide and conquer: crowdsourcing the creation of crosslingual textual entailment corpora. *Conference on Empirical Methods in Natural Language Processing, EMNLP '11. Proceedings of the Conference.* Stroudsburg, ZDA: Association for Computational Linguistics. 670–679.

Nesi, Hilary, 1996: The Role of Illustrative Examples in Productive Dictionary Use. *Dictionaries* 17. 198–206.

Nesi, Hilary, 2000: *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries.* Tübingen: Max Niemeyer Verlag.

Nesi, Hilary and Richard Haill, 2002: A Study of Dictionary Use by International Students at a British University. *International Journal of Lexicography* 15/4. 277–305.

Nesi, Hilary, 2011: The effect of e-dictionary font on vocabulary retention. *Electronic lexicography in the 21st century: new applications for new users (eLex), 10–12 November 2011, Bled, Slovenia.* http://videolectures.net/elex2011_nesi_effect/.

Neubach, Abigail and Andrew Cohen, 1988: Processing Strategies and Problems Encountered in the Use of Dictionaries. *Dictionaries: Journal of the Dictionary Society of North America* 10. 1–19.

Newman, Andrew, 2007: *A Relational View of the Semantic Web.* http://www.xml.com/pub/a/2007/03/14/a-relational-view-of-the-semantic-web.html.

Nidorfer Šiškovič, Mojca, 2013: Žanrskost funkcijskih besedilnih vrst. Žele, Andreja (ed.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 269–275.

Nivre, Joakim et al., 2015: *Universal Dependencies 1.0.* http://hdl.handle.net/11234/1-1464.

Oblak, Tanja, Gregor Petrič, Marko Pahor, Franc Trček and Slavko Splichal, 2005: *Splet kot medij in mediji na spletu.* Ljubljana: Fakulteta za družbene vede.

Ogrin, Matija, Jan Jona Javoršek and Tomaž Erjavec, 2013: A register of early modern Slovenian manuscripts. *Journal of the Text Encoding Initiative* 4. 1–13. http://jtei.revues.org/715.

Oyama, Satoshi, Yukino Baba, Yuko Sakurai and Hisashi Kashima, 2013: Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores. *Proceedings* of the *Twenty-Third International Joint Conference on Artificial Intelligence.* 2554–2560.

Paynter, Diane E., Elena Bodrova and Jane K. Doty, 2005: *For the Love of Words: Vocabulary Instruction that Works.* San Francisco: Jossey-Bass teacher.

Pečjak, Sonja, 2012: *Psihološki vidiki bralne pismenosti. Od teorije k praksi.* Ljubljana: Znanstvena založba Filozofske fakultete.

Petrylaitė, Regina, Diana Vaškelienė and Tatjana Vėžytė, 2008: Changing Skills of Dictionary Use. *Studies about Languages* 12. 77–82.

Pollak, Senja, 2014: *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Pomikálek, Jan, 2011: *jusText.* LINDAT/CLARIN Digital Library at Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11858/00-097C-0000-000D-F696-9.

Predmetnik osnovne šole, 2014. Ljubljana Ministrstvo za izobraževanje, znanost in šport RS. http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/devetletka/predmetniki/Pred_14_OS_4_12.pdf.

Prunč, Erich, 2009: Veliki čudež malega jezika. *Jezik in slovstvo* 54/1. 5–12.

Rayson, Paul and Roger Garside, 2000: Comparing Corpora Using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora.* Hong Kong. 1–6. http://www.comp.lancs.ac.uk/~rayson/publications/rg_acl2000.pdf.

Reffle, Ulrich, 2011: Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering* 17. 265–282.

Rigler, Jakob, 1971: H kritikam pravopisa, pravorečja in oblikoslovja v SSKJ. *Slavistična revija* 19/4. 433–462.

Rigler, Jakob, 1972: H kritikam pravopisa, pravorečja in oblikoslovja v SSKJ. *Slavistična revija* 20/1. 244–251.

RIS: Raba interneta v Sloveniji. http://www.ris.org/.

Rojc, Matej, Zdravko Kačič and Darinka Verdonik, 2002: Design and implementation of the Slovenian phonetic and morphology lexicons for the use in spoken language applications. *Proceeding od the Third international conference on language resources and evaluation, Las Palmas de Grand Canaria*. Grand Canaria: European Language Resources Association. 1296–1300.

Romih, Miro and Peter Holozan, 2002: Slovensko-angleški prevajalni sistem. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Tretje konference Jezikovne tehnologije*. Institut Jožef Stefan. 167.

Rozman, Tadeja, 2009: The Dictionary of Standard Slovenian – A(n) (Un)faithful Companion? Granić, Jagoda (ed.): *Jezična politika i jezična stvarnost/Language Policy and Language Reality*. Zagreb: HDPL. 126–136.

Rozman, Tadeja*,* 2010: *Vloga enojezičnega razlagalnega slovarja slovenščine pri razvoju jezikovne zmožnosti.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Rozman, Tadeja, Irena Krapš Vodopivec, Mojca Stritar, Iztok Kosem and Simon Krek, 2010: *Nova didaktika poučevanja slovenskega jezika – projekt »Sporazumevanje v slovenskem jeziku« – Kazalnik 15*. http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K15.aspx.

Rozman, Tadeja, Irena Krapš Vodopivec, Mojca Stritar and Iztok Kosem, 2012: *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Rumshisky, Anna, 2011: Crowdsourcing Word Sense Definition. *LAW* V. *Fifth* Linguitic Annotation Workshop. *Proceedings of the Workshop. Portland, ZDA: Association for Computational Linguistics.* 74–81.

Rumshisky, Anna, Nick Botchan, Sophie Kushkuley and James Pustejovsky, 2012: Word Sense Inventories by Non-Experts. *Proceedings* of the Eighth International Conference on Language Resources and Evaluation. LREC '12. Istanbul, Turky.

Rundell, Michael and Adam Kilgarriff, 2011: Automating the creation of dictionaries: where will it all end? Meunier, Fanny, Sylvie De Cock, Gaëtanelle Gilquin and Magali Paquot (eds.): *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam and Philadelphia: John Benjamins. 257–281.

Rundell, Michael, 2014: Macmillan English Dictionary: The End of Print? *Slovenščina 2.0* 2/2. 1–14. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_02.pdf.

Rupnik, Jan, Miha Grčar and Tomaž Erjavec, 2010: Improving Morphosyntactic Tagging of Slovene Language Through Meta-tagging. *Informatica* 34/2. 169–175.

Rychly, Pavel, 2007: Manatee/bonito - a modular corpus manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Univerza Masaryk. 65–70.

Sabou, Marta, Kalina Bontcheva, Leon Derczynski and Arno Scharl, 2014: Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC '14.* Reykjavik, Iceland. 859–866.

Salton, Gerard and Christopher Buckley, 1988: Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management: an International Journal* 24/5. 513–523.

Scherrer, Yves and Tomaž Erjavec, 2013: Modernizing historical Slovene words with character-based SMT. *The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. 58–62. http://hal.archives-ouvertes.fr/docs/00/83/85/75/PDF/13-scherrer-modernize.pdf.

Schiffrin, Deborah, 1987: *Discourse Markers.* Cambridge: Cambridge University Press.

Schutz, Rik, 2002: Indirect Offensive Language in Dictionaries. Braasch, Anna and Claus Povlsen (eds.): *Proceedings of the 10th EURALEX International Congress.* Copenhagen: Center for Sprogteknologi. 637–641.

Scott, Mike, 1997: PC analysis of key words – and key key words. *System* 25/2. 233–245.

Selva, Thierry and Serge Verlinde, 2002: L'utilisation d'un dictionnaire électronique: une étude de cas. Anna Braasch and Claus Povlsen (eds.): *Proceedings of the 10th EURALEX International Congress.* Copenhagen: Center for Sprogteknologi. 773–781.

Sharoff, Serge, 2010: Analysing Similarities and Differences between Corpora. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Sedme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 5–11.

Silberman, M. Six, Joel Ross, Lilly Irani and Bill Tomlinson, 2010: Sellers' problems in human computation markets. *Proceedings of the ACM SIGKDD Workshop on Human Computation.* 18–21.

Sinclair, John McH., 1991: *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Singleton, David, 1999: *Exploring the Second Language Mental Lexicon.* Cambridge: Cambridge University Press.

Skubic, Andrej, 1995: Klasifikacija funkcijske zvrstnosti in pragmatična definicija funkcije. *Jezik in slovstvo* 40/5. 155–168.

Skubic, Andrej E., 2004: Sociolekti od izraza do pomena: kultiviranost, obrobje in eksces. Kržišnik, Erika (ed.): *Aktualizacija jezikovnozvrstne teorije na Slovenskem: členitev jezikovne resničnosti. Obdobja* 22. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete UL. 297–320.

Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje. Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva, 2011. http://www.europass.si/files/userfiles/europass/SEJO%20komplet%20za%20splet.pdf.

Slovar slovenskega knjižnega jezika, 1970–1991. Ljubljana: DZS and Slovenska akademija znanosti in umetnosti, Inštitut za slovenski jezik.

Slovenski pravopis, 2001. Jože Toporišič (editor in chief). Ljubljana: Založba ZRC, ZRC SAZU.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng, 2008: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Conference on Empirical Methods in Natural Language Processing, EMNLP '11. Proceedings of the Conference.* Stroudsburg, ZDA: Association for Computational Linguistics. 254–263.

Sproat, Richard, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf and Christofer Richards, 2001: Normalization of Non-Standard Words. *Computer Speech and Language* 15/3. 287–333.

Stabej, Marko and Primož Vitez, 2000: KGB (korpus govorjenih besedil) v slovenščini. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 79–81.

Stabej, Marko, 2009: Slovarji in govorci: kot pes in mačka? *Jezik in slovstvo,* 54/3-4. 115–138.

Stabej, Marko, Tadeja Rozman, Nataša Pirih Svetina, Nina Modrijan and Boštjan Bajec, 2008: *Jezikovni viri pri jezikovnem pouku v osnovni in srednji šoli: končno poročilo z rezultati dela*. Ljubljana: Pedagoški inštitut. http://www.trojina.si/p/jezikovni-viri-pri-jezikovnem-pouku-v-osnovni-in-srednji-soli/.

Stritar, Mojca, 2012: *Korpusi usvajanja tujega jezika.* Ljubljana: Zveza društev Slavistično društvo Slovenije.

Suchomel, Vít and Jan Pomikálek, 2012: Efficient Web Crawling for Large Text Corpora. Kilgarriff, Adam and Serge Sharoff (eds.): *Proceedings of the 7th Web as Corpus Workshop (WAC7)*. Lyon. 39–43. http://nlp.fi.muni.cz/~xsuchom2/papers/PomikalekSuchomel_SpiderlingEfficiency.pdf.

Summers, Della, 1988, The Role of Dictionaries in Language Learning. Carter, Ron and Michael McCarthy (eds.): *Vocabulary and Language Teaching*. Longman. 111–125.

Ševčíková, Magda, Zdeněk Žabokrtský and Oldřich Krůza, 2007: Named Entities in Czech: Annotating Data and Developing NE Tagger. *Text, Speech and Dialogue. Lecture Notes in Computer Science* 4629. 188–195.

Šimková, Mária and Radovan Garabík, 2014: Slovenský národný korpus (2002–2012): východiská, ciele a výsledky pre výskum a prax. Gajdošová, Katarína and Adriána Žáková (eds.): *Jazykovedné štúdie XXXI: Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu)*. Bratislava: VEDA. 35–64.

Škiljan, Dubravko, 1999: *Javni jezik: k lingvistiki javne komunikacije*. Ljubljana: Studia Humanitatis.

Šnajder, Jan, 2013: Models for predicting the inflectional paradigm of Croatian words. *Slovenščina 2.0* 1/2. 1–34. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_02.pdf.

Štajner, Tadej, Tomaž Erjavec and Simon Krek, 2013: Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0* 1/2. 58–81. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_04.pdf.

Štebe, Janez, Sonja Bezjak and Sonja Lužar, 2013: *Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji*. Ljubljana: Založba Fakultete za družbene vede.

Tavčar, Aleš, Darja Fišer and Tomaž Erjavec, 2012: sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Inštitut Jožef Stefan. 197–202.

Teubert, Wolfgang, 2005: Korpusno jezikoslovje in leksikografija. Gorjanc, Vojko and Simon Krek (eds.): Študije o korpusnem jezikoslovju. Ljubljana: Krtina. 103–136.

Tiberius, Carole and Simon Krek, 2014: *Workflow of Corpus-Based Lexicography*. Deliverable COST-ENeL-WG3 meeting July 2014. Bolzano/Bozen.

Tiberius, Carole and Tanneke Schoonheim, 2015: The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process. Hildenbrandt, Vera (ed.): *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie"*. Mannheim: Institut für Deutsche Sprache.

Tono, Yukio, 1984: *On the dictionary user's reference skills*. B. Ed. Dissertation. University of Tokyo.

Tono, Yukio, 2000: On the effects of different types of electronic dictionary interfaces on L2 learners' reference behaviour in productive/receptive tasks. Heid, Ulrich, Stefan Evert, Egbert Lehmann and Christian Rohrer (eds.): *Proceedings of the 9th EURALEX International Congress, EURALEX 2000*. Stuttgart: Institut fur Maschinelle Sprach-verarbeitung, Universitat Stuttgart. 855–862.

Tono, Yukio, 2011: Application of Eye-Tracking in EFL Learners. Dictionary Look-up Process Research. *International Journal of Lexicography* 24/1. 124–153.

Toporišič, Jože, 1971a: Pravopis, pravorečje in oblikoslovje v Slovarju slovenskega knjižnega jezika I. *Slavistična revija* 19/1. 55–75.

Toporišič, Jože, 1971b: Pravopis, pravorečje in oblikoslovje v Slovarju slovenskega knjižnega jezika I. *Slavistična revija* 19/2. 222–229.

Toporišič, Jože, 1976 (1984, 2000, 2004): *Slovenska slovnica.* Maribor: Obzorja.

Trap-Jensen, Lars, Henrik Lorentzen and Nicolai Sørensen, 2014: An odd couple – Corpus frequency and look-up frequency: what relationship? *Slovenščina 2.0* 2/2. 94–113. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_07.pdf.

Učni načrt. Program Osnovna šola. Slovenščina, 2011: http://www.mizs.gov.si/fileadmin/ mizs.gov.si/pageuploads/podrocje/os/prenovljeni_UN/UN_slovenscina_OS.pdf.

Urdang, Laurence, 1984: A lexicographer's adventures in computing. *Dictionaries: Journal of the Dictionary Society of North America* 6.1. 150–165.

Venhuizen, Noortje J., Valerio Basile, Kilian Evang and Johan Bos, 2013: Gamification for word sense labeling. Erk, Katrin and Alexander Koller (eds.): *Proceedings of the 10th International Conference on Computational Semantics. IWCS 2013*. Potsdam, Nemčija. 397–403.

Verdonik, Darinka, Matej Rojc, Zdravko Kačič and Bogomir Horvat, 2002: Zasnova in izgradnja oblikoslovnega in glasovnega slovarja za slovenski knjižni jezik. Tomaž Erjavec and Jerneja Gros (eds.): *Zbornik konference Jezikovne tehnologije 2002*. Ljubljana: Institut Jožef Stefan. 44–48.

Verdonik, Darinka and Matej Rojc, 2004: Jezikovni viri projekta LC-STAR. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Jezikovne tehnologije: zbornik 7. mednarodne multi-konference Informacijska družba IS 2004*. Ljubljana: Institut Jožef Stefan. 24–47.

Verdonik, Darinka, Matej Rojc and Zdravko Kačič, 2004: Creating Slovenian language resources for development of speech-to-speech translation components. *Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC'04*. Lisbon, Portugal. 1399–1402.

Verdonik, Darinka and Ana Zwitter Vitez, 2011: *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Verdonik, Darinka, 2015: Govorjeni proti pisnemu ali katera leksika je »tipično govorjena«. Gorjanc Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 392–405.

Verlinde, Serge and Jean Binon, 2010: Monitoring dictionary use in the electronic age. Dykstra, Anne and Tanneke Schoonheim (eds.): *Proceedings of the XIV EURALEX International Congress*. Leeuwarden/Ljouwert: Fryske Akademy-Afûk. 1144–1151.

Vicknair, Chad, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen and Dawn Wilkins, 2010: A Comparison of a Graph Database and a Relational Database: A Data Provenance Perspective. *Proceedings of the 48th Annual Southeast Regional Conference* 42. New York: ACM.

Vičič, Jernej, 2012: *Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike*. Doktorska disertacija. Ljubljana: Fakulteta za racunalništvo in informatiko UL.

Vikør, Lars S., 2009: Lexicography and language planning in Scandinavia and the Netherlands. Nielsen, Sandro and Sven Tarp (eds.): *Lexicography in the 21st century. In honour of Henning Bergenholtz*. Amsterdam and Philadelphia: John Benjamins. 123−143.

Vintar, Špela, 2009: Samodejno luščenje terminologije – izkušnje in perspektive. Ledinek, Nina, Mojca Žagar Karer and Marjeta Humar (eds.): *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU. 345–356.

Vintar, Špela, 2010: Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16/2. 141–158.

Vintar, Špela and Tomaž Erjavec, 2008: iKorpus in lu**šč**enje izrazja za Islovar. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Zbornik* **Š***este konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 65–69.

Vintar, Špela and Darja Fišer, 2009: Adding multi-word expressions to sloWNet. Erjavec, Tomaž (ed.): *Mondilex Fifth Open Workshop: Research infrastructure for digital lexicography. Proceedings of the 12th International Multiconference Information Society 2009*. Ljubljana: Institut Jo**ž**ef Stefan. 56–63.

Vintar, Špela and Nataša Logar, 2015: Luščenje specializiranih izrazov za splošni slovar. Gorjanc, Vojko, Polona Gantar, Iztok Kosem and Simon Krek (eds.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 424–433.

Williams, John, 1996: Enough Said: The Problems of Obscurity and Cultural Reference in Learner's Dictionary Examples. Gellerstam, Martin (ed.): *Euralex ‚96 Proceedings I-II*. Gothenburg: Gothenburg University. 497–505.

Winkler, Birgit, 2001: English learners' dictionaries on CD-ROM as reference and language learning tools. *ReCALL* 13/2. 191–205.

Wright, Jon, 1998: *Dictionaries*. Oxford: Oxford University Press.

Zaidan, Omar F. and Chris Callison-Burch, 2011: Crowdsourcing Translation: Professional Quality from Non-Professionals. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, ZDA. 1220–1229.

Zakon o avtorskih in sorodnih pravicah (ZASP), 2007. *Uradni list RS* 16. https://www.uradni-list.si/1/content?id=78529.

Zgusta, Ladislav, 1971: *Manual of Lexicography*. The Hague.

Žganec Gros, Jerneja, Varja Cvetko-Orešnik and Primož Jakopin, 2006: SI-PRON: a comprehensive pronunciation lexicon for Slovenian. Erjavec, Tomaž and Jerneja Žganec Gros (eds.): *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006*. Ljubljana: Institut Jožef Stefan. 44–49.

Żmigrodzki, Piotr, 2014: Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. *Slovenščina 2.0* 2/2. 37–52.

II

Amazon Mechanical Turk. https://www.mturk.com/

ANW: Algemeen Nederlands Woordenboek. http://anw.inl.nl/

Clickworker. http://www.clickworker.com/en/

Collins. http://www.collinsdictionary.com/

Creative Commons. https://creativecommons.org/

Crowdcrafting. http://crowdcrafting.org/

CrowdFlower. http://www.crowdflower.com/

Daele: Diccionario de aprendizaje de español como lengua extranjera. http://www.daele.eu/

DDO: Den Danske Ordbog. http://ordnet.dk/ddo

OECD, 2004: *Declaration on Access to Research Data from Public Funding*. http://acts. oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157.

Diccionario de la lengua Española de la Real Academia Española. http://lema.rae.es/drae

Duolingo. https://www.duolingo.com/

DWDS: Das Digitale Wörterbuch der deutschen Sprache. http://www.dwds.de/

Eesti keele seletav sõnaraamat. http://en.eki.ee/dict/ekss

elexiko: Online-Wörterbuch zur deutschen Gegenwartssprache. http://www.owid.de/wb/ elexiko/start.html

Evroterm: večjezična terminološka zbirka. http://www.evroterm.gov.si/

FoldIt. https://fold.it/portal/

Groningen Meaning Bank. http://gmb.let.rug.nl/

Hrvatski enciklopedijski rječnik, 2003. http://hjp.znanje.hr (Croatian encyclopedic dictionary)

Igra besed. http://www.igra-besed.si/

Interactive Language Toolbox. https://ilt.kuleuven.be/inlato/

Islovar. http://www.islovar.org

Wiktionary. https://www.wiktionary.org/

GOS – a corpus of spoken Slovene. http://eng.slovenscina.eu/korpusi/gos

Korpus pisnih besedil: specifikacije postopkov za redno zbiranje tekstovnega gradiva za korpus, december 2008. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik1/ SSJ_Kazalnik_1_Specifikacije-pisni-korpus_v1.pdf.

Gigafida, a reference corpus of Slovene. http://eng.slovenscina.eu/korpusi/gigafida

IMP corpus of historical Slovene. http://nl.ijs.si/imp/index-en.html

SLD: Slovene Lexical Database. (Leksikalna baza za slovenščino). http://eng.slovenscina. eu/spletni-slovar/leksikalna-baza

Lexical Markup Framework. http://en.wikipedia.org/wiki/Lexical_Markup_Framework

Longman Dictionary of Contemporary English. http://www.ldoceonline.com/

Macmillan English Dictionary. http://www.macmillandictionary.com/

OECD Principles and Guidelines for Access to Research Data from Public Funding. http:// www.oecd.org/sti/sci-tech/38500813.pdf

Open Access Slovenia. http://www.openaccess.si/

Oxford dictionaries. http://www.oxforddictionaries.com/

Pedagoški slovnični portal. http://eng.slovenscina.eu/portali/pedagoski-slovnicni-portal

Phrase Detectives. http://anawiki.essex.ac.uk/phrasedetectives/

Phylo. http://phylo.cs.mcgill.ca/

Amebis Besana inflection dictionary (demo version). http://besana.amebis.si/pregibanje/

Priporočilo Komisije z dne 17. julija 2012 o dostopu do znanstvenih informacij in njihovem arhiviranju. http://eur-lex.europa.eu/LexUriServ/LexUriServ. do?uri=OJ:L:2012:194:0039:0043:SL:PDF.

PyBossa. http://pybossa.com/

Razvezani jezik. http://razvezanijezik.org/

Sloleks – Slovene morphological lexicon. http://eng.slovenscina.eu/sloleks

sloWNet – Slovene semantic lexicon. http://nl.ijs.si/slownet

Slovník spisovného jazyka českého. http://ssjc.ujc.cas.cz

sloWCrowd. http://nl.ijs.si/slowcrowd/

DNWSL: Dictionary of New Words of the Slovenian Language (Slovar novejšega besedja slovenskega jezika). http://www.fran.si/131/snb-slovar-novejsega-besedja

Slovene orthography, 2001. http://bos.zrc-sazu.si/sp2001.html

SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/

Fran dictionary portal. http://www.fran.si

DSLL: Dictionary of Slovene Literary Language (Slovar slovenskega knjižnega jezika). http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

DSLL2: Dictionary of Slovene Literary Language, 2nd edition (Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja). http://www.sskj2.si/

Šolar – *a developmental corpus of Slovene*. http://eng.slovenscina.eu/korpusi/solar

TEI P5: Guidelines for Electronic Text Encoding and Interchange, 2013. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Termania dictionary portal. http://www.termania.net

The Merriam-Webster Online Dictionary. http://www.merriam-webster.com

TheFreeDictionary. http://www.thefreedictionary.com/

Urban Dictionary. http://www.urbandictionary.com/

Vocabulary.com. http://www.vocabulary.com/

Web Content Accessibility Guidelines 2.0. http://www.w3.org/TR/WCAG20/.

Wielki słownik języka polskiego PAN. http://www.wsjp.pl

Wikislovar. http://sl.wiktionary.org/

Wordrobe. http://wordrobe.housing.rug.nl/

Wordsmyth.net. http://www.wordsmyth.net

All web pages in this book accessed 15. 3. 2016.

Name Index

Ž
Žabokrtský, Zdeněk
Žagar Karer, Mojca
Žáková, Adriána
Žele, Andreja
Žganec Gros, Jerneja

Ż
Żmigrodzki, Piotr