TRANSLATION STUDIES
AND APPLIED
LINGUISTICS

Univerza v Ljubljani
FILOZOFSKA
FAKULTETA

Edited by **Darja Fišer** and **Michael Beißwenger**

INVESTIGATING COMPUTER-MEDIATED COMMUNICATION: CORPUS-BASED APPROACHES TO LANGUAGE IN THE DIGITAL WORLD

Edited by Darja Fišer and Michael Beißwenger

INVESTIGATING COMPUTER-MEDIATED COMMUNICATION: CORPUS-BASED APPROACHES TO LANGUAGE IN THE DIGITAL WORLD

Book series Translation Studies and Applied Linguistics

Ljubljana 2017

INVESTIGATING COMPUTER-MEDIATED COMMUNICATION: CORPUS-BASED APPROACHES TO LANGUAGE IN THE DIGITAL WORLD
BOOK SERIES TRANSLATION STUDIES AND APPLIED LINGUISTICS
ISSN 2335-335X

Edited by: Darja Fišer and Michael Beißwenger
Reviewers: Martina Ožbot, Irena Stramljič Breznik
Editorial board: Špela Vintar, Vojko Gorjanc and Nike Kocijančič Pokorn
English language proofreading: Paul Steed
Layout: Jure Preglau

Published by: Ljubljana University Press, Faculty of Arts
Issued by: Department of Translation Studies
For the publisher: Branka Kalenić Ramšak, Dean of the Faculty of Arts, University of Ljubljana

Ljubljana, 2017
Digital edition.
Design: Kofein, d. o. o.
Publication is free of charge.
DOI: 10.4312/9789612379612

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. J6-6842).

The book is avaliable in e-form (PDF) at https://e-knjige.ff.uni-lj.si/

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=292036352
ISBN 978-961-237-961-2 (pdf)

Table of contents

Introduction

The increasing popularity of Web 2.0 has resulted in an unprecedented surge of user-generated and social media content which is becoming a major source of knowledge and opinion, and is considered a catalyst of bottom-up communication practices that contribute towards the democratization of language. As a consequence, we are seeing a growing need for a thorough multidisciplinary understanding of this type of communication that is significantly shaped by the specific social and technical circumstances in which it is produced: rich in colloquialisms and foreign language elements, non-canonical spelling variants and syntax, idiosyncratic abbreviations and neologisms.

What is more, this form of highly participatory, interactive and multimodal communication is accompanied by easily accessible and rich (sociodemographic) data, which open a wide range of new and exciting research opportunities, not only in linguistics and natural language processing, but also in the digital humanities and social sciences, as well as bringing about new technical, linguistic and ethical challenges for scholars.

The major bottleneck in the dissemination of corpora of computer-mediated content is not a technical one, as text retrieval from user-generated and social media platforms, such as chats, forums, weblogs and tweets, on social network sites and in wikis, is generally straightforward and sometimes even facilitated by native APIs. Instead, the main reason for the low number of publicly available corpora is the unclear legal status of computer-mediated communication (CMC) data when distributed as a resource to the scientific community, which is further exacerbated by the rapidly changing terms of service by content providers.

To address these issues, a growing number of projects all over Europe have started to create CMC corpora which are intended to be made available to the scientific community, and thus close the "CMC gap" in the corpus landscape (Beißwenger et al. 2017). Since 2013, the annual conference series *CMC and Social Media Corpora for the Humanities*[1] has been dedicated to the discussion of best practices on all aspects of open issues regarding the development, annotation, processing and analysis of CMC corpora among researchers who are building and processing these, along with representatives of language resource infrastructure initiatives such as CLARIN and DARIAH, and researchers in linguistics, digital humanities and social sciences who are using CMC data and corpora for the analysis of CMC phenomena in different languages and for different genres. The results of previous conferences have been published in the form of a special issue of the *Journal of Language Technology and Computational Linguistics* (Beißwenger et al. 2014), a monograph *Corpus de communication médiée par les réseaux: construction, structuration, analyse* (Wigham and Ledegen 2017) and as online conference proceedings (Fišer and Beißwenger 2016).

[1] http://www.cmc-corpora.org/

For the first time, the call for papers for this monograph was open also to authors who did not present their work at the conference. It includes eight contributions that have been selected from a total of 16 submissions based on a double-blind peer review. They are written by 16 authors from 13 institutions in 13 different countries dealing with the creation of CMC corpora and with the analysis of CMC phenomena in 10 different languages. Five of them are original papers and three are extended papers from the 2016 edition of the CMC-Corpora Conference that was held in Ljubljana, Slovenia. They tackle a diverse range of research questions and use a rich set of approaches, which is why we have organized them into four broad thematic and methodological parts: lexical analysis of CMC, sociolinguistic analysis of CMC, conversation and conflict in CMC, and building and processing CMC resources.

Part 1: Lexical analysis of CMC

Maja Miličević, Nikola Ljubešič and Darja Fišer investigate the universalities and specificities of communication in social media environments in a comparative analysis of spelling conventions on Twitter for three closely related languages: Slovene, Croatian and Serbian. This corpus-based study reveals that words from closed classes tend to be more often realized in non-standard spellings than words from open classes; that character deletions are more frequent than insertions or replacements; and that tweets in the three focal languages deviate from the written standard norms to different degrees. The datasets created for the study can be used as resources for further investigation of non-standard spelling conventions in the three languages.

Mohamed Tristan Purvis compiles a WhatsApp dataset to analyse the vocabulary that Hausa-speaking chatters adopt when consciously referring to their chat environment. The author shows that the interlocutors represented in his dataset not only code-mix with common English terms, but also widely employ Hausa words adapted for specialized reference to the online environment. The study analyses lexical, semantic and sociolinguistic factors that promote or constrain the adoption and use of Hausa words in chat terminology.

Part 2: Sociolinguistic analysis of CMC

Lieke Verheijen addresses the power conflict between the overt prestige of the (written) standard language and the covert prestige of the language used among young CMC users. In order to determine how the language used by the Dutch youth in CMC differs from Standard Dutch, the author presents an extensive register analysis of about 400,000 tokens of digital texts, produced by 12–23 year-old adolescents and young adults in SMS, instant messages and tweets. The study focuses on the orthographic, typographic, syntactic and lexical features of such texts. The results offer linguistic profiles of Dutch written CMC language for four new media genres and two age groups.

Steven Coats investigates the extent to which English is used on Twitter in the Nordic countries, with a special focus on the link between gender and grammatical or part-of-speech frequencies, a link which has hitherto been considered mainly in the context of data collected from L1 Anglophone contexts. The study uses a corpus of English-language messages originating from the Nordic countries which has been built using the Twitter Streaming API. It applies automatic methods to disambiguate author gender, assign part-of-speech tags, and determines the relative frequencies of grammatical types by gender and country. The analysis shows that Nordic English-language discourse on Twitter diverges according to gender for a number of grammatical features. The analysis supports L1 findings pertaining to gendered differences in feature frequencies in English.

Part 3: Conversation and conflict in CMC

Tatjana Scheffler examines the linguistic and structural features of German Twitter conversations. The study reveals that many well-known dialog phenomena can also be observed on Twitter, while at the same time the writers avail themselves of more formal, written-like options, while some spoken-like features take on new meanings. An analysis of the dialog structure shows that Twitter is not a homogeneous conversational genre, but that different types of conversations must be distinguished. Overall, the paper outlines several perspectives for further research on Twitter conversations.

Lydia-Mai Ho-Dac, Veronika Laippala, Céline Poudat and Ludovic Tanguy analyse the linguistic features of conflicts which occur on Wikipedia talk pages where authors of Wikipedia articles coordinate the collaborative writing task and process. Using a large corpus of talk pages from the French Wikipedia, they try to determine the linguistic cues that may help to identify and characterize conflicts on talk pages with two methods: supervised automatic classification of conflicting vs. harmonious discussion threads and multidimensional analysis of the data, to highlight key features on the genre of Wikipedia talk pages at a global level. The results open up perspectives for future work on automatic classification and analysis of conversational phenomena in large CMC corpora.

Part 4: Building and processing CMC resources

Solange Aranha and Paola Leone discuss the creation of a special type of learner corpus that contains Voice-over-IP (VoIP) interactions in which an L2 learner and an expert in the target language meet on a weekly basis, and which are conducted partially in the learner's L1 and partially in the learner's L2 (Teledandem interactions). Research on the Teledandem system is growing rapidly, as it can help to better understand and foster various language learning processes. based on the example of the DOTI database, which is currently composed of 700 hours of video data from Teledandem sessions, the authors discuss the relevant metadata, especially the characteristics of the learning scenarios, the tasks and activities observed in these, and the CMC environment.

Michael Beißwenger, Tobias Horsmann and Torsten Zesch discuss options for improving the treatment of sparsely represented linguistic phenomena that are of special interest for the annotation of linguistic corpora. The authors present a case study in which they used a PoS tagger to find one particular phenomenon of that type, and discuss several approaches for improving the identification of occurrences of this phenomenon in chats and tweets. The case study is Based on a PoS-tagged data set of 230 instances of German verb-pronoun contractions which can be retrieved from the CLARIN repository at IDS Mannheim.

We hope that this book is as inspiring and enjoyable to read as it was to edit. Our work would not have been possible without the dedicated work of all the authors who submitted their contributions, and without the careful and insightful comments of the reviewers who operated under a very tight deadline: Špela Arhar Holdt, Adrien Barbaresi, Tomaž Erjavec, Axel Herold, Nikola Ljubešić, Nataša Logar, Julien Longhi, Harald Lüngen, Maja Miličvić, Céline Poudat, Müge Satar, Tatjana Scheffler, Egon W. Stemle and Ciara R. Wigham. We would also like to thank the language editor Paul Steed for polishing the manuscripts, and for all the support and good spirits provided by Matevž Rudolf and Jure Preglau from the Faculty of Arts Publishing House.

Darja Fišer and Michael Beißwenger
Ljubljana, Slovenia and Essen, Germany
31 July 2017

References

Beißwenger, Michael, Nelleke Oostdijk, Angelika Storrer and Henk van den Heuvel, 2014: Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. *Journal of Language Technology and Computational Linguistics* 2/2014. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf.

Fišer, Darja and Michael Beißwenger (eds.), 2016: *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora2016)*. University of Ljubljana, Slovenia. http://nl.ijs.si/janes/cmc-corpora2016/proceedings/.

Wigham, Ciara R. and Gudrun Ledegen (eds.), 2017: *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse*. Paris: L'Harmattan (Humanités numériques).

Part 1
Lexical analysis of CMC

Birds of a feather don't quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese

Maja Miličević, *University of Belgrade*

Nikola Ljubešić, *Jožef Stefan Institute and University of Zagreb*

Darja Fišer, *University of Ljubljana and Jožef Stefan Institute*

Abstract

In this paper, we investigate the spelling conventions on the Twitter microblogging platform. In order to gain insight into the universalities and specificities of communication on social media, we perform a comparative analysis of three closely related languages: Slovene, Croatian and Serbian. The data collection and annotation protocols were developed jointly for all three languages, allowing for maximum interoperability and comparability of results. The analysis reveals differences in the amount of deviation from the norm in the three languages, with Slovene twitterese being the most inclined to using non-standard spelling, and Serbian the least. Overall, closed word classes, especially interjections and abbreviations, are found to be more non-standard than the open classes. In terms of types of standard > non-standard transformations, character deletions are more frequent than insertions or replacements, and transformations mostly occur in word-final positions. The discrepancies between languages are largely due to the pronounced tendency of Slovene and Croatian to use spoken-like, regional and dialectal forms characterised by vowel omissions, especially at the end of words. This analysis and the resulting datasets can be used to further study the properties of non-standard Slovene, Croatian and Serbian, as well as to develop language technologies for non-standard data in these languages.

Keywords: netspeak, Twitter, social media corpus, spelling variation, cross-lingual comparison

1 INTRODUCTION

Due to its increasing popularity and impact on society, computer-mediated communication (CMC) has been attracting a lot of attention in fields ranging from linguistics and communication studies to natural language processing and data analytics. CMC is seen as an important source of knowledge and opinions (Crystal 2011), as well as a prolific source of data on lexical and structural variation. CMC occurs under special technical and social circumstances (Noblia 1998), imposing specific communicative needs and practices (Tagg 2012). As a consequence, its language often deviates from the norms of traditional text production, instantiating numerous non-standard features at all levels, from unorthodox spelling to colloquial and other out-of-vocabulary lexis, as well as atypical syntax involving, for instance, frequent ellipsis and different uses, with and without syntactic value, of Twitter-specific elements such as @ mentions and hash tags (see, for example, Kaufmann and Kalita 2010, Arhar Holdt et al. 2016).

CMC has featured prominently in recent linguistic research, and of the three languages we focus on in this paper, Slovene CMC has been researched most extensively. An analysis of shortening strategies in tweets (Goli et al. 2016) showed a very strong tendency towards shortening among users, predominantly in the form of reductions at the orthographic level. Marko (2016), a study focused on neography, looked at letter/number homophones, showing that they occur equally frequently in foreign and Slovene words, and that the same symbol can have both a graphic (*g33k - geek*) and a phonetic use (*u3nek - utrinek / shooting star*). The influence of highly interactive and instantaneous communication platforms has been shown to blur the boundary between spoken and written discourse, resulting in the frequent use of phoneticised spelling, interaction words, deixis and non-standard lexis (Zwitter Vitez 2015).

When it comes to Croatian and Serbian, most attention in this field has centred on CMC in terms of SMS (Filipan-Žignić et al. 2012, Vrsaljko and Ljubomir 2013), Facebook (Vlajković 2010, Stamenković and Vlajković 2012), and chatroom messages (Radić-Bojanić 2007). The focus of such works has mostly been on the use of non-standard lexis (especially slang and Anglicisms) and deviations from orthographic rules, such as those concerning the use of capital letters and punctuation, as well as on non-standard spellings such as the use of *w* instead of *v*, or *sh* instead of *š*. Another prominent strand of research is the influence of new media language in the contexts of both education and literacy (Filipan-Žignić et al. 2015, Filipan-Žignić and Turk Sakač 2016), with the results showing that while pupils frequently use all the elements characteristic of new media in the texts written in their spare time, this does not interfere with their school assignments. Overall, even though some quantitative data have been reported, qualitative analysis and survey questionnaires prevail in these studies.

The two studies that are most directly related to the work presented in this paper are Fišer et al. (2015) and Miličević and Ljubešić (2016). The first compares tweets published in Slovene, Croatian and Serbian. It finds that, contrary to popular belief, most of the language used in tweets is fairly standard, especially in Slovene and Croatian. Another interesting finding was that the key characteristic of non-standard Slovene tweets is non-standard orthography, while non-standard lexis is more typical of Croatian, and especially Serbian. The second study looked only at Croatian and Serbian, detecting both similarities and differences between them. While some of the discrepancies were interpreted as being due to linguistic differences between the two languages (e.g. Croatian tends to drop final vowels to a higher extent than Serbian), others appear to be better explained by looking at extra-linguistic factors, such as user age, which seems to be lower in the case of Serbian, leading to a more chat-like format of messages. Both studies shared the finding that diacritics on letters such as *č*, *ć*, *š*, *ž* and *đ* are omitted more often in Serbian than in Croatian and Slovene.

In the present paper, we focus on posts from the Twitter microblogging platform written in Slovene, Croatian and Serbian. As one of the most widely used CMC platforms, Twitter has already received a lot of attention in linguistics. The average number of tweets published per day amounts to about 500 million,[1] and the content ranges from news broadcasts and official announcements by companies and institutions, to personal thoughts and opinions the users share, making Twitter a rich and easily accessible source of data for a wide range of (socio)linguistic inquiries. An additional component influencing the structural properties of its language is that tweets are limited to only 140 characters.

The analysis we report on is based on manually normalised, lemmatised and part-of-speech tagged samples of tweets in Slovene, Croatian and Serbian, created with the goal of developing tools for automatic CMC normalisation and tagging. In the remainder of the paper we first describe the corpora the tweets were sampled from and the samples themselves, moving on to the procedure and guidelines used in the manual normalisation. We then present the results of the analysis of normalisation. Specifically, we look at the distribution of standard-to-non-standard transformations across parts of speech and lemmas, as well as at the distribution of transformation types (deletions, insertions, and replacements), and compare these phenomena across the three datasets. Since very little related previous work is available for Slovene, Croatian and Serbian, our main goals are to give an overview of the key trends, and to compare them across languages. On the one hand, we investigate the degree to which spelling

[1] http://www.internetlivestats.com/twitter-statistics/

variations in the language of social media are universal, and on the other try to identify phenomena that are language-specific. In doing so, we treat all orthography-related phenomena as relevant for spelling, including word shortening and the expression of emphasis through letter repetitions.

2 CORPUS CONSTRUCTION AND SAMPLING

The corpora we employ comprise Slovene, Croatian and Serbian tweets harvested with TweetCat (Ljubešić et al. 2014), a custom-built tool for collecting tweets written in lesser-used languages. The collection of tweets for all three languages took place from 2013 to 2015, resulting in corpora of about 107 million tokens in Slovene, 25 million tokens in Croatian, and 205 million tokens in Serbian, after deduplication and filtering of foreign-language tweets and those without linguistically relevant content (i.e. those containing only mentions, links, or emoticons).

The initial samples used for the analysis presented in this paper were subsets of 4,000 tweets per language, each containing at least 100 characters, that were manually normalised, tagged and lemmatised (see Erjavec et al. 2016). These datasets were created to facilitate the development of processing tools for non-standard language, and for this reason they were sampled to represent tweets with different levels of technical and linguistic (non-)standardness (see Ljubešić et al. 2015). However, since the focus of this paper is on non-standard spelling variants, we only take into account the linguistically non-standard portion of the dataset, resulting in 1,983 tweets (54,688 tokens) in the original Slovene sample, 1,904 tweets (45,582 tokens) in the original Croatian sample, and 1,856 tweets (45,134 tokens) in the original Serbian sample.[2] After normalisation, the samples contain 54,955 Slovene tokens, 45,930 Croatian tokens and 45,322 Serbian tokens.

Examples of tweets containing non-standard features in Slovene, Croatian and Serbian are shown in Table 1. These features include phenomena typical of CMC in general, such as phonetic spelling of foreign words (e.g., *lajk* for *like*), omission of diacritics (e.g., *razrednicarka* for *razredničarka – teacher*), or shortenings (e.g., *yt* for *YouTube*), Twitter-specific phenomena like hashtags, @ name mentions and emoticons/emoji, as well as phenomena common in informal communication settings, such as the use of colloquial and dialectal non-standard forms (e.g., the Ikavian dialectal form *san* for *sam – am* in Croatian).

[2] A previous analysis of Croatian and Serbian (Miličević and Ljubešić 2016) was performed on tweets of all standardness levels.

Table 1: Sample tweets in Slovene, Croatian and Serbian (Original tweet [standard word form] // English translation).

Slovene
Original: @user99 vrjamm [Verjamem] ja :) nm [Nam] pa rece [reče] razrednicarka [razredničarka], da je naj do 6ihne [6-ih ne] budimo, in tko [tako] npr [npr.] smo bli [bili] ze [že] enkrt [enkrat] ob 4 zjutri [zjutraj] pred Louvrom :D
Translation: Yes, I believe you :) Our teacher told us not to wake her up before 6, so we were in front of the Louvre at about 4 a.m. already, for example. :D
Croatian
Original: Haha :-p nakon sta [što] san [sam] jucer [jučer] pricala [pričala] s iris [Iris] o supernaturalu, pocela [počela] sam sanjat [sanjati] one demone s creepy crnin [crnim] ocima [očima] ..... [...] brr
Translation: Haha :-p after talking to Iris about Supernatural yesterday, I started having dreams about those demons with creepy black eyes… Brr
Serbian
Original: Bad Copy i Sasa [Saša] Kovacevic [Kovačević] su skoro istovremeno objavili spotove veceras [večeras], a Bad Copy imaju vise [više] lajkova do sad na yt #geto #kvalitet
Translation: Bad Copy and Saša Kovačević published their videos almost simultaneously tonight, and up to now Bad Copy got more yt likes #ghetto #quality

3 NORMALISATION PROCEDURE AND GUIDELINES

The annotation process for all three languages was carried out using the web-based annotation platform Webanno (Eckart de Castilho et al. 2014). The annotation guidelines were first developed for the Slovene Twitter data within the Janes project (see Čibej et al. 2016), and then adapted for Croatian and Serbian based on the differences between the orthography and grammar manuals of the languages concerned. This resulted in a unified set of guidelines for the three languages, which is a big advantage in data-driven linguistics, as it enables direct cross-lingual comparisons.

For each language, each tweet was annotated independently by two annotators. A curation procedure followed, in which disagreements in the annotators' decisions were resolved. Tweets were annotated on five levels: token (i.e., corrections of word boundaries), sentence (sentence segmentation corrections), normalisation (i.e., standardisation of non-standard language features), lemmatisation (i.e., assignment of the canonical form to each word form in the running text, e.g., *objavili > objaviti – publish*) and morphosyntactic description (assignment of a morphosyntactic tag to each word in the running text following the MULTEXT-East v5.0 standard,[3] e.g., *demone – demons* > *Ncmsay* for *noun, common, masculine, singular, accusative, animate*). The complete annotation guidelines are available in the CLARIN repository,[4,5] and these are also summarised in the following subsections.

3.1 Segmentation and tokenisation

The samples were pre-tokenised and split into sentences with standard tools, and then checked manually by the annotators. Corrections at the sentence segmentation level relied on punctuation, if present, and on other symbols (e.g., name mentions designated with @, emoticons/emoji, and hashtags), in cases when they occupied a position where punctuation would normally be found. As for tokenisation, guidelines were provided for cases known to be problematic: hyphenated inflectional endings for abbreviations (e.g., *BMWu* for *BMW-u – at BMW* [locative]), cases where a vowel omission is marked by an apostrophe (e.g., in Serbian *pos'o* for *posao – job*), and abbreviations ending with a dot (e.g., *dr.* for *drugi – other*), which often lead to incorrect automatic splitting of a single token into two or three separate ones. An opposite case was that of word combinations containing hyphens, which are sometimes not separated into multiple tokens when they should be (e.g., in Slovene *Nemčija-Grčija* for *Nemčija – Grčija*).

3.2 Linguistic normalisation

In this paper we are most interested in the level of linguistic normalisation. In our case, the main goal of manual normalisation was to provide training data for building tools for automatic normalisation of CMC data. However, normalisation is also important for the end users of CMC corpora, as it enables them to perform queries based on standard forms, much along the lines of dialectal or diachronic data.

Normalisation was restricted to the word level, while word order, syntax, punctuation, ellipses, usernames, hashtags, emoticons/emoji and lexical choice (e.g., colloquial *komp* for *kompjuter – computer*) were not normalised. Normalisation

[3] http://nl.ijs.si/ME/V5/msd/html/

[4] Janes-smernice-v1.0.pdf at: http://hdl.handle.net/11356/1084

[5] ReLDI-NormTag-Guidelines.pdf at: http://hdl.handle.net/11356/1121

included the standardisation of non-standard spelling variants (e.g., in Slovene *jst > jaz – I*), as well as spelling and typing errors (e.g., in Croatian *popodme > popodne – afternoon*) and diacritic restoration (e.g., in Serbian *veceras > večeras – tonight*). A minimal intervention approach was adopted (e.g., in Slovene the non-standard variant *pucajne – cleaning* is normalised into the canonical non-standard variant *pucanje*, not into its standard equivalent *čiščenje*). In other words, we focused on non-standard forms that can be seen as spelling deviations, and not on style, grammar, or Twitter-specific phenomena. Context was to be taken into account when resolving unclear and ambiguous cases; if an issue could not be resolved from the available context, no normalisations were made.

While in most cases each non-standard token was normalized to one standard token, on rare occasions one non-standard token had to be split into multiple standard tokens (1:n mapping, *nevem – ne vem*, *do not know* in Slovene), and vice versa (n:1 mapping, *ni jedno – nijedno*, *neither* in Croatian). The percentage of tokens with the 1:n mapping is 0.47% in Slovene, 0.7% in Croatian and 0.39% in Serbian, while the n:1 mapping is observed with 0.06% Slovene tokens, 0.14% Croatian tokens and 0.07% Serbian tokens.

The following normalisation rules were applied in all languages (with the examples below coming from all three):

- Insert missing diacritics: *noz > nož – knife*
- Normalise foreign letters or letter combinations: *kavizza > kavica – coffee*
- Normalise non-standard spellings (regardless of whether they are regional forms, phonetic adaptations, or forms containing an obvious typo): *maš > imaš – have*
- Normalise cases of vowel omission or merging: *al > ali – but*
- Normalise non-standard inflectional endings: *živin > živim – I live*
- Normalise cases of missing sound assimilations: *rijedkost > rijetkost – rarity*
- Normalise lexical words in which some letters or syllables are repeated for emphasis; the same rule was applied to foreign words: *kaakooo > kako – how*
- Normalise interjections in which some letters or syllables are repeated for emphasis to two repetitions; the same rule was applied to foreign interjections: *hahaha > haha*
- Normalise words containing numbers instead of letters: *je2 > jedva – barely*

- Separate/merge words non-standardly written together/apart: *nebo* > *ne bo – will not*
- Add a hyphen before inflectional endings attached to abbreviations: *DS* > *DS-u – to DS*
- Add a dot to abbreviations missing one: *min* > *min. – minute*

Specific rules were applied to only one or two of the languages, due to linguistic differences, available reference resources or the need for upstream processing:

- Slovene: Do not normalise common deviations from prescriptive rules, such as incorrect preposition choice between *z/s – with*, or incorrect modal verb choice between *moči/morati – can/must*
- Croatian and Serbian: Spell out non-standard shortenings for words other than proper nouns: *msm* > *mislim* (*I think*) (in Slovene, this was not performed)
- Croatian and Serbian: Change *bi* (*would*) into standard inflectional forms *bih/bismo/biste* for the 1st person singular, 1st person plural and 2nd person plural respectively
- Slovene and Croatian: Normalise short infinitives into long infinitives (with the exception of future tense forms in Croatian): *vjerovat* > *vjerovati* (*believe*)
- Croatian: Normalise synthetic future forms into non-synthetic future forms: *biće* > *bit će* (*will be*)
- Croatian: Normalise long infinitives into short infinitives within future tense forms: *potpisivati ću* > *potpisivat ću* (*I will sign*)
- Croatian: Normalise dialectal interrogative pronoun forms *kaj* and *ća* to the standard form *što* (in Slovene, this was not performed)

Note that we distinguish between abbreviations, which tend to have a standard form (e.g. *min.* for *minute*), and shortenings, which are idiosyncratic. In the normalisation process, abbreviations were not expanded to their full form in either of the languages, while shortenings were kept in Slovene, and expanded in Croatian and Serbian. This is one of the very few differences in the guidelines, introduced due to the different needs related to the future use of the datasets in various different projects. In addition, abbreviations were assigned a dedicated PoS tag (see Section 4.2.1), while tags assigned to shortenings depended on what PoS classes they were normalised to (e.g. *msm* stands for *mislim – I think*, and was tagged as a verb).

4 DATA ANALYSIS

In this section we present the results of the analyses conducted on the normalised Slovene, Croatian, and Serbian Twitter datasets. Given that our normalisation guidelines were largely based on descriptive categories that are difficult to identify automatically (e.g., phonetic transcription or incorrect spelling), the analyses had to be adjusted to look at more readily identifiable criteria. We therefore decided to focus on transformations, i.e. character-level modifications that took place in non-standard language use compared to the standard. Note that this is the opposite from the normalisation process described in Section 3, where standard language forms were assigned to non-standard ones. For instance, in Section 3 we gave an example of the Croatian Ikavian verb form *živin*, which was normalised to the standard *živim* (*I live*); in the analyses presented in the remainder of the paper we treat this as a transformation of the standard *živim* into non-standard *živin* through character replacement.

We take into account the following: (1) original tokens, comparing them to (2) normalised tokens;[6] (3) morphosyntactic descriptions assigned to normalised tokens; and (4) lemmas assigned to normalised tokens. We study the frequency distribution of transformations by part of speech, and single out the most frequently transformed lemmas and surface forms. In addition, when looking at surface forms of normalised and original tokens, we classify the differences in terms of Levenshtein transformation types (deletions, insertions, replacements),[7] and we also look at the position of specific transformations within words.

Where appropriate, we use the log-likelihood (LL) statistical test to compare the frequencies of transformations between the three corpora. It has been argued that the LL test, similar to the chi-square test, is inappropriate as an inferential test for comparing corpus frequencies, given that word choice in corpora is not random, and words are not independent of one another (see Kilgarriff 1996). However, LL can be very useful as a measure for ranking differences between corpora, e.g. for finding words and/or tags that are distinctive of a corpus (Granger and Rayson 1998, Rayson 2002); we thus use the LL to identify those part-of-speech classes and transformation types on which non-standard Slovene, Croatian, and Serbian differ most, or look most alike.[8] To calculate the LL values, we use the pre-prepared Excel sheet created by Paul Rayson.[9]

[6] One original token could be normalised to up to four tokens, and multiple original tokens could be merged into a single normalised token (see Section 3.2).

[7] We do not include the transposition transformation from the Damerau-Levenshtein distance, as it has no linguistic grounding, but rather resolves non-intentional misspellings.

[8] Due to the shortness of individual tweets, alternatives such as the Mann-Whitney test, which takes individual texts rather than whole corpora as the unit of analysis, making sure that at least texts are independent of each other (Lijffijt et al. 2016), are not applicable in our case.

[9] http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx

Lastly, we should mention that in this study we do not control for sociolinguistic variables such as user age, education and location, or tweet topic; this is an additional reason for using the statistical tests for describing our samples rather than for drawing inferences. More specifically, while we are aware of the likely influence of at least some extra-linguistic variables, our initial goal was to provide a general overview of non-standard spelling in Slovene, Croatian and Serbian Twitter data. We leave a closer inspection of the contributions made by specific additional variables for future work.

4.1 Overall transformation frequency

The overall percentage of transformed tokens equals 17.39% (9,555 tokens) in Slovene, 13% (5,969 tokens) in Croatian, and 10.32% (4,679 tokens) in Serbian. However, many transformations are merely diacritic omissions (*č, ć, š, ž, đ* > *c, c, s, z, dj*), present for technical rather than linguistic reasons (possibly because typing on smartphones and international computer keyboards is faster without diacritics). After these are filtered out from the sample, we are left with 15.56% (8,552) transformed tokens in Slovene, 10.08% (4,628) transformed tokens in Croatian, and 3.96% (1,793) transformed tokens in Serbian. In line with the findings of previous works by Fišer et al. (2015) and Miličević and Ljubešić (2016), these numbers show that diacritics are most often omitted in Serbian, while Croatian and Slovene have a greater tendency towards non-standard forms beyond diacritic omission.[10]

4.2 Analysis by part of speech

The first analysis we focus on is based on the part-of-speech information assigned to each token in the normalised sample. We first compare the distributions of transformations by part of speech (i.e. among all transformations, how many belong to each PoS class) in Slovene, Croatian, and Serbian. We also look at the percentage of forms that have been transformed for each part of speech (i.e. out of all words that belong to a given PoS class, how many have undergone transformation) in each language. Both analyses are limited to the tokens that have undergone transformations other than diacritic omissions.

[10] The cross-lingual difference in the amount of diacritic omissions is most likely to be due to different rates of use of international keyboards on computers and the (non)availability of localized keyboards on smartphones. The reasons are unlikely to have a linguistic nature, so we do not look into this issue further, and focus on transformations that go beyond diacritic omission.

4.2.1 Distribution of transformations by part of speech

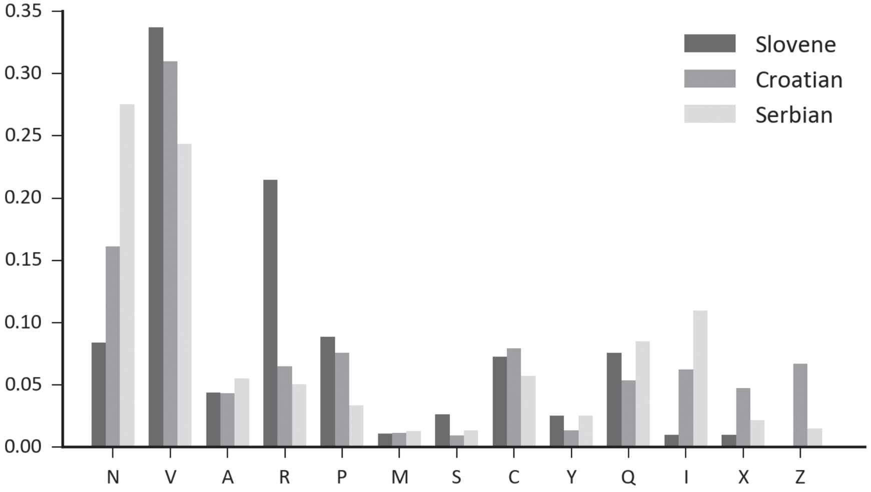


Figure 1: Distribution of transformed forms by part of speech in the Slovene, Croatian, and Serbian Twitter datasets.[11]

The relative frequencies of transformations by PoS are shown in Figure 1. It can be seen that despite the close relatedness of the three languages, some interesting differences emerge: while most transformations concern verbs in Slovene and Croatian, Serbian shows a more marked tendency towards noun transformation, with verbs coming second. Nouns occupy the second position in Croatian, but in Slovene they are preceded by adverbs (by a large margin) and pronouns (to a much lesser extent). It is also interesting to note that the rates of transformation in pronouns and prepositions are higher in Slovene than in the other two languages. Croatian takes the lead in the number of transformations of residuals, punctuation and conjunctions, whereas this is the case for adjectives, interjections and particles for Serbian.

The trends in Figure 1 are confirmed by log-likelihood values, which show that the difference between the three languages is most pronounced for adverbs (LL=649.66), with interjections coming second (LL=475.09), and nouns third (LL=412.03). On the opposite end of the spectrum, Slovene, Croatian and Serbian pattern together on numerals (LL=0.43), adjectives (LL=4.33), and conjunctions (LL=9.03). LL values for all parts of speech, as well as the raw frequencies they are based on, are reported in the Appendix (Table A1).

As will be shown in Section 4.3, verbal transformations in all three languages mostly belong to the auxiliary/copula *biti* (*be*), especially its 1st person singular form *sem*

[11] The tag values are as follows: N – noun, V – verb, A – adjective, R – adverb, P – pronoun, M – numeral, S – preposition, C – conjunction, Y – abbreviation, Q – particle, I – interjection, X – residual, Z – punctuation.

(often rendered as *sm*) and 3rd person singular past participle *bilo* (shortened to *blo*) in Slovene, and its 1st person singular preterite form *bih* (frequently realised as *bi*) in Croatian and Serbian. In addition, Slovene and Croatian are characterised by frequent transformations of other verbs through the shortening of the infinitive, e.g., *gledat* for *gledati – watch*, which is highly atypical of Serbian. Slovene adverbs are mostly shortened (e.g., *tako – so* frequently shortened to *tko*), but other kinds of transformations occur too. An interesting case is *zdaj – now*, which is transformed in three different ways in the dataset: *zdej*, *zdj* and *zj*. The transformations of interjections are mostly due to repeated vowels or syllables (e.g., *hahahaha*). Here, the differences across the languages are in all probability caused by minor differences in the application of the normalisation guidelines (e.g., despite the shared instructions, *ahaha* was normalised to *haha* in Croatian and Serbian, but left as *ahaha* in Slovene).

4.2.2 Shares of transformed forms within parts of speech

As for the percentages of forms that have been transformed within each part-of-speech class, Figure 2 shows that, overall, closed-class parts of speech tend to undergo more transformations than the open-class ones, with some differences between languages. The log-likelihood values indicate that Slovene, Croatian and Serbian differ the most on verbs (LL=1702.49), followed by adverbs (LL=1390.43) and pronouns (LL=734.56), while the classes that differ the least are numerals (LL=20.87), particles (LL=36.69), and abbreviations (LL=47.39). More detailed information is again provided in the Appendix (Table A2).

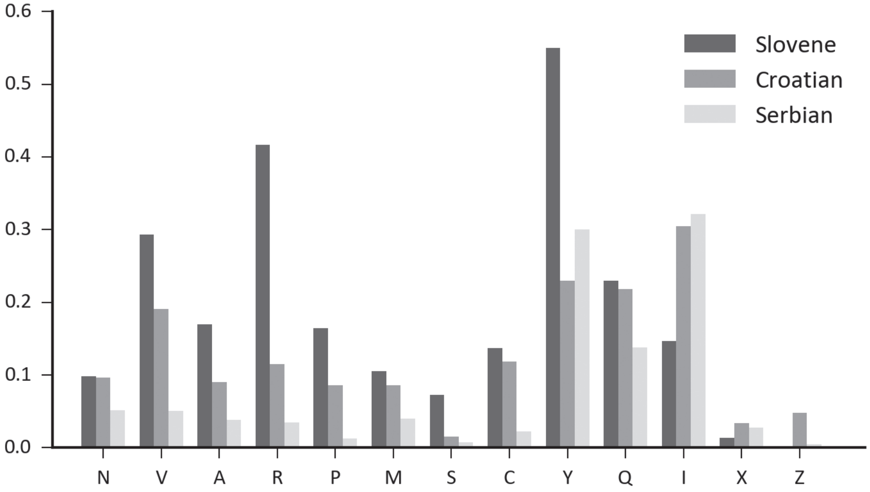


Figure 2: Shares of transformed forms within part-of-speech classes in the Slovene, Croatian, and Serbian Twitter datasets.

The highest percentage of transformed tokens in Slovene is found among abbreviations (mostly due to omissions of the final full stop, as in *slo*, used instead of *slo.* for *slovenski – Slovene*). In Croatian and Serbian it is the interjections that take the lead (mostly due to the aforementioned vowel or syllable repetitions, as in *hahahahaha*), followed by abbreviations (for the same reason as in Slovene), and particles (e.g., *neka – let it* is shortened to *nek*, and *je li – is it*, often merged and shortened to *jel*). Particles are transformed more in Croatian than in Serbian due to the more pronounced tendency of Croatian to omit final vowels in informal communication settings (cf. Sections 4.4 and 4.5). Conjunctions are another interesting case, as they have an overall low percentage of transformed tokens, but with about five times as many transformations in Slovene and Croatian as in Serbian. Similar to particles, most instances of transformed conjunctions are shortened versions with a (mostly final) vowel omitted. Some examples are *al* (from *ali – or* in Slovene / *but* in Croatian and Serbian), *il* (Croatian and Serbian *ili – or*), *kak* (in Slovene and Croatian, from *kako – how*), *ak* (Croatian, from *ako – if*). Pronouns are also transformed more often in Slovene and Croatian than in Serbian, but here the difference between Croatian and Serbian is mostly due to the frequent non-standard *ko* in place of the standard *tko – who*, and *šta* being used instead of *što* (*what*), while in Serbian *ko* and *šta* are the standard forms. In Slovene, the most frequent form is the 1st person singular personal pronoun *jaz - I*, commonly rendered as *jst*, *js, jest,* or *jz* instead.

Among the open part-of-speech classes, most transformations were detected for adverbs in Slovene, verbs in Croatian, and verbs and nouns in Serbian, which is consistent with the tendencies outlined for the distribution of transformations by PoS in Section 4.2.1. The trend of Slovene using more non-standard forms than Croatian, and especially Serbian, persists for adverbs, verbs, and adjectives. Interestingly, even though nouns prevail in the total percentage of transformations in Serbian, a look at within-PoS distributions reveals that more nouns actually undergo transformations in Slovene and Croatian, which can be traced back to the overall higher frequency of transformations in these two languages.

Overall, lexical word classes take up most transformations in the first comparison, while functional words take the lead in the second. In other words, despite the fact that lexical words are more frequent, a lower percentage of these are transformed, and this is why they dominate in Figure 1 but not Figure 2. From a linguistic point of view, however, this conclusion should be interpreted with caution, as some of the closed classes included in our analysis (abbreviations, residuals and punctuation), are not typically treated as PoS classes in linguistic analyses. While they do constitute a traditional PoS class, interjections too are a special case, as in our samples they mostly instantiate transformations based on repetitions, which have to do with emphasis and emotion and are not phonetic in nature (and were in addition normalised slightly differently in the three languages).

Finally, the PoS-based analyses confirm the initial observation that more non-standard spelling variants are used in Slovene and Croatian than in Serbian CMC. Multiple examples of the transformed tokens indicate that this might at least in part be due to a marked tendency of Slovene and Croatian towards vowel dropping. Before looking at this issue through Levenshtein transformations, we next present the results of the lemma- and surface form-based analyses.

4.3 Analysis by lemma and surface form

The set of analyses presented in this section focuses on the most frequently transformed lemmas (4.3.1) and surface forms (4.3.2).

4.3.1 Lemma analysis

The lemmas that underwent most transformations in each of the three datasets are shown in Table 2, where for each lemma we report the overall percentage of the transformed forms this lemma covers (% total), on which the lemma ranking is based, as well as the percentage of all forms of that lemma that were transformed (% lemma). We again disregard transformations due to diacritic omissions.

There is a high overlap among the lemmas on the lists of all three languages, with some variation in rank. The overall most frequently transformed forms come from the auxiliary verb *biti* (*be*), first-ranked in Slovene and Serbian, and second-ranked in Croatian. The full stop, ranked first in Croatian, does not make it to the Slovene list, and is ranked 17th in Serbian. Function words and interjections follow. The interrogative particle *li*, the conjunction *kao* (*as*), and the interjections *haha* and *hajde* (*let's*) are some examples of lemmas shared by Croatian and Serbian, while the conjunction *ali* (*or* in Slovene / *but* in Croatian/Serbian) appears in all three lists. Another interesting indirect match is between the Slovene and Croatian interrogative pronouns *kaj* and *što* (*what*), the former mostly appearing as *kej* or *kj*, and the latter as either *šta* (non-standard) or *kaj* (dialectal).[12]

As for the lexical words, adverbs dominate the Slovene lemma list, while verbs are equally present in all three lists. The verbs present in the Slovene and Croatian lists (other than *biti*) undergo most transformations in the infinitive form, where their final *i* is often omitted. The situation is more varied in Serbian, where the

[12] Recall from Section 3.2 that dialectal forms of the interrogative pronoun were normalised in Croatian (as an exception to the general ban on lexical intervention), but not in Slovene.

transformations of *hteti* (*want*) are mostly due to the drop of the initial *h*, as in *oću* (*hoću – I want*), while those of the slang verb *jebati* (*fuck*) are mostly caused by the high frequency of its non-standard past participle forms *jebo* and *jeb'o* (for *jebao*). Interestingly, another two forms of the same verb, functioning as interjections, also make it to the list (*jebote* and *jebiga, fuck* and *fuck it*), due to often being shortened to *jbt* and *jbg* respectively.[13] As for nouns and adjectives, none appear in any of the three lists.

Table 2: The 20 most frequently transformed lemmas in the Slovene, Croatian, and Serbian Twitter datasets.

Slovene			Croatian			Serbian		
Lemma	% total	% lemma	Lemma	% total	% lemma	Lemma	% total	% lemma
biti#V	8.33%	17.02%	.#Z	6.59%	15.16%	biti#V	7.53%	6.12%
jaz#P	3.24%	33.90%	biti#V	5.56%	12.21%	li#Q	6.53%	61.26%
tudi#Q	3.13%	82.21%	što#P	3.35%	62.50%	haha#I	2.90%	81.25%
imeti#V	3.09%	66.50%	haha#I	2.87%	77.78%	hajde#I	2.84%	92.73%
saj#C	1.61%	79.77%	ne#Q	2.38%	24.55%	hteti#V	2.01%	9.78%
potem#R	1.49%	73.41%	kao#C	2.33%	57.45%	ali#C	1.73%	19.38%
tako#R	1.39%	74.38%	li#Q	2.01%	61.18%	kao#C	1.51%	14.21%
zdaj#R	1.34%	76.16%	ali#C	1.71%	38.35%	jebati#V	1.45%	27.08%
malo#R	1.30%	82.22%	hajde#I	1.19%	93.22%	ne#Q	1.34%	4.86%
samo#Q	1.29%	61.45%	moći#V	1.17%	27.84%	jebote#I	1.23%	68.75%
lahko#R	1.20%	52.82%	htjeti#V	1.10%	12.78%	da#C	0.84%	1.07%
toliko#R	1.09%	91.18%	ako#C	0.84%	32.23%	jebiga#I	0.84%	83.33%
ne#Q	1.06%	11.15%	znati#V	0.82%	21.35%	moći#V	0.78%	8.19%
kaj#P	1.05%	36.29%	tko#P	0.82%	45.78%	min.#Y	0.78%	77.78%
kar#R	1.04%	70.08%	gdje#R	0.73%	87.18%	ja#P	0.73%	1.35%
ali#C	1.03%	63.77%	kako#C	0.65%	33.71%	u#S	0.67%	1.36%
videti#V	0.83%	76.34%	nešto#P	0.63%	34.12%	.#Z	0.61%	0.62%
misliti#V	0.81%	62.73%	ići#V	0.61%	30.43%	?#Z	0.61%	3.30%
kot#C	0.72%	32.46%	ili#C	0.58%	21.09%	ili#C	0.56%	8.85%
danes#R	0.70%	61.86%	tako#R	0.58%	36.99%	odmah#R	0.56%	50.00%

4.3.2 Surface form analysis

Moving on to surface forms, the 20 most frequent pairs of standard forms and their transformations are given in Table 3, omitting once again those that

[13] Note that idiosyncratic shortenings were expanded in Croatian and Serbian but not in Slovene.

only lack diacritics. The specific transformations are given in brackets, and the percentages these forms account for in the total number of transformations are also shown.

Table 3: The 20 most frequently transformed surface forms in the Slovene, Croatian, and Serbian Twitter datasets.

Slovene		Croatian		Serbian	
Form	% total	Form	% total	Form	% total
sem (sm)	3.37%	... (..)	5.68%	je li (jel)	3.99%
tudi (tud)	2.29%	kao (ko)	1.94%	li (l’)	1.81%
samo (sam)	1.93%	ali (al)	1.71%	ali (al)	1.56%
bilo (blo)	1.68%	je li (jel)	1.61%	hajde (aj)	1.50%
potem (pol)	1.39%	što (sta)	1.47%	jebote (jbt)	1.31%
saj (sej)	1.30%	što (šta)	1.40%	jebiga (jbg)	0.87%
tako (tko)	1.28%	bih (bi)	1.10%	min. (min)	0.87%
jaz (jst)	1.21%	... (....)	0.96%	kao (k’o)	0.81%
malo (mal)	1.21%	ako (ak)	0.89%	kao (ko)	0.78%
kar (kr)	1.10%	gdje (di)	0.86%	hajde (ajde)	0.75%
ali (al)	1.07%	što (kaj)	0.86%	bismo (bi)	0.62%
jaz (js)	1.03%	tko (ko)	0.77%	hajde (ae)	0.62%
zdaj (zdej)	0.97%	kako (kak)	0.72%	haha (hahaha)	0.56%
tudi (tut)	0.89%	haha (hahaha)	0.63%	odmah (odma)	0.50%
imam (mam)	0.76%	tako (tak)	0.61%	haha (hahah)	0.44%
pri (pr)	0.70%	hajde (ajde)	0.58%	bih (bi)	0.44%
ko (k)	0.70%	sam (san)	0.51%	ili (il)	0.44%
kaj (kej)	0.70%	ili (il)	0.51%	jebao (jebo)	0.44%
nekaj (neki)	0.66%	biti (bit)	0.49%	u stvari (ustvari)	0.44%
toliko (tolk)	0.66%	haha (hahah)	0.40%	li (l)	0.37%

The conjunction *al* is the only form shared between all three lists. While Slovene – expectedly – does not have any other forms in common with the other two languages, multiple additional forms are present in both Croatian and Serbian lists – for instance *jel* (*je li – is it*), *bi* (*bih – would*), and *ko* (*kao – like*). In Slovene *js* and *jst* instead of *jaz* (*I*) are very frequent, while all other forms instantiate either vowel replacement (typically *a>e*) or vowel omission, in different positions within words. In terms of PoS classes, most of the listed forms are adverbs. Ikavian forms (e.g., *di* for *gdje – where* and *san* for *sam – am*), as well as some final vowel omissions (*kak* for *kako – how*, *tak* for *tako – like that*, *ak* for *ako – if*, *bit* for *biti – be*) are specific to Croatian, while abbreviations such as *min* (*min.* for minute), and shortenings such as *jbt* (*jebote – fuck*) and *jbg* (*jebiga – fuck it*) are frequent only in Serbian.

4.4 Analysis by transformation type

In this section we present the probability distribution of the three types of Levenshtein transformations – deletions, insertions and replacements (Levenshtein 1966) for each language, again going from the normalised forms to the forms actually found in tweets. The results are summarised in Figure 3. The left half of the figure captures all transformations, and shows that while deletions are more frequent in Slovene than in Croatian, and in particular Serbian, the exact opposite is true of replacements. Insertions are most often found in Croatian, followed by Serbian, while they are very rare in Slovene. The high replacement rate in Serbian can be explained by its already mentioned pronounced tendency towards diacritic omission. Indeed, the right half of the figure, obtained after we discarded the tokens in which the transformation(s) consisted solely in the omission of diacritics, shows partly reversed trends: deletions and insertions become more frequent in Serbian than in Croatian (with deletions still less frequent than in Slovene), while Croatian outranks Serbian in the frequency of replacements. Overall, the most frequent transformation type is character dropping, followed by replacements, while insertions are the least frequent manifestation of the non-standard language used on Twitter.

We also performed log-likelihood tests on the data relative to the distribution of transformation types (without diacritics), confirming that insertions are the type

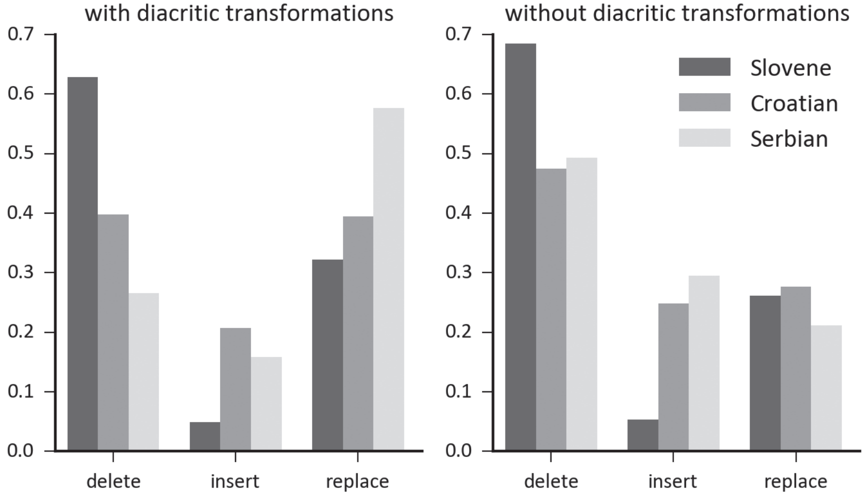


Figure 3: Comparison of transformation distributions in the Slovene, Croatian and Serbian Twitter datasets, with (left) and without (right) diacritic transformations.

that differs most between languages (LL=1723.79). Deletions occupy the second position (LL=400.71), while replacements reach the highest level of similarity in Slovene, Croatian and Serbian (LL=40.52). The raw frequencies that the LL values are based on are shown in Table A3 in the Appendix.

The next step in the analysis is to look at the most frequent specific transformations in each of the studied languages (again disregarding diacritic omissions). In Table 4 we show the top 10 transformations for each Levenshtein transformation type per language, together with a common example illustrating that particular transformation. The transformations are analysed at the level of single letters, so that digrams such as *lj* /lj/ are treated as two separate letters. However, special rules are added for treating 1:2 letter correspondences *đ* > *dj* and *ks* > *x* as single replacements rather than a replacement plus an insertion/deletion, as the latter approach would create a linguistically irrelevant bias in the frequency of *d* insertions and *k* deletions.[14] Moreover, an important and unavoidable consequence of the letter-by-letter approach is that many tokens contain multiple transformations defined on purely technical grounds (e.g. the definition of the Slovene transformation *potem* > *pol* is delete_t, delete_e, replace_m-l). Such transformations are not always linguistically relevant, and in some cases reflect technical decisions rather than linguistic regularities. The relative frequencies reported in Table 4 should thus be interpreted as primarily reflecting the technical side of the process, to which we add linguistic explanations in those cases where such explanations seem justified based on a qualitative analysis.

Table 4: The 10 most frequent transformations by language and type (with examples).

Slovene								
Deletions			Insertions			Replacements		
i	35.04%	tudi > tud	a	25.8%	pa > paa	l-u	14.65%	mogel > mogu
e	17.83%	sem > sm	h	14.97%	haha > hahah	a-e	13.32%	zdaj > zdej
o	13.30%	lahko > lahk	e	14.17%	ne > neee	j-i	5.21%	zjutraj > zjutri
a	11.23%	tako > tko	j	9.24%	ne > nej	o-u	4.37%	ono > uno
j	3.88%	skoraj > skor		4.62%	odkar > od kar	a-s	4.19%	jaz > jst
	3.10%	ne bi > neb	o	4.14%	zelo > zelooo	m-l	4.09%	potem > pol
.	2.79%	npr. > npr	s	3.98%	imate > maste	a-o	3.98%	danes > dons
t	2.73%	potem > pol	i	3.82%	vsak > saki	z-s	3.95%	jaz > js
d	1.77%	tudi > tut	u	3.82%	super > suuuper	z-t	3.88%	jaz > jst
u	1.26%	tule > tle	m	2.71%	bi > bim	i-t	3.57%	tudi > tut

[14] *Dj* is an alternative, non-standard spelling of the grapheme *đ*, while *x* is completely absent from the alphabets of the languages we study, which use *ks* instead (as in *maksimum* rather than *maximum*).

Croatian								
Deletions			Insertions			Replacements		
i	24.08%	kupiti > kupit	a	26.20%	na > naa	o-a	10.89%	što > šta
	9.51%	je li > jel	h	15.85%	haha > haahhhaaa	e-i	9.59%	treba > triba
.	9.07%	2013. > 2013	o	13.46%	to > tooo	m-n	7.45%	sam > san
a	8.49%	neka > nek	e	10.73%	najviše > najvišeee	o-j	3.27%	što > kaj
j	8.14%	vridi > vrijedi	.	6.40%	npr > npt.	a-e	3.16%	pasje > pesje
o	7.39%	kao > ka	i	6.23%	ti > tii	t-a	2.99%	što > kaj
e	7.10%	čovik > čovjek	u	3.39%	Au > Auuu	š-k	2.93%	što > kaj
h	5.84%	hajmo > ajmo	j	2.56%	falio > falija	o-l	1.86%	kupio > kupil
t	3.90%	netko > neko		2.17%	A ha > Aha	ć-č	1.64%	već > več
d	2.50%	budeš > buš	s	2.00%	sereš > seress	i-’	1.52%	velike > vel’ke
Serbian								
Deletions			Insertions			Replacements		
i	13.62%	li > l	a	22.51%	jao > jaao	i-’	7.49%	ali > al’
e	10.95%	hajde > aj	h	12.63%	hehe > heheheh	a-’	5.05%	ostao > ost’o
a	10.67%	kao > ko	e	11.59%	umrla > umrela	ks-x	3.06%	faks > fax
	10.33%	je li > jel	.	9.97%	… > ………	i-e	2.45%	zaspi > zaspe
h	5.96%	hladan > ladan	o	6.36%	Alo > Aloo	š-h	2.29%	šiša > shisha
o	5.90%	jebote > jbt	i	5.03%	ima > iiima	h-’	2.14%	hoće > ’oće
d	4.03%	hajdmo > hajmo		3.89%	trebaće > treba će	e-i	2.14%	živce > živci
j	3.97%	mi je > mie	!	3.61%	!!! > !!!!	a-e	1.99%	nove > nova
u	3.58%	ne mogu > nmg	u	3.04%	juhu > juhuuuu	h-x	1.83%	hehe > xexe
-	3.46%	sms-a > smsa	?	2.85%	?! > ??!!	r-v	1.53%	smrde > smvde

4.4.1 Analysis of deletions

The most frequent deletions in all three languages are those of vowels and blank spaces. In Slovene, most deletions concern the vowel *i* (taking up over one third of all deletions), followed by *e*, *o*, and *a*. The vowels are omitted both word-finally (*tudi* > *tud* – *also*) and word-internally (*tako* > *tko* – *both*). They are followed by *j*, deletions of which are much less frequent, and similar in number to those of the blank space, full stop, *t*, *d*, and *u*. In Croatian, too, the most frequent cases, close to one quarter, are omissions of *i* (as in *al* for *ali* – *but*, and *kupit* for *kupiti* – *buy*). *I* is followed by the blank space (due to the merging of words such as *jel* for *je li* – *is it*), the dot (either within punctuation, or in abbreviations, as in *npr* for *npr.* – *e.g.*), *a* (e.g. in shortenings such as *ko* for *kao* – *like* and *nek* for *neka* – *let it*), and *j* (often due to the use of the Ikavian yat reflex *i* instead of the Ijekavian *(i)je*, as in *di* for *gdje – where*, or *uvik* for *uvijek – always*). In Serbian, the most frequent omissions are those of *i* (as in *jel* for *je li – is it*, *al* for *ali – but*), *e* (in shortenings like *aj* for *hajde – come on*, or *jbg* for *jebiga – fuck*), *a* (in shortened forms such as *ko* for *kao – like*, or *reko* for *rekao – said*), and the space (in merged words like *jel* for *je li – is it*, or *ustvari* for *u stvari – actually*). However, Serbian does not have a dominant deletion pattern similar to that of *i* in Slovene and Croatian.

4.4.2 Analysis of insertions

Insertions are mostly the result of expressive multiplication of syllables (e.g., *hahahahaha*) or vowels (e.g., in Slovene *zelooo – very*), in interjections and lexical words. The second most frequent category of insertions are strings of two words that were erroneously spelled as separate (e.g., *treba će* instead of *trebaće – will need* in Serbian). What follows are words that use foreign or idiosyncratic spelling for domestic words (e.g., in Croatian *bass* for *baš – very; right*), non-canonical abbreviation expansions (e.g., *esemes* for *sms* in Serbian), and dialectal forms that are longer than the standard ones (e.g., *falija* instead of *falio – lacked; missed* in Croatian).

4.4.3 Analysis of replacements

As for replacements, the most frequent case in Slovene is the *l* > *u* transformation in verbal past participles (*napisal* > *napisu – wrote*, *mogel* > *mogu – could*, *mislil* > *mislu – thought,* etc.); the second in frequency is *a* > *e* (*kaj* > *kej – what*, *zdaj* > *zdej – now*). In Serbian, replacements mostly cover the marking of character omissions with an apostrophe (as in *je l'* for *je li – is it*, or *ost'o* for *ostao – he stayed*), a phenomenon virtually non-existent in Croatian and Slovene. In Croatian, there are three frequent cases: *e-i* (due to the use of the Ikavian yat reflex, as in *triba* for *treba – needs*), *o-a* (in the substandard pronoun variant *šta* (*što – what*), and the southern dialectal endings of present participles like *falija* (*falio – lacked; missed*)), and *m-n* (transformation of the standard ending *m* in the southern dialect, as in *san* (*sam – I am*) or *van* (*vam – to you*)).

4.5 Analysis by position of transformation

In this section we focus on the position of transformations (deletions, insertions, and replacements) within words (with diacritic omissions once again excluded). In

Figure 4 we show the overall positional distributions of all transformations for Slovene, Croatian, and Serbian, while the following three panels (Figures 5, 6 and 7) show the results for the relative positions of deletions, insertions, and replacements.

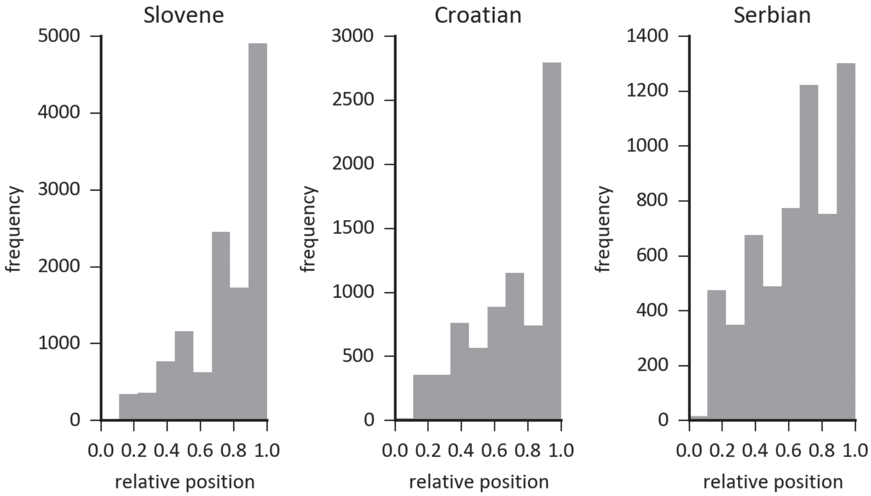


Figure 4: Distributions of transformations by position, for Slovene, Croatian, and Serbian.

The overall trend that emerges in the first set of histograms (Figure 4) is that transformations mostly occur at the word end, and only rarely at the beginning. The same trend is evident in all three languages, with Serbian standing out for its least marked bias towards word-final modifications in non-standard language.

Fairly similar trends are also found in all three languages for specific types of transformations. Deletions, as can be seen in Figure 5, are very biased towards the word end in Slovene, and even more so in Croatian, largely due to final vowel deletions (mostly in function words and infinitives, as outlined in Sections 4.2 and 4.3). Deletions are somewhat more evenly distributed across the word in Serbian, and not only because final vowel dropping is not as common in this language. Recall that in Serbian some of the most frequently transformed surface forms are rendered as shortenings, involving deletions at various positions within words, e.g., *jbg* < *jebiga*, *nzm* < *ne znam* (see Table 3 in Section 4.3). A tendency towards reducing words and entire phrases to shortenings is less present in Croatian, while in Slovene such phenomena were not normalised (see Section 3.2).

Insertions (Figure 6) and replacements (Figure 7) show similar distributions in all three languages, having overall an even stronger tendency towards the end of the

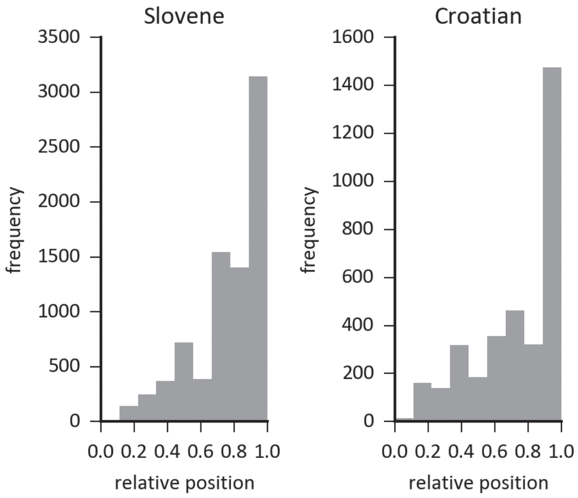


Figure 5: Distributions of deletions by position, for Slovene, Croatian, and Serbian.

word. For insertions, a closer inspection reveals that most cases are in fact expansions via repetitions of the final vowel. End-of-word replacements are largely accounted for by the *l* > *u* verb ending transformation in Slovene, the *o* > *a* in *što* > *šta* (*what*) and *m* > *n* in ending transformations on verbs in Croatian, and word-final vowel-to-apostrophe transformations in Serbian (e.g., *ali* > *al' – but*).

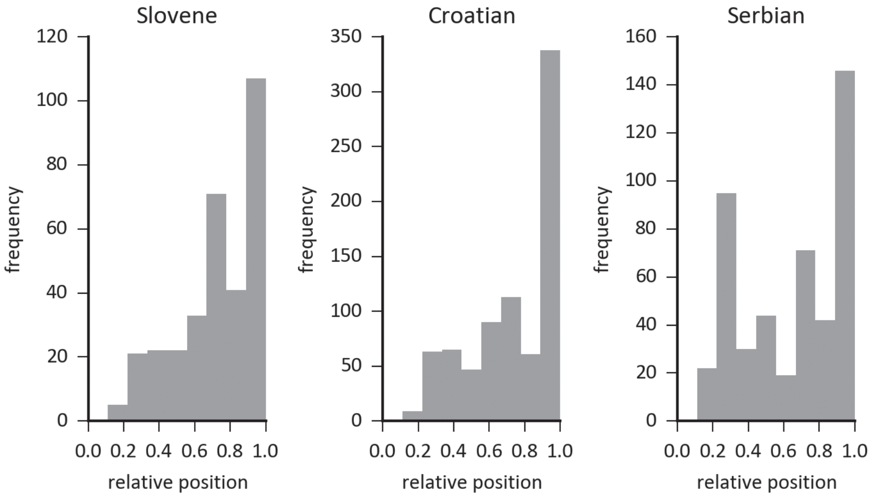


Figure 6: Distributions of insertions by position, for Slovene, Croatian, and Serbian.

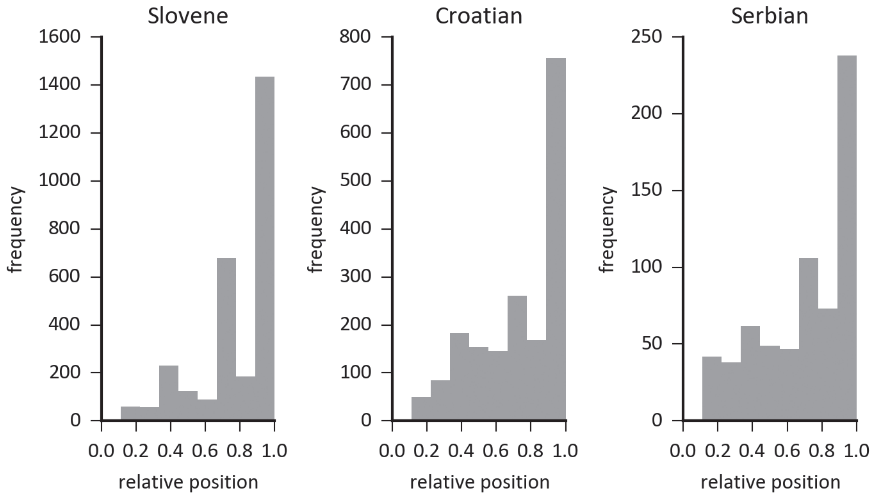


Figure 7: Distributions of replacements by position, for Slovene, Croatian, and Serbian.

5 CONCLUSION

In this paper we analysed a sample of Slovene, Croatian and Serbian tweets that were manually normalised by following unified annotation guidelines. Looking at the overall frequency of transformations, we established that the non-standard Serbian used on Twitter shows a greater tendency towards omitting diacritics, while its Slovene and Croatian equivalents are more prone to using other kinds of non-standard forms. The distribution of transformations by part of speech is such that the largest portion is occupied by open word classes (adverbs in Slovene, verbs in Croatian, and nouns in Serbian). However, looking within specific parts of speech, the most prominent transformations are those on closed classes, as confirmed by the lemma-based analysis, which revealed that the most frequently transformed lemmas belong to the classes of auxiliary verbs, interjections, and conjunctions.

By calculating the frequencies of Levenshtein transformations we observed that, leaving aside diacritic omissions, the most frequent transformations are deletions, as expected not only based on the general principle of language economy, but also due to the informal, highly interactive communication setting and frequent use of portable communication devices with suboptimal keyboards. Deletions are particularly present in Slovene, where insertions are less common than in

Croatian and Serbian. Across languages, deletions mostly consist of vowel droppings that resemble colloquial spoken language, while insertions are largely cases of expressive/emphatic vowel and syllable repetitions, especially in interjections. The picture is more varied for replacements, which also differ the most among the languages, and mostly include transformations into colloquial forms (especially in Serbian) and regional/dialectal variants (especially in Slovene and Croatian). Finally, we found that transformations are mostly word-final and very infrequently word-initial, especially in Slovene and Croatian, which is again characteristic of the colloquial spoken varieties.

While the goal of this paper was not to test specific linguistic hypotheses, we did identify some interesting spelling variation patterns. First of all, even though deletions were found to be the most typical transformation in all three languages, and vowels were consistently dropped the most in non-standard language, we also confirmed the tendency of Slovene and Croatian twitterese to omit these more often than their Serbian counterpart, especially in word-final positions. This tendency appears to be largely linguistic in nature, and mirrors the properties of the spoken varieties of the languages in question, and some historical dialectal differences (e.g. the wide presence of short infinitives in some dialects, see Stevanović 1986).

On a more sociolinguistic side, more shortenings seem to be used in non-standard Serbian than in non-standard Croatian (no data is available for Slovene, as its shortenings were not normalised). The exact reasons for this are yet to be established, given that the communicative and practical constraints are shared. One possible technical explanation is that shortenings are used in Serbian in order to gain the space that Croatian frees through single-vowel droppings. Another hypothesis is that Serbian twitterese is more "playful," and that its users (who might belong to a different demographic than those in Croatia or Slovenia) use language in a particularly creative way. On the other hand, more regional and dialectal forms are used in Slovene and Croatian twitterese than the Serbian version, which could perhaps be traced back to differences in the official language policies of the three countries, and in how much different dialects are used and how they are viewed.

The overall picture thus seems to be one of a (socio-)linguistic non-standardness continuum going from Slovene to Serbian. What is particularly interesting is that Croatian patterns with Slovene in several respects when it comes to the non-standard language, despite the standard language of Croatian being overall much closer to Serbian, linguistically and historically. These conclusions should of course be tested in a more controlled manner in future work, and while some of the results that lead us to them might have been affected by minor discrepancies in the normalisation guidelines for the three languages, the tendencies seem robust enough to provide motivation for further studies.

In sum, given the relative scarcity of large-scale empirical data on Slovene, Croatian and Serbian CMC, the analyses reported in this work are intended to provide a valuable first insight into the nature of deviations from their norms, and to serve as a starting point for more focused studies of the linguistic phenomena at hand. In the future, our study could be complemented with an analysis of the impact of socio-demographic factors, such as user age or geographic location, on the observed transformations. Another topic that would be interesting to explore in future work would be a lexical analysis of CMC, i.e. a study of standard > non-standard lexical transformations. Such cases are not captured in our current normalisation guidelines, but previous work by Fišer et al. (2015) indicates that they are highly relevant for cross-linguistic comparisons, as Slovene was found to make less use of non-standard lexis than Croatian and Serbian.

Acknowledgements

The research described in this paper was funded by the Slovenian Research Agency national basic research project J6-6842 *Resources, Tools and Methods for the Research of Nonstandard Internet Slovene,* the Swiss National Science Foundation (through *ReLDI – Regional Linguistic Data Initiative*, an institutional partnership among the Universities of Zurich, Belgrade and Zagreb, within the Scientific Co-operation between Eastern Europe and Switzerland programme; project No. 160501, 2015-2017), and the Ministry of Education of the Republic of Serbia (through the national research project *Standard Serbian Language: Syntactic, Semantic and Pragmatic Explorations*; project No. 178004, 2011-2016).

We are also grateful to our annotators, who made the analyses reported in this paper possible, and to the reviewers for their valuable comments.

References

Arhar Holdt, Špela, Darja Fišer, Tomaž Erjavec and Simon Krek, 2016: Syntactic annotation of Slovene CMC: First steps. Fišer, Darja and Michael Beißwenger (eds.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Academic Publishing Division of the Faculty of Arts. 3–6. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Arhar_et_al_Syntactic-Annotation-of-Slovene-CMC.pdf. (Last accessed 29 June 2017.)

Crystal, David, 2011: *Internet Linguistics: A Student Guide*. New York: Routledge.

Čibej, Jaka, Darja Fišer and Tomaž Erjavec, 2016: Normalisation, tokenisation and sentence segmentation of Slovene tweets. Andrius, Utka, Jurgita Vaičenonienė and Rita Butkienė (eds.): *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe)*, *LREC 2016*. 5–10. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf. (Last accessed 29 June 2017.)

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych and Seid Muhie Yimam, 2014: WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*. Soesterberg, Netherlands. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf. (Last accessed 29 June 2017.)

Erjavec, Tomaž, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić and Darja Fišer, 2016: Gold-standard datasets for annotation of Slovene computer-mediated communication. *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2016)*. Brno, Czech Republic. https://nlp.fi.muni.cz/raslan/2016/paper06-Erjavec_etal.pdf. (Last accessed 29 June 2017)

Filipan-Žignić, Blaženka, Katica Sobo and Damir Velički, 2012: SMS communication – Croatian SMS language features as compared with those in German and English speaking countries. *Revija za elementarno izobraževanje* 5. 5–22.

Filipan-Žignić, Blaženka, Vladimir Legac, Tea Pahić and Katica Sobo, 2015: New literacy of young people caused by the use of new media. *Procedia – Social and Behavioral Journal* 192. 172–179.

Filipan-Žignić, Blaženka and Marija Turk Sakač, 2016: Utjecaj novih medija na jezik mladih u pisanim radovima. *Slavistična revija* 4. 463–474.

Fišer, Darja, Tomaž Erjavec, Nikola Ljubešić and Maja Miličević, 2015: Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. Smolej, Mojca (ed.): *Simpozij Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)*. Ljubljana: Filozofska fakulteta. 225–231.

Goli, Teja, Eneja Osrajnik and Darja Fišer, 2016: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Erjavec, Tomaž and Darja Fišer (eds.): *Proceedings of the Language Technologies and Digital Humanities Conference*. Ljubljana, Slovenia. 77–82. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Goli-et-al_Analiza-krajsanja-slovenskih-sporocil.pdf. (Last accessed 29 June 2017.)

Granger, Sylviane and Paul Ryson, 1998: Automatic profiling of learner texts. Granger, Sylviane (ed.): *Learner English on Computer*. London: Longman. 119–131.

Kaufmann, Max and Jugal Kalita, 2010: Syntactic normalization of Twitter messages. *International Conference on Natural Language Processing (ICON 2010)*. Kharagpur, India. 149–158.

Kilgarriff, Adam, 1996: Which words are particularly characteristic of a text? A survey of statistical approaches. Evett, Lindsay J. and Tony G. Rose (eds.): *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex University. 33–40.

Levenshtein, Vladimir I., 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8. 707–710.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki and Heikki Mannila, 2016: Significance testing of word frequencies in corpora. *Literary and Linguistic Computing* 31/2. 374–397.

Ljubešić, Nikola, Darja Fišer and Tomaž Erjavec, 2014: TweetCaT: a tool for building Twitter corpora of smaller languages. Calzolari, Nicoletta et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2279–2283. http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf. (Last accessed 29 June 2017.)

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak and Iza Škrjanec, 2015: Predicting the level of text standardness in user-generated content. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*. 371–378. https://aclweb.org/anthology/R/R15/R15-1049.pdf. (Last accessed 29 June 2017.)

Marko, Dafne, 2016: The use of alphanumeric symbols in Slovene tweets. Fišer, Darja and Michael Beißwenger (eds.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Ljubljana University Press (Faculty of Arts). 48–53. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Marko_Use-of-Alphanumeric-Symbols-in-Slovene-Tweets.pdf. (Last accessed 29 June 2017.)

Miličević, Maja and Nikola Ljubešić, 2016: Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0* 4/2. 156–188. http://dx.doi.org/10.4312/slo2.0.2016.2.156-188. (Last accessed 29 June 2017.)

Noblia, Maria Valentina, 1998: The computer-mediated communication: A new way of understanding the language. *Proceedings of the 1st Conference on Internet Research and Information for Social Scientists (IRISS'98)*. 10–12.

Radić-Bojanić, Biljana, 2007: *neko za chat?! Diskurs elektronskih ćaskaonica na engleskom i srpskom jeziku*. Novi Sad: Filozofski fakultet.

Rayson, Paul, 2002: *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD dis., University of Lancaster.

Stamenković, Dušan and Ivana Vlajković, 2012: Jezički identitet u komunikaciji na društvenim mrežama u Srbiji. Mišić-Ilić, Biljana and Vesna Lopičić (eds.): *Jezik, književnost, komunikacija: zbornik radova. Jezička istraživanja*. Niš: Filozofski fakultet. 212–224.

Stevanović, Mihailo, 1986: *Savremeni srpskohrvatski jezik (gramatički sistemi i književnojezička norma. I Uvod, fonetika, morfologija* (5th ed.). Belgrade: Naučna knjiga.

Tagg, Caroline, 2012: *Discourse of Text Messaging*. London: Continuum.

Vlajković, Ivana, 2010: Uticaji engleskog jezika na srpski na planu pravopisa, leksike i gramatike u komunikaciji na Fejsbuku. *Komunikacija i kultura online* 1. 183–196.

Vrsaljko, Slavica and Tea Ljubomir, 2013: Narušavanje pravopisne norme u ranojezičnoj neformalnoj komunikaciji (na primjeru SMS poruka i internetske društvene mreže Facebook). *Magistra Iadertina* 8/1. 155–163.

Zwitter Vitez, Ana and Darja Fišer, 2015: From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. *Electronic lexicography in the 21st century: linking lexical data in the digital age: Proceedings of eLex 2015 Conference*. Ljubljana: Trojina, Institute for Applied Slovene Studies, Brighton: Lexical Computing. 250–267.

APPENDIX

Table A1: Raw frequencies and log-likelihood values for transformations by part of speech in the Slovene, Croatian, and Serbian Twitter datasets.

PoS	Slovene	Croatian	Serbian	LL
M	94	53	23	0.43
A	376	201	99	4.33
C	623	368	103	9.03
Y	219	62	46	23.94
Q	647	248	153	28.82
V	2883	1435	437	44.41
S	227	43	24	54.12
P	760	351	60	70.82
Z	0	311	27	84.31
X	86	220	39	171.55
N	718	746	494	412.03
I	84	288	197	475.09
R	1835	302	91	649.66
Total	**8552**	**4628**	**1793**	---

Table A2: Raw frequencies and log-likelihood values for transformations within part-of-speech classes in the Slovene, Croatian, and Serbian Twitter datasets.

PoS	Number of transformations			Total number of tokens			LL
	Slovene	Croatian	Serbian	Slovene	Croatian	Serbian	
M	94	53	23	891	619	575	20.87
Q	647	248	153	2814	1136	1110	36.69
Y	219	62	46	398	270	153	47.39
I	84	288	197	572	944	613	48.28
X	86	220	39	6415	6420	1416	61.31
N	718	746	494	7291	7745	9531	161.26
A	376	201	99	2215	2219	2611	221.98
S	227	43	24	3137	2739	3146	229.69
Z	0	311	27	7828	6526	5695	243.20
C	623	368	103	4553	3103	4508	444.18
P	760	351	60	4617	4065	4797	734.56
R	1835	302	91	4401	2623	2592	1390.43
V	2883	1435	437	9823	7521	8575	1702.49

Table A3: Raw frequencies and log-likelihood values by transformation type in the Slovene, Croatian, and Serbian Twitter datasets.

Transformation type	Slovene	Croatian	Serbian	LL
Deletions	7962	3439	1762	400.71
Insertions	628	1798	1053	1723.79
Replacements	3038	1998	758	40.52
Total	**11628**	**7235**	**3573**	**---**

CMC terminology in Hausa as found in a corpus of WhatsApp chats

***Mohamed Tristan Purvis,** American University of Nigeria*

Abstract

A corpus of WhatsApp chats reveals how Hausa-speaking youth have adopted and spread homegrown Hausa terms, via semantic extension, for the actions (e.g. chatting, forwarding), objects (e.g. image) and space (e.g. group, online/offline) associated with computer-mediated communication rather than strictly borrowing from English chat jargon. This study reviews the linguistic forms (including source language), range of terminology, and frequency of occurrence of chat environment-related terminology found in this corpus, representing 56 different interlocutors in 40 different dyads of chat excerpts. Primary consideration is given to lexical and semantic factors that promote or constrain the adoption of Hausa words in chat terminology, but preliminary consideration is also given to sociolinguistic factors.

Keywords: Hausa, chat jargon, semantic extension, lexical borrowing, corpus development

1 INTRODUCTION

This study analyses the vocabulary that Hausa-speaking chat participants (chatters) adopt when consciously referring to the chat environment itself. In particular, I analyse the extent to which chatters either draw on English-based chat jargon or employ equivalent Hausa terms for this purpose. Observations are drawn from a freshly developed corpus of WhatsApp chats between Hausa speakers. The corpus includes 40 different dyads of chats involving 56 different interlocutors. Sixty-four terms (lemma), including 22 inherent Hausa items and 42 instances of English loanwords or code-mixing, were tracked as terms used in reference to the actions (e.g. *chat(ting)*, *forward(ing)*), objects (e.g., *image*), and space (e.g. *group*, *online/offline*) associated with the chat environment. The results reveal members of the Hausa-speaking community to be quite innovative when it comes to drawing on their language's own lexical resources for use as chat terminology.

2 BACKGROUND

2.1 Increasingly Multilingual Cyberspace

English has long been recognized as the dominant, established lingua franca of the Internet (Danet and Herring 2007) as well as SMS communication. Nonetheless, through a combination of pure necessity—as smartphones and wireless technology spread to the remotest areas of the world—and users' sense of cultural identity, more and more languages have been adapted for computer-mediated communication (CMC), and by now the Internet and cybersphere can truly be recognized as a relatively diversified, multilingual environment.

Before looking at the example of Hausa WhatsApp chat in particular, let us first consider what it takes to truly adapt to the medium of cyberspace. To the extent that online chat and SMS messaging, presumably the most widely used applications of CMC, are similar to spoken conversation, one might think that adapting to the new technology is a simple matter of typing words as they are spoken. However, this naturally comes with various challenges, and the result is that English's influence in computer-mediated communication is partly reinforced by these obstacles.

First of all, of course, users must be literate and share some basic standards of orthography with their interlocutors. For languages lacking an established literate tradition, bilingual speakers may end up preferring to use English, thus reinforcing its continued dominance as the language of the Internet.

Furthermore, languages using non-Latin scripts face challenges. Although Internet and cell-phone technology can accommodate different language scripts, we still find users adapting their native language to Latin scripts. For example, "Greeklish" is a Latin script-based rendering of Greek that developed rapidly when the Internet came to Greek society (Androutsopoulos 2012). Similarly, Palfreyman and Khalil (2007) study the use of a so-called "ASCII-ized Arabic"—where Latin characters along with numerals and other symbols represent different Arabic letters—among college students in UAE. As such, even though the language of communication may not be English, the implicit hegemony of English as the language of the Internet is still reflected in the choice of script.

Third, in the online chat environment, at least, it is desirable to express oneself as rapidly as possible. This is largely facilitated by the development of abbreviated forms such as the iconic trends seen in the English-speaking world of CMC, with phrases like *y r u so l8* (in place of the 15-character phrase *Why are you so late?*). While any given language can be used for online chatting without such abbreviations, certain bilingual speakers might again opt for English as the language that gives them a ready-made, established medium for rapid, not to mention playful, communication.

2.2 CMC Terminology

Even where a language has successfully adapted to the CMC environment, there is yet another area where one might expect to see remnant signs of the dominance of English as the global language of technology—namely, in the use of specialized chat terminology. Though meant to mirror in many ways spoken conversation, chatters must on occasion refer to actions, objects, and space that are unique to the computer-mediated medium. In fact, presence in the chat environment often serves as a topic of conversation, as chatters make reference to *profile pictures* that they have *uploaded* to their *account* and request one another to *forward snapshots*, for example. Thus, inevitably, chat participants will have a need and desire for jargon for conscious reference to the virtual electronic environment itself—terms like *email*, *attachment*, *profile*, *upload*, and *online*. For example, one chatter switching to English in the Hausa chat database writes, "Where did u knw dem?@ur dp."

With such chat jargon logically taking cues from the field of information technology, and with online chat being a product of globalization in its own right, one might therefore expect, to begin with, bilingual chatters to resort to code-mixing in English (as the dominant language of globalization and IT). Furthermore, even monolingual chatters would be influenced by the multilingual community, and languages might fully adopt (borrow) English-based loanwords for such terms as *chat*, *forward*, and *online*.

Indeed, technical communication is often cited among the motivations for code-switching (i.e., bilingual speakers switching back and forth between different languages) and code-mixing (i.e., linguistic borrowing) (Daulton 2012, Wong 2006). In general, technological terms, such as those used in chat jargon, are prone to spread from the originating or dominant language to other cultures where they get adopted as loanwords. For example, when checking for translation equivalents for the word *computer* in Google Translate, 76% (77 of 101) of the languages supported present a word that is clearly derived from the Latin-cum-English term. Daulton (2012) further confirms that "the most borrowed words refer to technology (e.g. engine) and names for new artifacts (e.g. taxi)."

2.3 Alternatives to English Loanwords

The use of chat jargon might be inevitable, but the spread of terminology as loanwords is not. After all, the English language itself has drawn on various word-building strategies in the development of jargon dealing with computer technology—from reviving an old term like *cursor* (which itself had been borrowed from Latin, like so many English words), to repurposing common words like *mouse* and *web* via semantic extension, to use of acronyms like *PC*. Similarly, other languages can draw on their own resources.

In many cases, when languages are found using intrinsic strategies for technological lexical development, it is understood in part as a conscious effort to defend linguistic purity (Blommaert 2002, Haspelmath 2009). For example, the Académie française has long been active with moderating the development and documentation of new French terms, with moderate success thanks to government backing in matters of broadcasting and publication. Examples include recommending the use of *logiciel* and *courriel* in place of *software* and *e-mail* (Daulton 2012). Similar efforts at linguistic purification can be seen in other parts of the world, such as with Korean and various Eastern European languages (Haspelmath 2009).

2.4 Hausa

Hausa, an Afro-asiatic language spoken widely in West Africa, is an example of a language that has successfully been adapted for computer-mediated communication.[1] For one thing it does have an established, printed literary tradition using a Latin-based script. Although the Latin-based script was only introduced

[1] More details on the Hausa chat community are provided in later sections.

early in the 20th century, it has overtaken Ajami (an Arabic-based script, whose use with Hausa dates back to the 15th century) as the dominant orthographic standard. While many speakers might not be familiar with official standards of orthography, they get by well enough with predictable pronunciation and influence from mixed levels of literacy in English. Furthermore, within the corpus of Hausa chats described in this article, the Hausa speakers collectively use a variety of abbreviated forms such as *wlh* for *wallahi* ('by God') and *ya kk* for *yaya kake/kike/kuke* ('How are you?'—covering masculine, feminine, and plural forms of second-person reference which are otherwise distinguished in Hausa grammar).

But what about chat jargon in Hausa? Returning to the discussion in the preceding section, I will begin by noting that the Hausa community is not documented as one that is prone to efforts at language purification. First of all, the Hausa language has frequently drawn upon languages it comes into contact with to expand its lexicon. For example, words like *burodi* ('bread'), *tebur* ('table'), and *famfo* ('pump') have come from English, while terms like *albarka* ('blessing'), *hankali* ('wisdom'), and *wallahi* ('by God') come from Arabic. Some words traced to these two languages were transmitted to Hausa via yet other languages— such as *tasha* ('station') coming into Hausa from Yoruba (or possibly other languages spoken south of Hausa speaking areas), and *kasuwa* ('market'), having been introduced via another language of northern Nigeria, Kanuri, which had its own lexical borrowing from the Arabic word *suq* (Newman 2000). Secondly, and more directly relevant to this study, many of the Hausa speakers in the Hausa chat corpus frequently code-switch between Hausa and English (and less frequently, Arabic, Fulfulde, and Kanuri) in addition to using English borrowings (code-mixing) within Hausa texts. Though I earlier clarified the use of the terms code-mixing/lexical borrowing versus code-switching in parenthetical comments, the following example from a Hausa text serves to illustrate the difference (note: the examples reflect the original chat text, not standard Hausa orthography):

(1) Illustration of code-mixing versus code-switching in a Hausa chat text

Original chat:	MTN-na	nakasa	recharging	wlh
English gloss:[2]	MTN-my	1s.cont.-refuse	recharging	by.God

da	tuni	nakira	d	ntwrk	is	damn	bad	wlh
in.the.past	long.ago	1s.comp.-called	the	network	is	damned	bad	by.God

Translation: 'My MTN [SIM card] isn't recharging, I swear. I would have called long ago. The network is damned bad, I swear.'

[2] I try to avoid abbreviations in the English glosses of the linguistic examples presented in this article, to make them more self-explanatory. In example (1) 1s stands for first-person singular, cont. stands for continuative, and compl. stands for completive and in example (2) (presented later in Section 5) neg. stands for negative, 2s stands for second-person singular, f. stands for feminine, m. stands for masculine, and rel. stands for relative.

In the first line, the chat participant has code-mixed by inserting the English word *recharging* within his Hausa syntax, whereas at the end of the second line he completely code-switches to English.

As a language open to lexical borrowing, one might expect the largely bilingual chatters to naturally draw on established English terms for chat jargon. Indeed, many do draw on English both for emotive jargon (as seen in the 206 instances of *lol* and three instances of *l8r*, 'later'), which is not analysed in this study, and for the specialized terminology referring to the chat environment, which is examined in this paper. Yet, interestingly, within this relatively new medium, young Hausa speakers appear to have spontaneously adopted and spread homegrown terms, via semantic extension or metaphor, for the actions or processes (e.g. chatting, forwarding), objects (e.g. image) and space (e.g. group, online/offline) associated with phone- and Internet-based communication. Hausa thus shows itself to be a language with robust semantic extension, among other strategies for lexical expansion.

3 METHODOLOGY

3.1 Corpus Development

Data collection. The corpus was originally targeted as a database of SMS texts with the goal of collecting a minimum of 60 texts from at least 50 participants.[3] WhatsApp chats were ultimately adopted for the following reasons:

- it is more widely used for extended communication than SMS in Nigeria;
- the data is more practical to collect;
- it is a roughly comparable form of computer-mediated communication.

University students and some other community members shared excerpts of chats for which their interlocutors (friends, family members, colleagues) also agreed for the texts to be used in the database. To meet the originally targeted volume of data, chats were collected such that the contribution from each participant was at least 4,200 characters (based on an estimated average SMS length of 70 characters)—although for six additional participants included in the study the volume of texts fell short of this number. At the time of this study, the corpus included 56 participants (representing excerpts for 40 conversations between two individuals), and the total volume of the corpus was 21,693 lines (about 90,000 words or 380,000 characters).

A short survey of sociolinguistic/contextual information was conducted for each participant, the details of which are summarized in Table 1. All the participants

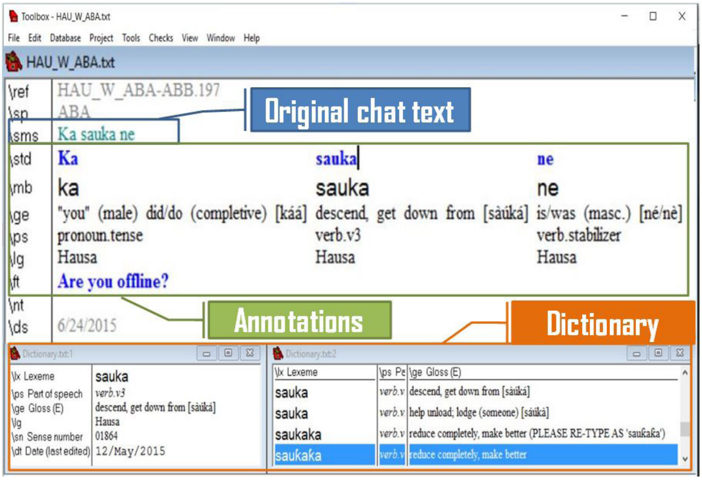
[3] This objective came from University of Maryland Center for Advanced Study of Language (CASL), who conceived of and funded the creation of this corpus.

claimed to speak English, with a handful of them also claiming fluency in other languages. As noted earlier, the participants were all bilingual, essentially fluent speakers of both Hausa and English (the Nigerian standard, which is largely based on the British standard).

Table 1: Chat Participant Demographics.

Factor	Details
Gender:	Female, 24; Male, 32
Age:	Average, 22; Mode, 20; Range of 14-35
Education:	Mostly undergraduate; but ranging from high school to Master's
Occupation:	Student, 48; Teacher, 2; Nurse, 1; Entrepreneur, 1; Musical artist, 2; Film maker, 1; Unemployed, 2
Origin (/Birthplace):	Adamawa, 10 (/0); Borno, 1 (/5); Gombe, 2 (/1); Jigawa, 2 (/1); Kaduna, 4 (/5); Kano, 20 (/19); Katsina, 7; Kogi, 0 (/1); Niger, 0 (/1); Sokoto, 1 (/0); Taraba, 2 (/1); Yobe, 6 (/5)
Residence:	Adamawa, 22; Borno, 2; Gombe, 1; Jigawa, 2; Kaduna, 6; Kano, 10; Katsina, 4; Yobe, 4; Sudan, 2
Mother Tongue:	Hausa, 27; Fulfulde, 16; Kanuri, 3; Yoruba, 1; Margi, 1; Nupe, 1; Other, 5
Language at Home:	Hausa, 45; Fulfulde, 9; English, 1; Yoruba, 1; Kanuri, 2;
Relationship to Interlocutor:	(Close/Best/Family) Friend, 29; Brother, 3; Sister, 3; Cousin, 3; Uncle, 1; Colleague, 3

Corpus processing. Each line of chat was annotated for standardized spelling, word translation, parts-of-speech, language (in case of code-switching) and a free translation of the entire comment. This was facilitated through the use of the Linguist's Toolbox (SIL), as illustrated in Figure 1.



\ref	HAU_W_ABA-ABB.197		
\sp	ABA		
\sms	Ka sauka ne		
\std	Ka	sauka	ne
\mb	ka	sauka	ne
\ge	"you" (male) did/do (completive) [káá]	descend, get down from [sàùká]	is/was (masc.) [né/nè]
\ps	pronoun.tense	verb.v3	verb.stabilizer
\lg	Hausa	Hausa	Hausa
\ft	Are you offline?		
\nt			
\ds	6/24/2015		

\lx Lexeme	sauka
\ps Part of speech	*verb.v3*
\ge Gloss (E)	descend, get down from [sàùká]
\lg	Hausa
\sn Sense number	01864
\dt Date (last edited)	12/May/2015

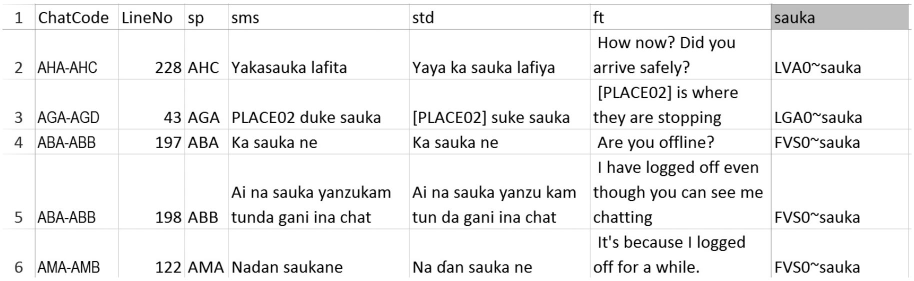
\lx Lexeme	\ps Pa	\ge Gloss (E)
sauka	*verb.v*	descend, get down from [sàùká]
sauka	*verb.v*	help unload; lodge (someone) [sáùkà]
saukaka	*verb.v*	reduce completely, make better (PLEASE RE-TYPE AS 'saukaka')
saukaka	*verb.v*	reduce completely, make better

Figure 1: Example of Data Annotation.

The methodology called for the primary participants from whose phone the data was collected to carry out the initial annotations on their own chat data with appropriate training. However, some participants were unable to complete this task and it was outsourced to other Hausa-speaking assistants. I subsequently vetted all annotations for accuracy and consistency, checking with the original chatter and/or other native Hausa speakers to resolve discrepancies.

3.2 Data Preparation

A convenient means to evaluate the context of each line of text was needed in order to analyse the use of chat terminology in the Hausa texts. Standard concordancing software (including the concordancing feature built into the SIL Toolbox software) was not appropriate, as I needed to view English translations along with the Hausa texts. So, for this step, I extracted essential contextual information (original and standardized Hausa and English free translation along with identifying information (file, line, speaker)) from the text files using Regular Expressions option in Funduc Search & Replace program,[4] and then exported these into an Excel spreadsheet (as seen in the first six columns in Figure 2).



1	ChatCode	LineNo	sp	sms	std	ft	sauka
2	AHA-AHC	228	AHC	Yakasauka lafita	Yaya ka sauka lafiya	How now? Did you arrive safely?	LVA0~sauka
3	AGA-AGD	43	AGA	PLACE02 duke sauka	[PLACE02] suke sauka	[PLACE02] is where they are stopping	LGA0~sauka
4	ABA-ABB	197	ABA	Ka sauka ne	Ka sauka ne	Are you offline?	FVS0~sauka
5	ABA-ABB	198	ABB	Ai na sauka yanzukam tunda gani ina chat	Ai na sauka yanzu kam tun da gani ina chat	I have logged off even though you can see me chatting	FVS0~sauka
6	AMA-AMB	122	AMA	Nadan saukane	Na ɗan sauka ne	It's because I logged off for a while.	FVS0~sauka

Figure 2: Excel Table Used to Verify Chat Jargon Usage.

Subsequently, all instances of targeted chat terminology (keywords dealing with the chat environment and presumed to be potential candidates for chat terminology used by this speech community) could be searched for in the "standardized spelling" field and evaluated in terms of contextual variables that were then coded as shown in the seventh column in Figure 2. Each occurrence of the targeted terms was tagged for the following contextual features: (1) Usage and language

[4] The following search and replace strings, respectively, were used to identify all data fields found in the text files and extract just the data needed for analysis: Search: \\ref*\r\n\\sp*\r\n\\sms*\r\n\\std*\r\n\\mb*\r\n\\ge*\r\n\\ps*\r\n\\lg*\r\n\\ft*\r\n\\nt*\r\n\\ds*\r\n; Replace: %1~%2~%3~%4~%9. As illustrated in Figure 1, the 'ref' and 'sp' fields contain the identifying information, while 'sms,' 'std,' and 'ft' contain the Hausa text and corresponding English translation.

choice (Hausa chat jargon versus other use of Hausa term, and English loanword versus English term used in full instance of code-switching; English words were likewise ascertained as being used as chat jargon or otherwise); (2) part-of-speech (Noun, Verb, Gerund/Verbal-noun, Adjective); (3) field of use (Action, Object, Space); (4) number of Hausa suffixes appearing on words; (5) whether or not the instance was a typo, correction, or immediate repetition of a previous instance; and (6) original spelling employed by the chat user.

In the sample shown in Figure 2, for example, the first two instances of the word *sauka* (a Hausa verb that literally means 'to descend or get down,' and which has been extended to refer to 'logging off or going offline') are coded as instances of a literal use of the word ("L" for literal Hausa usage). The next three examples, on the other hand, are instance of the figurative use that counts as chat terminology. Most of the examples in Figure 2 involve a word Hausa employs as a basic verb (V), but in one instance the gerund form (spelled exactly the same in this case) is used. The two instances with the literal reference to arriving/alighting from public transportation principally deal with an action (A)—irrelevant in any case, since these are not instances of chat terminology—whereas the three instances referring to 'going offline' are coded as relating to space (S) in the chat environment. None of the examples in Figure 2 have any morphological affixes (hence the 0); and none of the examples count as repetitions or corrections (in which case an additional code would have appeared after the 0).

Regarding the specific chat terms targeted for this study, I mainly relied on intuition when searching for concepts commonly used in everyday chat and relating to the immediate chat environment, and I also benefitted from knowledge of specific words being employed by chat users in this corpus (both Hausa and English), which I gained through the course of vetting the data annotations. The English translation field also served to identify potential Hausa chat jargon of this sort that I was not already aware of. For example, an instance of the Hausa word *taɓa* (literally, 'touch') had been glossed as 'text' by the Hausa-speaking annotator, drawing attention to an apparent specialized use of this word for the chat environment (discussed later in Section 4). There was thus no attempt to exhaustively search all possible terms that might qualify as specialized terminology used in reference to the CMC environment—as might be drawn from a resource like netlingo.com, for example, with over 6,000 entries (including abbreviations of general expressions like *lol* and *b4*, academic terms like *asynchronous learning* and *cyberterrorism*, and highly technical terms like *LAN* and *microsite*, as well as common terms like *upload* and *offline*).[5] The set of words ultimately included in the study (i.e., terms relating to common chat

[5] For example, two instances where a chat user incorporates English *hack* within Hausa utterances (as *hacko* and *hacking*) in reference to hacking into someone's camera (presumably from Internet connection) are not included. Here a chatter with IT training was referring to activities outside of the chat environment.

environment concepts for which at least one instance was found to occur in the texts) is presented in Table 2.

Table 2: List of Words Tracked (that appear in the corpus).

Theme Group	Jargon Terms[6]
Group A ('talk'):	*chat(ting)*, *'gist'* (Nigerian English term for casual/playful chat), *talk(ing)*, *[kuke] whatsapp*, *hira*, *magana*, *surutu*, *taɗi*, *zance*
Group B ('message'):	*answer*, *comment*, *link*, *mail*, *message*, *reply(ing)*, *respond(ing)/ response*, *text*, *ping*, *amsa*, *saƙo*, *taɓa(wa)*
Group C ('send'):	*email, forward(ing)*, *send(ing)*, *transfer(ing)*, *tura(wa)*, *turo(wa)*
Group D ('file operations'):	*attach(ing/ment)*, *copy(ing)*, *download(ing)*, *screenshot*, *snapping*, *delete*, *saving*, *goge*
Group E ('image'):	*image*, *(display/profile) picture (dp/pp, pic/pix)*, *photo*, *hoto*
Group F ('post'):	*post(ing)*, *upload(ing)*, *sa*, *saka(wa)*
Group G ('enter'):	*enter*, *launch*, *buɗe*, *shiga*
Group H ('online/ offline'):	*offline*, *online*, *[tana] on*, *fita*, *hau/hawa*, *sauka*
Group I ('Internet'):	*Internet*, *network*, *website*, *yanar gizo-gizo*
Group J ('group')	*account*, *group*, *username*, *password*, *code(s)*, *shafuffukan yaɗa zumunta*, *zaure*

As seen in the table, the terms have been categorized by field of use ('Theme group') to help track patterns of choice between Hausa terms and English code-mixing or code-switching. Some relevant and/or interesting cases may have been overlooked without a more systematic approach drawing upon a full dictionary of Internet terminology. For instance, the examples of *username* and *password* (presented later) were overlooked in the first round of analysis. However, the list used here is now a fairly exhaustive collection of the chat jargon I intended to target in this study.

4 RESULTS

4.1 Tally of Chat Jargon Terms

A total of 1,582 instances of the targeted terms were found to occur in the Hausa chat database. This initial tally included all instances, whether used as specialized chat terminology or polysemous terms used in other senses (as in an

[6] Glosses for Hausa terms are provided in the tables in Section 5.

English chatter referring to an actual spider web or a web of lies, as opposed to the World Wide Web.)

Of the 1582 instances of the target terms, 754 were identified as being used as intentional instances (i.e., not corrected typos leading to repetition) of chat jargon within Hausa texts. The remaining instances were excluded on one of the following grounds: (a) the term was not used as a chat term in the particular context (for example, as in the literal use of *sauka* in the sense of 'to descend or alight'—as opposed to going offline—as seen in the first two lines of Figure 2 presented earlier); (b) the term appeared in a full instance of code-switching—i.e., a text entirely or predominantly expressed in English or, more rarely, some other language; (c) the term appeared as a correction to a typing error (thus already counted in an immediately preceding instance).

Tables/Figures 3-12 present the results of these tallies for each of the 10 theme groups. Each group is presented and discussed in turn.

4.2 Group A: 'Talk'

Admittedly, the notion of *chat* or *talk* is a relatively problematic theme to track distinctly as a jargon term, since communication (and thus terms referring to verbal exchange) is a natural part of the chat environment. In any case, as seen in Table 3/Figure 3, for the instances identified as counting as chat jargon under this theme, the Hausa chatters in this corpus draw predominantly on Hausa vocabulary—using Hausa terms over twice as often as corresponding loanwords from English.

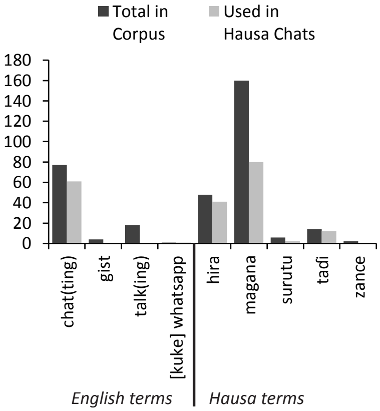
The frequency of using these Hausa terms might actually be a little higher than that shown here. I was relatively conservative in the inclusion of instances of the word *magana*, which carries the sense of 'matter, issue' in addition to 'talk, discussion' (the latter often in combination with the verb *yi* ('do')). I thus treated it as 'matter' where the interpretation was not clear, and excluded it from the chat jargon tally.

Though appearing less frequently than *magana* overall, the word *hira* appears to be the principle Hausa word used as a specialized term to refer to 'chat.' While *magana* is a frequently occurring word in Hausa in any context, *hira* has a more specialized original meaning: 'chat of an evening' (i.e. speakers making a special point to take time to chat casually), and reportedly it now refers to chatting in more general terms. In a similar vein, online forums for chatting present a space for very purposeful yet casual discussion between individuals, and thus the term *hira* must have been a natural choice for semantic extension to refer to this act.

An apparent relatively higher frequency of occurrence of *hira* in these chats compared to spoken communication (according to informal input from Hausa speakers)—as well as the higher frequency of instances used as jargon versus other uses in the corpus—underscores its use as a chat jargon term.

Table 3/Figure 3: Frequency of Occurrence for Words in Group A – 'Talk'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*chat(ting)* (77 total; 23.5%)	61 (31.0%)
	gist (4 total; 1.2%)	0 (0.0%)
	talk(ing) (15 total; 4.6%)	0 (0.0%)
	[kuke] whatsapp ('you guys are on WhatsApp') (1 total; 0.3%)	1 (0.5%)
		N=62 (31.5%)
Hausa terms	*hira* ('chat'; lit. 'informal chat of the evening, gist') (48 total; 14.7%)	41 (20.8%)
	magana ('talk, chat'; lit. 'talking, matter, issue') (160 total; 48.9%)	80 (40.6%)
	surutu ('chatting') (6 total; 1.8%)	2 (1.0%)
	taɗi ('chatting') (14 total; 4.3%)	12 (6.1%)
	zance ('talk, chat') (2 total; 0.6%)	0 (0.0%)
		N=135 (68.5%)



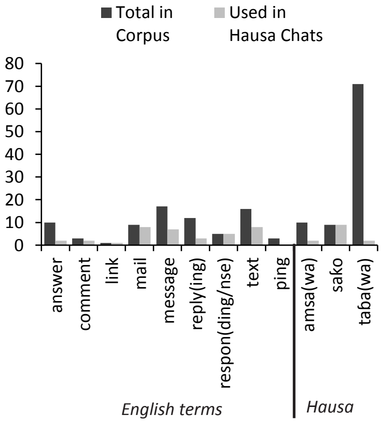
4.3 Group B: 'Message'

Group B includes a wider range of terms—various formats or methods of messaging by which chat users communicate with one another. In this case, it is the use of English code-mixing that is over twice as frequent, as seen in Table 4/Figure 4. I speculate this is due to the readily distinguishable nuances available with the well-established English terms.

Among the Hausa terms found in use, *amsa* ('respond'/'response') and *saƙo* ('message') are relatively general ones. Though it was hard to tell the exact intended sense of the instances of *taɓa* (verb form) and *taɓawa* (gerund/verbal noun), judging from the basic meaning of this term ('touch'), it seems likely that this is a budding extension of this term to refer to something like 'poking,' as used on social media platforms.

Table 4/Figure 4: Frequency of Occurrence for Words in Group B – 'Message'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*answer* (10 total; 6%)	2 (4.1%)
	comment (3 total; 1.8%)	2 (4.1%)
	link (1 total; 0.6%)	1 (2.0%)
	mail (9 total; 5.4%)	8 (16.3%)
	message (17 total; 10.2%)	7 (14.3%)
	reply(ing) (12 total; 7.2%)	3 (6.1%)
	respon(ding/nse) (5; 3%)	5 (10.2%)
	text (16 total; 9.6%)	8 (16.3%)
	ping (3 total; 1.8%)	0 (0.0%)
		N=36 (73.5%)
Hausa terms	*amsa(wa)* ('reply(ing)') (10 total; 6%)	2 (4.1%)
	saƙo ('message') (9 total; 5.4%)	9 (18.4%)
	taɓa(wa) ('poke'?; lit. 'touch') (71; 42.8%)	2 (4.0%)
		N=13 (26.5%)



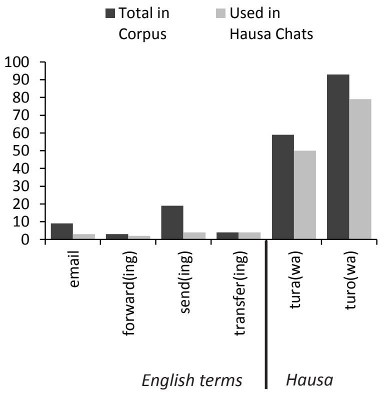
4.4 Group C: 'Send'

Compared to the various *formats* of message represented in Group B, the *means* of conveying them is more or less constant. Although English has various terms like *send*, *forward*, *email*, and *transfer*, these basically all boil down to sending. Incidentally, it is a Hausa word (*tura(wa)/turo(wa)*) that is overwhelmingly the term of choice when referring to the action of sending, as seen in Table 5/Figure 5.

The adoption of this term also illustrates a noteworthy case of semantic extension. The term *tura* literally means 'to push.' (The difference between *tura* and *turo* is that of directionality ('push away' vs. 'push towards,' respectively); and the *–wa* suffix creates a nominalized form of the verb or gerund, as pointed out earlier with *taɓawa*.) Outside of the chat environment, the term already carries an extended meaning of sending packages physically. So, again, it is a logical choice for conveying the notion of 'sending' messages, pictures, attachments, etc. by electronic means.

Table 5/Figure 5: Frequency of Occurrence for Words in Group C – 'Send'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*email* (9 total; 4.8%)	3 (2.1%)
	forward(ing) (3 total; 1.6%)	2 (1.4%)
	send(ing) (19 total; 10.2 %)	4 (2.8%)
	transfer(ing) (4 total; 2.1%)	4 (2.8%)
		N=13 (9.2%)
Hausa terms	*tura(wa)* ('send(ing)'; lit. 'push (outwards)') (59 total; 31.6%)	50 (35.2%)
	turo(wa) ('send(ing)'; lit. 'push (hither)') (93 total; 49.7%)	79 (55.6%)
		N=129 (90.8%)

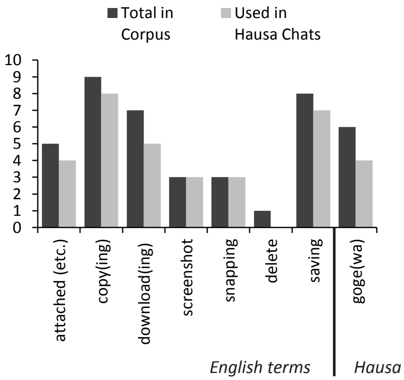


4.5 Group D: 'File-Operations'

Compared to 'sending,' which is a straightforward and common action regardless of what we call it, the chat environment involves numerous other specialized file operations. This is an area where we do find the Hausa speakers almost exclusively code-mixing in English, as shown in Table 6/Figure 6.

Table 6/Figure 6: Frequency of Occurrence for Words in Group D – 'File-operations'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*attached/attaching/ attachment* (5 total; 11.9%)	4 (11.8%)
	copy(ing) (and paste) (9 total; 21.4%)	8 (23.5%)
	download(ing) (7 total; 16.7%)	5 (14.7%)
	screenshot (3 total; 7.1%)	3 (8.8%)
	snapping (3 total; 7.1%)	3 (8.8%)
	delete (1 total; 2.4%)	0 (0.0%)
	saving (8 total; 19%)	7 (20.6%)
		N=30 (88.2%)
Hausa	*goge(wa)* ('delet(ing)'; lit. 'rub clean, polish') (6 total; 14.3%)	4 (11.8%)
		N=4 (11.8%)



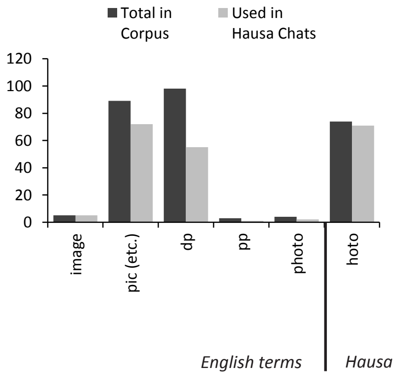
The only specialized file operation for which a Hausa term is found to be used is the notion of 'deleting' (a picture/file), which is expressed by the word *goge* (literally meaning 'to rub, wipe' and with an extended meaning of 'erase'). Next to the four instances of *goge*, the only instance of the English word *delete* occurs where a speaker has shifted to a full English utterance. All other distinctive file operations referenced in this corpus (attaching, copying, downloading, taking a screenshot, snapping (a picture), saving) draw on English terms.

4.6 Group E: 'Image'

The most prominent object discussed in the WhatsApp environment is the image—especially the so-called *dp* (display picture) on a user's profile, but also other images that are shared. In this case, abbreviated English forms *pic* (including related forms like *pix*) and *dp* are extremely common, accounting for 61.7% of references to images (Table 7/Figure 7).

Table 7/Figure 7: Frequency of Occurrence for Words in Group E – 'Image'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*image* (5 total; 1.8%)	5 (2.4%)
	pic & related forms (e.g. *pix*) (89 total; 32.6%)	72 (35.0%)
	dp (display pic) (98 total; 35.9%)	55 (26.7%)
	pp (profile pic) (3 total; 1.1%)	1 (0.5%)
	photo (4 total; 1.5%)	2 (1.0%)
		N=135 (65.5%)
Hausa	*hoto/foto* ('photo, picture') (74 total, including 7 spelled as *photo*; 27.1%)	71 (34.5%)
		N=71 (34.5%)



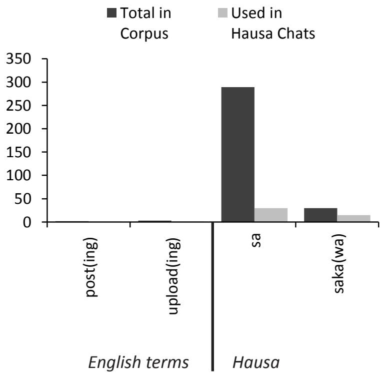
However, the Hausa term for picture (*hoto/foto*) appears about as often as the most common English term (*pic*). Obviously, the Hausa term is already an English borrowing, although here we are dealing with a loanword that entered the Hausa language at least more than 80 years ago (Bargery 1934) in reference to physical photographs, and it has since been fully adopted as a Hausa term carrying the same general scope as the English term *picture*. Included within the tally of Hausa *hoto* (alternative spelling *foto*) are a handful of instances that had been spelled as 'photo' but that otherwise pattern as the Hausa word based on clues like use of the Class II plural ending (as in *photuna*, compared to *hotuna* ('images')) and the definite marker *-n* (as in photon ('the image')). Although some speakers apply possessive pronoun suffixes when code-mixing in English, as seen in Example (1) presented earlier (*MTN-na* 'my MTN [SIM card]'), there is no evidence of other nominal suffixes such as those noted above (plural and definite markers) being attached to any English nouns appearing within the Hausa texts.

4.7 Group F: 'Post'

A specialized operation not included in Group D deals more specifically with images as opposed to other file types: posting. For this operation, which again is both common and straightforward (as there are not really any nuanced ways to post an image), a Hausa term is almost exclusively used: *sa(ka)*. This verb has the basic meaning of 'put, place.' The short form, *sa*, is also used in common expressions like *Me ya sa?* ('What happened?') and is a very frequently occurring word in general, with 289 total instances in this corpus (as shown in Table 8/Figure 8), of which 30 refer to posting in the chat environment. Technically, *sa* is just a reduced form of *saka*, but in practice the full form is used more rarely, and (according to informal input from Hausa speakers) it tends to be used in reference to a very deliberate act like placing a poster or sign on a wall or bulletin board. Given that *saka* is also heard more rarely in speech (based on impressions of Hausa speakers consulted on the difference between *sa* and *saka*), it seems the 1:2 frequency in this corpus relative to the more common short form *sa* is noteworthy—potentially indicative of its status as a specialized chat term.

Table 8/Figure 8: Frequency of Occurrence for Words in Group F – 'Post'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*post(ing)* (2 total; 0.6%)	1 (2.1%)
	upload(ing) (3 total; 0.9%)	1 (2.1%)
		N=2 (4.2%)
Hausa terms	*sa* ('post'; lit. 'put, place') (289 total; 89.2%)	30 (63.8%)
	saka(wa) ('post(ing)'; lit. 'put, place') (30 total; 9.3%)	15 (31.9%)
		N=45 (95.8%)

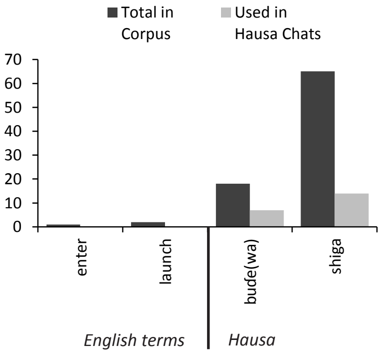


4.8 Group G: 'Enter'

Another type of action that is referenced in the chat environment has to do with navigating the space, as in clicking on a link. Somewhat surprisingly, the English term *click* (a likely candidate as a jargon loanword in the IT environment) is not found to be used at all—only appearing in shared links (with text copied from some other source). As shown in Table 9/Figure 9, the only other English terms found anywhere are two instances of *launch* and one of *enter*, used only when fully switching to English. All references to navigating the WhatsApp space (as in guiding an interlocutor through account settings) are carried out with two Hausa terms: 14 instances of *shiga* ('enter') and seven of *buɗe* ('open').

Table 9/Figure 9: Frequency of Occurrence for Words in Group G: 'Enter'.

	Total uses of target word in corpus	Used as jargon in Hausa
English	*enter* (1 total; 1.2%)	0 (0.0%)
	launch (2 total; 2.3%)	0 (0.0%)
		N=0 (0%)
Hausa terms	*buɗe(wa)* ('open') (18 total; 20.9%)	7 (33.3%)
	shiga ('enter') (65 total; 75.6%)	14 (66.7%)
		N=21 (100%)



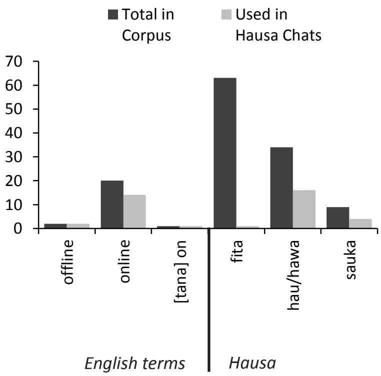
4.9 Group H: 'Online/Offline'

Another concept that comes immediately to mind as a likely candidate for borrowing from English chat jargon is the notion of being online or offline. In this case, as seen in Table 10/Figure 10, the English term *online* is indeed frequently used, along with a few instances of *offline*. However, these terms see strong competition from Hausa equivalents, with the Hausa terms being favoured overall (55.3% versus 44.7%).

The word for offline (*sauka*) and its original meaning of 'to descend' was introduced earlier, with the examples of data processing in Section 3. Similarly, the concept of being online draws on the Hausa antonym for *sauka*: *hau* ('to mount, climb'). These two terms are clearly on their way to being spread as the principle Hausa chat jargon terms for online/offline. However, in one instance the verb *fita* ('to exit/go out') was used in reference to going offline.

Table 10/Figure 10: Frequency of Occurrence for Words in Group H: 'Online/ offline'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*offline* (2 total; 1.6%)	2 (5.3%)
	online (20 total; 15.5%)	14 (36.8%)
	[tana] on (i.e.'[she is] on[line]') (1 total; 0.8%)	1 (2.6%)
		N=17 (44.7)
Hausa terms	*fita* ('enter') (63 total; 48.8%)	1 (2.6%)
	hau/hawa ('go(ing) online'; lit. 'mount') (34 total; 26.4%)	16 (42.1%)
	sauka ('go offline'; lit. 'descend') (9 total; 7%)	4 (10.5%)
		N=21 (55.3%)



4.10 Groups I & J: 'Internet' & 'Group'

The two remaining theme groups involve direct reference to virtual spaces: from one's personal account, to exclusive online groups, to the broader Internet itself. Frequency data for relevant jargon terms found in this corpus are presented in Table 11/Figure 11 (Group I – 'Internet') and Table 12/Figure 12 (Group J – 'Group'). Virtual accounts also have objects of sorts associated with them (username and password), and instances where these were referred to in the Hausa texts are also incorporated into Table 12/Figure 12.

Table 11/Figure 11: Frequency of Occurrence for Words in Group I – 'Internet'.

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*internet* (1 total; 4.5%)	1 (20.0%)
	network (18 total; 81.9%)	2 (40.0%)
	website (2 total; 9.1%)	1 (20.0%)
		N=4 (80%)
Hausa	*yanar gizo-gizo* ('Internet') (1 total; 4.5%)	1 (20.0%)
		N=1 (20%)

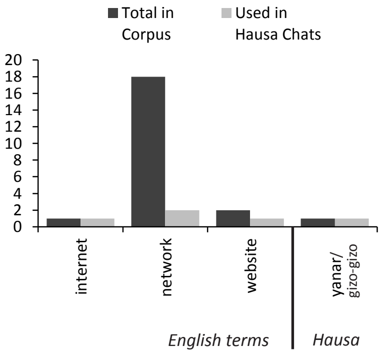
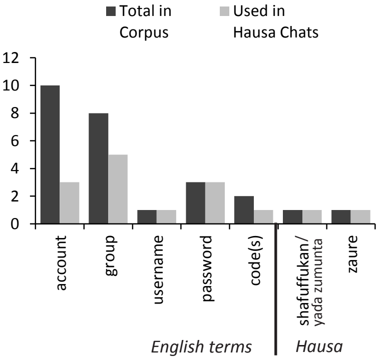


Table 12/Figure 12: Frequency of Occurrence for Words in Group J – 'Group'

	Total uses of target word in corpus	Used as jargon in Hausa
English terms	*account* (10 total; 38.5%)	3 (20%)
	group (8 total; 30.8%)	5 (33.3%)
	username (1 total; 3.8%)	1 (6.7%)
	password (3 total; 11.5%)	3 (20%)
	code(s) (2 total; 7.7%)	1 (6.7%)
		N=13 (86.7%)
Hausa terms	*shafuffukan yada zumunta* ('social network') (1 total; 3.8%)	1 (6.7%)
	zaure ('group'; lit. 'entry hall to a compound') (1 total; 3.8%)	1 (6.7%)
		N=2 (13.3%)



Two similar observations can be made for the two theme groups represented here. First, in both instances, English terms are more frequently drawn upon, but Hausa equivalents also appear with reference to the space-associated terms. Secondly, the number of occurrences of any term is quite low, thus reducing the significance of the relative frequency between English versus Hausa terms. The fact that the Hausa alternatives exist means that they could conceivably be or become more widespread, especially if there is a trend to continue to draw on indigenous terms to fill the role of chat jargon.

The Hausa terms adopted in these cases are especially creative. The word for group (*zaure*) comes from the word for entry hall in the traditional Hausa housing compound, where guests wait to be received by the host. This ends up being a fitting extension of this particular word, if not as obvious a choice as jargon terms like *hira* ('chat') and *sa(ka)* ('post'). Its simple, one-word format also makes it a good candidate to catch on as a chat term. The other creative Hausa terms in these groups are built from compounding. The phrase *shafuffukan yada zumunta* was used in place of the term 'social media.' The breakdown in meaning is as follows: *Shafuffukan* is the plural form of the word *shafi* (along with the linking suffix *–n*). *Shafi* has a variety of senses having to do with a 'sheet' of something (the lining of a garment, page of a book, coat of paint); *yada* is a verb meaning 'to spread (news, info, rumours)'; and *zumunta* means 'close relations, intimacy.' So, the literal translation is 'sheets (media) for spreading good relationships.' Surely, a phrase of this length is not likely to catch on without an abbreviated form, which is somewhat hard to imagine from this rather complex phrase. Similarly, the term for the Internet is a relatively lengthy compound: *yanar gizo* ('spider web')—actually appearing as *yanar gizo-gizo* in this corpus. In this case, however, it is conceivable that this term could be reduced to *yana*, for example, even though in its original sense *yana* on its own refers to a film or scum covering a surface and does not convey the sense of 'web' without being combined with the word *gizo* ('spider'). For the younger generation, the sense of 'web' comes more readily.

5 DISCUSSION

5.1 Analysis of results

From the results presented above, we see that Hausa-speaking chat users are employing a mixture of English code-mixing and Hausa words as chat jargon. That bilingual speakers (or non-English speakers in a multilingual speech community) end up using English loanwords from the IT field is not surprising. It is, however, somewhat striking to see the degree to which Hausa terms have quickly been adapted for use as chat jargon in a relatively new medium, and one that otherwise tends to be dominated by English at a global level.

When organizing the results by theme groups, we see that the likelihood of finding an English term versus a Hausa alternative is not entirely random. First, a number of Hausa terms emerge as natural candidates to fulfil the role of key chat jargon where the referenced meaning is clear, either having a literal sense or applying only a light metaphorical extension: *hira* ('chat'), *tura* ('send'), *hoto* ('image'), *sa* or *saka* ('place' = 'post'), and a combination of *shiga* ('enter') and *buɗe* ('open') for clicking on links. In the case of *tura*, *sa* and *shiga/ buɗe* (and variant forms), the Hausa terms are used almost exclusively.

With a number of other terms, a wider leap of semantic extension is called upon to repurpose Hausa words to expand the Hausa-based chat jargon. For example, the notion of going or being online and offline is aptly equated to climbing on and descending, employing the Hausa verbs *hau* and *sauka* (and variant forms), respectively. Though extremely rare in this corpus (and thus not substantial enough to draw meaningful conclusions about the relative frequency of use), we also find innovative semantic extension with terms for online 'group' and Internet, as well as an innovative compound term to refer to social media: *zaure* ('entry hall' = 'group'), *yanar gizo(-gizo)* ('spider web' = 'Internet'), and *shafuffukan yaɗa zumunta* (= 'social media').

Where English still dominates to a great extent are areas where the widely established English IT terms account for important distinctions or nuances in specialized actions and objects—including various file operations (like *attaching*, *copying*, *downloading*, *deleting*, and *saving*) and message types (like *comment*, *response*, *link*, and *text*) as well as terms like *username* and *password*. Nonetheless, we do find speakers drawing on Hausa resources for purposes of this sort—such as *buɗe* ('open'), mentioned above as a logical choice for clicking a link or opening a file, and *goge* (literally 'rub, wipe') being used in reference to the deletion of a virtual object. It may just be a matter of time before the innovative Hausa-speaking community repurposes other Hausa words for more specialized IT concepts.

Short of drawing on purely indigenous Hausa lexical items to fulfil the role of chat jargon, another possibility is for English code-mixing to lead to fully incorporated lexical adoption. Recall an example of this was pointed out in the case of *hoto*, a loanword from English dating back to the colonial period which almost all Hausa speakers would now consider as a Hausa word. The status of the word *hoto* within the Hausa lexicon is reflected by adjustments in phonological form and morphological behaviour. A hint at such a development among chat jargon today appears among the instances of the specialized 'file-operations' terms. Consider the following example:

(2) Illustration of English loanword adapting to Hausa phonology?

Original chat (Speaker A):	Shine	kika	copa	maganata	ko
English gloss:[7]	it-be	2s.f.rel.compl.	copy	talk-my	or?

Translation: 'So, you have copied my words, eh?'

Original chat (Speaker B):	Ai	ba	kai na	copa	ba
English gloss:	oh!	neg.	2s.m. 1s.rel.compl.	copy	neg.

Translation: 'Well, it's not *you* I copied'

In this example, one speaker introduces a word spelled as *copa* when accusing the interlocutor of copying his words. Rather than use the English spelling *copy*, or even mapping English pronunciation onto Hausa orthography (e.g. <kopi>), the vowel at the end has changed. Hausa has a complex set of verb classes or 'grades,' but the three most common basic grades start with the form CVCa—that is a sequence of consonant, vowel, consonant, and *–a* as the final vowel (along with distinctive patterns with vowel length and tone which are not reflected in standard orthography). Though the spelling is flawed—< *c* > in Hausa orthography corresponds to a "ch" sound—we see here an apparent attempt to adapt the English loanword to Hausa morphophonology, whether intentionally or subconsciously. Incidentally, the addressee uses the same form in his response. This exchange either suggests the *Hausafied* form is already spreading, or it captures a moment where one speaker succeeds in influencing the lexical choice of another. In either case, the implications are interesting, and it would be informative to track further development of this form by these or other speakers. For example, a tendency towards incorporation of this loanword into Hausa lexicon could be confirmed if a nominalization like *<copawa> ends up appearing instead of the English gerund *copying*, or if the use of a form like *<kopa> in spoken communication reflects the tonal and vowel length patterns of a particular verb grade.

[7] compl. stands for completive.

5.2 Considerations for Extended Research

Sociolinguistic Factors. When it comes to analysing lexical choices by bilingual speakers, we should also account for sociolinguistic factors. Previous studies have reported mixed results regarding the relationship between certain sociolinguistic characteristics and code-mixing or code-switching. With regard to sex, for example, Rabbani and Hammad (2012) find no difference in patterns of code-mixing by Urdu-English bilingual undergraduates, while Das and Gambäck (2013), drawing on populations of Bengali-English and Hindi-English university students, find that females code-switch more while males code-mix more. However, a greater variety of studies have found women to code-mix more, including Ahmed, Ali, and Xiang's (2015) study of SMS texting by Urdu-English speakers, Hamdani's (2012) study of language use among Sundanese-Bahasa teens, and Wong's (2006) broad-based research examining code-mixing by Chinese-English speakers. However, there is less research on the effect of other sociolinguistic factors on code-mixing or code-switching. Nonetheless, Wong (2006), for example, finds a strong correlation between education and code-mixing but no noteworthy correlation with age.

The relatively homogenous nature of this corpus of Hausa chats (mostly composed of texts from college students around 20 years old), precludes the ability to analyse the effects of variables like age, education, and occupation. Likewise, although factors such as region of origin and mother tongue were tracked and some variation is reflected in the corpus, the corpus size and spread of data are not conducive for analysing any impact they may have on language choice. On the other hand, with the data largely controlled for the above-mentioned factors, we can more confidently analyse the effect of gender. In terms of gender, the corpus is relatively balanced (24 females and 32 males, as shown earlier in Table 1, with 70% of the chat jargon terms coming from females and 30% coming from males).

Table 13 presents the frequency by which instances of chat jargon terms (a) appear as Hausa-based lexical items, (b) involve English code-mixing, or (c) occur within English code-switching. In addition to the chat terms analysed in Section 4, presented above, this sociolinguistic analysis also includes 80 instances of references to specific social media apps (BBM, Facebook, Instagram, Skype, Snapchat, Viber, YouTube, and WhatsApp). From this distribution, we see that females seem to prefer a combination of code-mixing (41.5%) and code-switching (19.6%) to Hausa-based jargon (38.9%), compared to their male counterparts: 46.5% Hausa terms versus 36.2% English code-mixing and 17.2% code-switching (Chi-square = 4.284; *p*-value = .038473., significant at $p < .05$). Incidentally, this tends to support those studies that found female speakers to code-mix and code-switch more than men (Ahmed Ali and Xiang 2015; Hamdani 2012; Wong

2006). In any case, however, it is of interest for future works to pursue a fuller, more systematic account of the relation between different sociolinguistic factors and the use of chat jargon.

Table 13: Cross-tabulation of Gender and Lexical Choice for Instances of Chat Jargon.

Group	Hausa	Code-mix	Code-switch	Total	%
Male	325 (46.6%)	253 (36.2%)	120 (17.2%)	698	69.5%
Female	119 (38.9%)	127 (41.5%)	60 (19.6%)	306	30.5%
Total	444 (44.2%)	380 (37.8%)	180 (17.9%)	1004	

Notes. Chi-square = 4.284; *p*-value = .038473. Significant at *p* < .05 (but not at p<.01)

Degree of Specialization of Jargon Terms. Another important question that remains to be addressed more systematically is the relation between the chat jargon terms and the use of the same words in various other contexts. For example, while still focusing on chat space, how do the dynamics of a chat group (instead of just one-on-one exchanges) affect word choices and the promotion of particular jargon terms? To what extent are the various IT jargon terms found elsewhere on the Internet? Can we get a more accurate estimate of the relative frequency of the target terms in spoken communication versus online communication? In the earlier presentation of results, I relied on impressions from native speakers for rough judgments. However, future extensions of this research should aim for a more systematic data-driven approach to such issues.

Origin and Spread of Hausa-based Jargon. Finally, this article necessarily attributes the spread of Hausa chat jargon to the Hausa-speaking chat participants. But where has this community drawn its inspiration? For example, the term *yanar gizo* had been documented as referring to the Internet as early as 2007 (Newman 2007). More recently, this phrase has even been used as the title of a "Kannywood" film which focuses on the use of social media: "Yanar Gizo" (A.Y.A Media, Nigeria 2014). (The hub of the Hausa film industry is the city of Kano—hence the industry nickname of "Kannywood".) By nature of most Kannywood films, the word also features in song and multiple film instalments—all of which are likely to reinforce or spread its use among Hausa speakers. Other chat conventions might be traced to popular Hausa literature. For example, several speakers use the sequence *mtsw* as an ideophone for a lip-pursing/inward sucking sound used to express disapproval, and one of the users claimed this spelling convention can be traced to Hausa romance novels. While it is quite conceivable that many innovations have and will continue to come directly from within the chat community itself, inspiration by and reinforcement in other media will surely help spread the fuller development of a Hausa-based chat jargon that already appears to be robust, based on the patterns found in the corpus presented in this study.

6 CONCLUSION

In this article, applying data from a newly compiled corpus of WhatsApp chats in Hausa, I have analysed the language choices of Hausa-speaking chat users when drawing on terminology used to refer to the chat environment. While the bilingual speakers represented in this corpus do code-mix with common English terms like *chat*, *text*, *pic*, *download*, *online*, and *username*, as might be expected, they also widely employ Hausa words adapted for specialized reference to cyberspace, such as *hira* ('chat'), *saƙo* ('message'), *hoto* ('image'), *tura* ('forward, send'), and *hau* ('go online'). English terms were predominant where nuanced meaning is more important—as in types of messages (e.g. *comment*, *link*, *reply*) and distinct file operations (e.g. *attach*, *copy*, *save*). On the other hand, in some cases where reference is made to common, general actions, like sending and posting, the Hausa terms—*tura* ('send') and *sa* ('post') were predominant. However, with some other general concepts the ratio of occurrence was relatively balanced—as in reference to images (English *pic* versus Hausa *hoto*) or being connected to the Internet (English online/offline versus Hausa *hau* ('go online'; lit. 'mount') and *sauka* ('go offline'; lit. 'descend, dismount'). Preliminary sociolinguistic analysis reveals that the female chat users tended to code-mix and code-switch to English more than the males, reinforcing similar findings in other speech communities. In a field of study dominated by the major world languages, it would be of interest to track the evolution of underrepresented languages, like Hausa, along with other African languages that are adapting to cyberspace. The present study is a step in this direction, and hopefully presages the wider cross-linguistic study of computer-mediated communication in future works.

References

Ahmed, Khalid, Ihsan Ali and Hua Xiang, 2015: Code-mixing as a marker of gender identity in SMS language in Pakistan. *Journal of Humanities and Social Science* 20/1. 58–65.

Androutsopoulos, Jannis, 2012: 'Greeklish': Transliteration practice and discourse in the context of computer-mediated digraphia. Jaffe, Alexandra, Jannis Androutsopoulos, Marka Sebba and Sally Johnson (eds.): *Orthography as social action: Scripts, spelling, identity and power*. Berlin: De Gruyter. 359–392.

Bargery, George P., 1934: *A Hausa-English dictionary and English-Hausa vocabulary*. London: Oxford University Press.

Blommaert, Jan, 2002 [1994]: The metaphors of development and modernization in Tanzanian language policy and research. Fardon, Richard and Graham Furniss (ed.): *African languages, development and the state*. London: Routledge. 213–226.

Danet, Brenda and Susan Herring, 2007: Introduction: Welcome to the multilingual Internet. Danet, Brenda and Susan Herring (eds.): The multilingual Internet: Language, culture, and communication online. Oxford and New York: Oxford University Press. 3–39.

Das, Anupam and Björn Gambäck, 2013: Code-mixing in social media text: The last language identification frontier? *TAL* 54/3. 41–64.

Daulton, Frank E., 2012: Lexical borrowing. Chappelle, Carol A. (ed.): *The Encyclopedia of Applied Linguistics*. Blackwell Publishing. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0687/abstract. (Last accessed 29 June 2017.)

Hamdani, Fakry, 2012: The influence of gender in determining the language choice of teenagers: Sundanese versus Bahasa. *International Journal of Basic and Applied Science* 1/1. 40–43.

Haspelmath, Martin, 2009: Lexical borrowing: Concepts and issues. Haspelmath, Martin and Uri Tadmor (eds.): *Loanwords in the world's languages: A comparative handbook*. Berlin: De Gruyter. 35–54.

Newman, Paul, 2000: Comparative linguistics. Heine, Bernd and Derek Nurse (eds.): *African languages: An introduction*. Cambridge: Cambridge University Press. 259–271.

Newman, Paul, 2000: *A Hausa-English dictionary*. New Haven: Yale University Press.

Newman, Paul, 2007: *The Hausa language: An encyclopedic reference grammar*. New Haven: Yale University Press.

Palfreyman, David and Muhamed al Khalil, 2007: A funky language for teenzz to use: Representing Gulf Arabic in instant messaging. Danet, Brenda and Susan Herring (eds.): *The multilingual Internet: Language, culture, and communication online*. Oxford and New York: Oxford University Press. 43–63.

Rabbani, Rida and Mushtaq Hammad, 2012: Difference in code-switching and code-mixing in text messages of undergraduate students. *Language in India* 12/1. 346–356.

Wong, Kwok-Lan Jamie, 2006: Gender and codemixing in Hong Kong. Honours Thesis, University of Sydney Linguistics Department.

Software Used

Field Linguist's Toolbox (SIL International): http://www-01.sil.org/computing/toolbox/. (Last accessed 29 June 2017.)

Goldvarb X: http://individual.utoronto.ca/tagliamonte/goldvarb.html. (Last accessed 29 June 2017.)

Search & Replace Pro (Funduc, Inc.): http://www.funduc.com/search_replace.htm. (Last accessed 29 June 2017.)

Part 2
Sociolinguistic analysis of CMC

WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication

Lieke Verheijen, *Radboud University*

Abstract

Communication via new media or social media, i.e. computer-mediated communication (CMC), is now omnipresent. The 'CMC language' that youngsters use in such media often diverges from the 'official' spelling and grammar rules of the standard language. Many parents and teachers are thus critical of CMC language, because they view Standard Dutch as a strict norm. Yet among youths it enjoys a certain status, and is regarded as playful, informal, and cool. So an interesting power conflict exists between the overt prestige of the standard language and the covert prestige of CMC language among youngsters. To determine how Dutch youths' language use in computer-mediated messages differs from Standard Dutch, an extensive register analysis was conducted of about 400,000 tokens of digital texts, produced by youths of two age groups – adolescents (12-17 years old) and young adults (18-23 years old), in four social media – SMS text messages; instant messages, viz. MSN chats and WhatsApp messages; and microblogs, namely tweets. This corpus study focuses on various linguistic features of four writing dimensions: orthography (textisms, misspellings, typos), typography (emoticons, symbols), syntax (omissions), and lexis (borrowings, interjections). The results suggest that the variables of age and medium are of crucial importance for (Dutch) youths' online language use.

Keywords: social media, computer-mediated communication (CMC), youth language, writing, WhatsApp

1 INTRODUCTION[1]

The use of social media has increased massively in recent years, both worldwide and in the Netherlands. Communication via these new media is called 'computer-mediated communication,' abbreviated to CMC. This has been defined as "the practice of using networked computers and alphabetic text to transmit messages between people or groups of people across space and time" (Jacobs 2008: 470). A growing number of communication tools are now at our disposal on computers, mobile phones, and tablets, and their users appear to get younger by the day. In informal CMC, young people often use what can be called 'CMC language' (in Dutch: '*digi-taal*'). The definition of this, as used in this paper, is as follows:

CMC language is a digitally written language variant that is especially used by youths in informal communication via new media, and is characterized, to a greater or lesser extent, by deviations from the standard language norms at different levels of writing, such as spelling, grammar, and punctuation.

In fact, CMC language is an umbrella term which encompasses great variation in itself, depending on various characteristics such as the user who composed the text, the circumstances under which it was written, and the medium that was used to produce it (see section 1.2). So even though language use in CMC has several prominent linguistic peculiarities, computer-mediated texts do not always display the same features to the same extent. Yet because CMC language overall diverges markedly from the standard language, this has caused feelings of resistance among some people, particularly from older generations, as it is feared that these new media pave the way to 'language corruption' or 'language deterioration'. Such sentiments are based, however, on superficial observations, anecdotal evidence, and personal experiences with CMC – not on empirical research. To find out whether these fears are in any way justified, a large-scale systematic register analysis was conducted of digital texts composed in four new media, namely SMS text messaging, instant messaging via MSN Messenger, microblogging on Twitter, and instant messaging via WhatsApp Messenger, written by Dutch youths from two age groups, i.e. adolescents and young adults.

The research question that is central to this paper is as follows: how does the language used by Dutch youths in these social media differ from Standard Dutch? In addition, the following question is addressed: is this language dependent on age group and/or medium? In other words, is the linguistic variation within written CMC by youths from the Netherlands dependent on social and medium-related factors?

[1] This chapter is a translated, extended, revised, and updated version of a Dutch conference paper by the author (Verheijen 2016).

1.1 New media

Research into new media requires clarity about what this term encompasses. In this day and age, numerous new media exist. Two relatively 'old' new media are text messaging and email, which first became popular two decades ago. Online chats are of a similar vintage, and two main kinds exist: chat rooms hosted on the Internet and instant messaging services, with the latter occurring via four kinds of technologies: mobile phone applications (e.g. *WhatsApp Messenger*, *Telegram*), Internet applications (*Google Hangouts*, *Skype*, formerly *MSN Messenger*), social networking sites (*Facebook chat*), and online gaming networks or virtual worlds (*World of Warcraft*, *Second Life*). Other new media include social networking sites (*Facebook*, *Google+*) and platforms for sharing visual media (*YouTube*, *Instagram*, *Pinterest*). Blogs and microblogs (*Twitter*, *Tumblr*) are also forms of new media. The concept further includes online forums or discussion boards (*4chan*, *FOK!forum*, *VIVA Forum*). This list indicates that new media are extremely varied, and thus the communication that takes place via these various platforms can also be surmised to be rather diverse. That is, each of these media differ in multiple characteristics that may affect the language used in CMC. Table 1 gives an overview of the various media analysed in this paper.

Table 1: Characteristics of four new media.

Medium characteristics	Instant messaging: MSN	Text messaging: SMS	Microblog: Twitter	Instant messaging: WhatsApp
message size limit	no	yes (max 160 characters)[2]	yes (max 140 characters)	no
synchronicity of communication	near-synchronous (real-time)	asynchronous (deferred time)	asynchronous (deferred time)	near-synchronous (real-time)
visibility	private	private	public, sometimes private (direct message)	private
interactivity	one-to-one or some-to-some (group chat)	one-to-one, sometimes one-to-many (broadcast message)	one-to-many, sometimes one-to-one (direct message)	one-to-one or some-to-some (group chat)
technology	computer	mobile phone (or computer)	mobile phone or computer	mobile phone (or computer)
communication channel	multimodal	textual or multimodal[3]	multimodal	multimodal

[2] With the exception of concatenated text messages, in which messages are joined if the limit is exceeded.

[3] The use of emoticons (see section 2.2) in SMS is textual, because they are composed of typographic characters. Smartphones, however, allow the use of emoji in SMS (but not in the present corpus): this leads to multimodality, because emoji are small images.

1.2 Computer-mediated communication

Certain attributes of CMC language, on various levels of writing, have cross-linguistically emerged from previous research. As for orthography, CMC language is prototypically known for the use of unconventional, non-standard spelling, 'textisms'; that is, transformations of conventionally spelt words.[4] As for typography, emoticons are a key novel feature of such communication (e.g. Silva 2011). Moreover, a frequently mentioned syntactic attribute is the omission of words, in particular function words (Ferrara et al. 1991, Werry 1996, Hård af Segerstad 2002, Crystal 2006, Frehner 2008, Bergs 2009, Winzker, Southwood and Huddlestone 2009, Herring 2012, Wood, Kemp and Plester 2013). A lexical attribute is the use of many English borrowings (Crystal 2008, Frehner 2008, De Decker and Vandekerckhove 2012). Graphical attributes are, for example, the use of hyperlinks and the incorporation of images, sound files, or videos; there can also be multimodality, a "blending of graphic with grapheme" (Carrington 2004: 218).[5]

CMC language thus tends to deviate from the standard language, a phenomenon that has roots in four main causes. Firstly, efficiency and speed are of great importance when communicating via new media, and tempo thus overrules 'correctness.' In addition, some media are limited in message size. For example, a single text message can only contain up to 160 characters, and a tweet no more than 140, so succinctness is crucial in these media. Secondly, words are often typed in computer-mediated messages as they are pronounced in informal spoken language (phonetic writing), to make the writing more like casual speech. Deviations from the standard language can, furthermore, increase expressivity: they can compensate for the lack of paralinguistic and prosodic elements in written (digital) language, such as stress, intonation, and volume, as well as the lack of body language, such as gestures and facial expressions. Androutsopoulos (2011: 149) summarizes these three principles as economy, orality, and compensation. Lastly, many youths like to be creative and original when communicating via new media, and such playing with language can contribute to their social identities. We can infer from this that many deviations in CMC language are functional: they are often resourceful, practical adaptations for which youths, in the context of the current study, make optimal use of the linguistic possibilities of written CMC in order to reach their communicative goals, despite the technological limitations of new media and the pragmatic limitations of written language.

[4] The term 'textism' is obviously derived from the phrase 'text messaging,' but these unconventional spellings also occur in CMC via other media.

[5] Bergs (2009) rightly stresses that not all of these deviations from the standard language were first invented during communication via new media. Some features of CMC language were already present in earlier writing genres, such as telegrams, postcards, informal personal letters, and newspaper headlines.

Still, Crystal (2006: 128) is right when he remarks that “the graphological deviance noted in [new media] messages is ... not universal”: digital texts diverge from the standard language to different extents. Such differences stem from a variety of factors (Herring 2001, Hård af Segerstad 2002, Crystal 2006, Crystal 2008, Drouin and Davis 2009, Proudfoot 2011):[6]

- user characteristics, such as age, gender, region, ethnic background, familiarity with textisms, personal preferences;
- situational characteristics, such as conversational topic, (social distance to) receiver of the message, communicative intent;
- medium characteristics, such as a possible message size limit, (a)synchronicity, interactivity, visibility.

All this makes CMC language stylistically diverse. That is why, as Hård af Segerstad (2002: 234) rightly argues, CMC should not be regarded as “one single mode of communication.” Rather, each new media user determines their own unique way of communicating every time they compose a digital message, depending on their personal profile, the medium they use for communication, and various situational features.

1.3 Polarization and prestige

CMC language has evoked a range of sentiments. A so-called ‘Gr8 Db8’ (great debate) exists about CMC language and its impact on reading, writing, and spelling (Crystal 2008), and it has become quite polarized. On the one hand, the language used in new media is negatively described by critics, with terms such as ‘language corruption’, ‘modern scourge’, ‘linguistic ruin’, ‘vandalism’, ‘foe of literacy’ and ‘bane’, while on the other hand, positive terms are used by those who are optimistic about the linguistic potential of CMC, such as ‘language enrichment’, ‘opportunity’, ‘resource’, ‘valuable’, ‘frNd of literacy’ and ‘blessing’.

Dutch youths’ CMC language is thus, as it were, embroiled in a power conflict with Standard Dutch. The standard language has overt prestige, because it is openly esteemed by many as the norm (Labov 1966): ‘official’ Dutch is dominant within the Netherlands. Although what used to be known as ‘Civilized Dutch’ (in Dutch: ‘*Algemeen Beschaafd Nederlands*’) is nowadays perhaps less used in spoken language, for one reason due to the rise of ‘Polder Dutch’ (‘*Poldernederlands*’: a speech variant that has increased in popularity in the last decades, especially among young highly-educated women, Stroop 2010), many people still regard

[6] Many of these factors are not exclusive to new media texts: they also explain (in part) other forms of language variation.

Standard Dutch as a strict norm in its written form. They consider the ‘incorrect’ and inconsistent language use in social media as a detrimental influence on their beloved language. The following reactions by parents and teachers, prompted by an article about ‘language errors’ by youths, illustrate this:

> “Got the feeling that language deterioration has been going on for years …, particularly among youths, and is getting worse. Some seem to just enjoy communicating in a kind of semi-slang. Maybe also caused by modern communication tools WhatsApp, Facebook etc ... in which it is not so important whether something is spelled correctly as long as it is understood by friends.”
>
> (‘*Heb het idee dat er al jaren … taalverloedering is, met name onder jongeren, en steeds erger wordt. Sommigen lijken het ook gewoon leuk te vinden om in een soort semi-straattaal te communiceren. Misschien ook veroorzaakt door huidige communicatiemiddelen Whatsapp, Facebook etc…waarin het niet zo van belang is of iets juist gespeld is als het maar door vrienden begrepen wordt.*’) (TN 2014)
>
> “Social media such as Facebook and WhatsApp definitely affect language deterioration”
>
> (‘*Sociale media zoals Facebook en Whatsapp hebben zeker invloed op taalverloedering*’) (Robin F 2014)

The following example from a public Internet forum shows similar concerns. A contributor is convinced that social media “cause language corruption”: they “sometimes get the impression that with the advent of Facebook & Co, the Netherlands spontaneously became dyslexic collectively” (social media ‘*leid[en] tot taalverloedering (krijg soms de induk dat met de komst van Facebook & Co Nederland spontaan collectief dyslectisch is geworden)*’) (w00t00w 2015). Another forum participant shares this critical outlook and when comparing language use in old and new media, he observes, “With newspapers and publishers, contributors could hardly afford to make a spelling error back then. With social media, this does not matter anymore at all” (‘*Bij kranten en uitgevers konden de inzenders zich toen nauwelijks een spelfoutje permitteren. Bij de sociale media maakt dat nu allemaal geen bal meer uit*’) (EricMM 2015). In short, non-standard language use on social media is criticized openly and often, and in various contexts. The overt prestige of Standard Dutch is also clear from the success of non-academic publications about language ‘errors,’ such as the immensely popular books and online communities of *Taalvoutjes* (Bogle and Hollebeek 2013), in which Dutch ‘language errors’ are made fun of.

By contrast, unconventional CMC language enjoys covert prestige among many youths, who value this non-standard language variety. They consider it as playful, informal, and cool. The use of CMC language is thus part of youth culture (Bergs 2009), may express humour, rebelliousness, and youthfulness (Shaw 2008), and is often used to mark one's social identity (Wood, Kemp and Plester 2013). In this way, CMC language bears resemblances to so-called street language (in Dutch: '*straattaal*'), an urban youth language which is spoken in the streets, particularly in multi-ethnic cities, and is characterized by influences from immigrant languages and American slang. That, too, is an informal youth language which deviates from Standard Dutch, and is therefore regarded with suspicion by many (older) people, whereas many youths consider it as fashionable and cool.[7] Street language and CMC language foster a sense of belonging to a group and help youths create their own social space (De Rooij, in Truijens 2009), and this covert prestige of CMC language also reveals itself through creativity with language in new media, such as novels and poetry written in the form of text messages or tweets. This paper examines the linguistic characteristics to which Dutch youths' CMC language owes its covert status. Put differently, this work investigates in which ways this language variant diverges from Standard Dutch, and whether these divergences are dependent on the variables of medium and age group.

2 MATERIALS AND METHODOLOGY

2.1 Data collection

For this register analysis of new media messages produced by Dutch youths, texts written in three media were selected from SoNaR ('*STEVIN Nederlandstalig Referentiecorpus*', Oostdijk et al. 2013), an existing reference corpus of written Dutch, while additional texts from one further medium, WhatsApp, were also collected. The WhatsApp messages were gathered especially for the present study: a website was created with instructions on how Dutch youths could voluntarily contribute their authentic (private) messages by sending them to a specific email address (Verheijen and Stoop 2016). Data collection was promoted via diverse national and regional media, and an added incentive for young people to donate their messages was a prize raffle among all contributors with the chance to win gift certificates. The final corpus used for this study contains 392,169 tokens of instant messages (MSN chats and WhatsApp messages), text messages, and tweets, composed by youths aged 12 to 23. These were divided into two age groups: adolescents (between the ages of 12 and 17) and young adults (18 up to 23 years old). The specifics of the corpus, and the distribution of tokens over the

[7] Just like CMC language, street language is a heterogeneous phenomenon. CMC texts from different media and by different users are distinct; likewise, there are different kinds of street language, which cannot be simply lumped together in any formal analysis.

media and age groups, are shown in Table 2. To be clear, messages in the different media – not only those in the added WhatsApp component, but overall – came from different individuals, so the corpus was not longitudinal. Due to the distribution of new media texts in SoNaR, the corpus is unfortunately imbalanced for the independent variables of medium and age group, but this does not skew the tables and figures presented below, because the frequencies of the linguistic features have been normalised per 10,000 words.

Table 2: Corpus of new media texts for analysis.

Medium	Year(s) of collection	Age group	Mean age	# tokens	# chats or contributors[8]
Instant messaging: MSN	2009-2010	12-17	16.2	45,051	106
		18-23	19.5	4,056	21
		total		49,107	127
Text messaging: SMS	2011	12-17	15.4	1,009	7
		18-23	20.4	23,790	42
		total		24,799	49
Microblogging: Twitter	2011	12-17	15.9	22,968	25
		18-23	20.6	99,296	83
		total		122,264	108
Instant messaging: WhatsApp	2015	12-17	14.0	55,865	11 / 84
		18-23	20.4	140,134	23 / 132
		total		195,999	34 / 216
		grand total		**392,169**	

2.2 Data coding

The new media texts were examined quantitatively for various linguistic features that have been found in prior research, carried out on languages other than Dutch, to be relevant for CMC: the orthographic features of textisms, misspellings, and typos; the typographic features of emoticons and symbols; the syntactic feature of omissions; and the lexical features of borrowings and interjections.

The following spelling deviations of Standard Dutch have been classified in the analysis as textisms (adapted from Plester, Wood and Joshi 2009; see also Verheijen 2013):

- **initialism**: first letters of each word/element in a compound word, phrase, (elliptical) sentence, or exclamation (cf. Daniëls' (2009) 'lettero'), e.g. *hw* < *huiswerk* ('homework'), *gmj* < *goed, met jou* ('fine, how are you'), *hjb* < *houd je bek* ('shut up'), *wtf* < *what the fuck*

[8] Number of chats: MSN, WhatsApp; number of contributors: SMS, Twitter, WhatsApp.

- **contraction**: omission of letters (mostly vowels) from middle of word (cf. Daniëls' (2009) 'shortje'), e.g. *ltr* < *later* ('later'), *hzo* < *hoezo* ('why'), *sws* < *sowieso* ('in any case')
- **clipping**: omission of final letter of word (mostly silent *-n* or *-t*), e.g. *morge* < *morgen* ('tomorrow'), *bes* < *best* ('rather'), *naa* < *naar* ('to')
- **shortening**: dropping of ending or occasionally beginning of word, e.g. *miss* < *misschien* ('maybe'), *opdr* < *opdracht* ('assignment'), *ns* < *eens* ('some time')
- **phonetic respelling**: substitution of letter(s) of word by (an)other letter(s), while applying accurate grapheme-phoneme patterns of the standard language (resulting in abbreviation, replacement, or extension), e.g. *sgool* < *school* ('school'), *meel* < *mail*, *owkeej* < *oké* ('okay')
- **single letter/number homophone**: substitution of entire word by a phonologically resembling or identical letter/number, e.g. *k* < *ik* ('I'), *m* < *hem* ('him'), *2* < *too*/*to*
- **alphanumeric homophone**: substitution of part of word by phonologically resembling or identical letter(s) and/or number(s), e.g. *opdr8* < *opdracht* ('assignment'), *id* < *idee* ('idea'), *hh* < *haha*
- **reduplication**: repetition of letter(s) (cf. De Decker's (2015) 'flooding' and Darics' (2013) 'letter repetition'), e.g. *cooool* < *cool*, *doeii* < *doei* ('bye'), *jaaa* < *ja* ('yes')
- **visual respelling**: substitution of letter(s) by graphically resembling non-alphabetic symbol(s) (special characters or numbers), e.g. *w00t* < *woot*, *j@n* < *Jan*
- **accent stylisation**: words from casual, colloquial, or accented speech spelled as they sound, e.g. *hoessie* < *hoe is het* ('how are you'), *das* < *dat is* ('that's'), *eik* < *eigenlijk* ('actually')
- **inanity**: miscellaneous spelling deviations, e.g. *eeyz* < *ey*, *duz* < *dus* ('so'), *chilliej* < *chill*
- **standard language abbreviation**: abbreviation that is part of the standard language,[9] e.g. *jan* < *januari* ('January'), *uni* < *universiteit* ('university'), *min* < *minuut* ('minute')
- unconventional use of spacing, punctuation, diacritics and capitalisation (incl. 'all caps,' i.e. entire words or utterances typed in capital letters).

[9] Of course, standard language abbreviations do not deviate from the 'official' spelling; after all, they are included in dictionaries that codify Standard Dutch. Yet these abbreviations were still included in the present analysis of textisms, since they are also typical of the succinctness and speed of CMC.

In the classification of 'misspellings,' only a number of distinct spellings deviating from Standard Dutch have been coded, and these are deviations that are strongly denounced by prescriptivist linguists or language users. These concern 'spelling errors' with *d/t*, *ei/ij*, *is/eens*, *jou/jouw*, *n* (the letter *n* used to connect two words in Dutch, or final *n*), obsolete spelling, and with borrowings. Only a select group of deviations has thus been classified as 'misspelling'; the rest has been interpreted as textism, despite the fact that these are also regarded as 'incorrect' by those who hold the standard language as the norm for all writing.

The analysis only contains manifest typos (typing errors), where the writer clearly intended to type another word, given the context. Such deviations often differed by only one letter, e.g. *hey boek* ('thy book') instead of *het boek* ('the book').

Emoticons – a portmanteau word of the words 'emotion' and 'icon' – are understood to mean symbols composed of typographic characters (punctuation marks, letters, and/or numbers) which represent facial expressions with emotions, such as :-) (a smiling face, or 'smiley') to indicate joy. These help to express the writer's feelings. Both Western variants, which should be understood by tilting one's head, and Asian/Japanese variants, which can be interpreted at face value (e.g. ^^ and -_-), have been included, although the latter (also called 'kaomoji') only occurred rarely in the present corpus. Emoticons do not include the nowadays popular emoji – which, in the new media analysed here, only occur in WhatsApp: these small, standardised images are not part of typography.

The symbols encountered in this new media corpus are as follows: *&* (and), + (and, plus), = (is, equals), <, >>, --> (arrow), € (euro, money), *<3* (heart), *X* or *x* (kiss), *K* or *k* (kiss), *(K)* or *(k)* (kiss), *o* (hug, as in xoxo), *(L)* or *[L]* (love), *(H)* or *(h)* (heart or cool), *(A)* or *(a)* (angel), *(Y)* or *(y)* (yes, okay), and * (correction, emphasis, or action).

All omissions have been coded and subsequently classified on the basis of the part of speech of the omitted elements: articles, subject pronouns (personal or demonstrative pronouns that function as the grammatical subject), other pronouns (personal/demonstrative pronouns with another grammatical function, such as object or possessive pronouns), auxiliary verbs, copula verbs, lexical verbs, combinations of subject pronoun and verb (plus possibly object pronouns), conjunctions, prepositions, and other elements (e.g. adverbs). Further analysis of these types of omissions was outside the scope of this paper.

The following lexical elements have been classified as borrowings: borrowed words, borrowed phrases, borrowed sentences, borrowed interjections, and borrowed textisms. Words that originate from other languages, but have now been officially acknowledged as part of Standard Dutch, have not been coded. The criterion used to objectively determine whether a word has been acknowledged as part of Standard Dutch was inclusion in the *Dikke Van Dale Online* dictionary, an authority among Dutch lexicons.

Interjections are expressions or utterances that do not constitute a grammatical constituent of a sentence, but stand on their own. They are mainly used to express sentiment or to imitate sounds, for example onomatopoeias conveying laughter.

These features were identified and classified entirely manually. To increase the reliability of the results, all data were checked twice by the first coder (the author). Moreover, a subset of the data (*n* = 10,010 tokens, a random sample of at least 1,000 from each subcorpus) was also coded independently by a second coder, who before this process began took part in two training sessions with the first coder to get a full grasp of the codebook. The intercoder reliability for this subset of the data was measured with Cohen's κ. It was calculated per linguistic feature, to ensure acceptable levels of reliability (except for the omissions, which were only coded by a single coder). Values ranged from 0.68 to 0.92 (see Table 3); the average intercoder reliability was κ = 0.83.

Table 3: Reliability coefficients per linguistic feature.

Linguistic features	Kappa
textisms	.92
misspellings	.70
typos	.68
emoticons	.98
symbols	.85
omissions	-
borrowings	.82
interjections	.83

2.3 Data analysis

The results reported here have been separated for medium and age group and normalised per 10,000 words, because the total number of words analysed differs per medium and age group. The results have also been subjected to statistical testing with IBM SPSS Statistics, through seven loglinear analyses and one chi-square test.

The loglinear analyses were performed on the raw frequencies, taking into account the total sample sizes. A hierarchical model was used for these analyses, containing all the lower-order interactions and main effects of the interactions examined. Seven of the eight linguistic features – textisms, misspellings, typos, emoticons, symbols, borrowings, and interjections – were treated as variables in their own loglinear analyses. Textisms, for instance, were a variable in one analysis (NB: it was thus *not* the case that 'linguistic feature' was a variable in an overall analysis and the different features, such as textisms and misspellings etc., were its levels). For each of the seven linguistic features analysed with loglinear analyses, a separate analysis was conducted with the following variables: 'medium' (MSN, SMS, Twitter, or WhatsApp), 'age group' (adolescent or young adult), and 'linguistic feature' (feature present or absent), which were all weighted by the raw frequencies. The raw frequencies of 'feature absent' were computed as follows: the total number of words per medium and age group, minus the raw frequency of linguistic feature per medium and age group, e.g. for textisms in MSN by adolescents: 45,051 - 8,398 = 36,653. As an example, Table 4 shows what the SPSS data file for the statistical analysis of textisms looked like:

Table 4: Example data file for loglinear analysis: textisms.

MSN_SMS_Twitter_or_WhatsApp	Adolescent_or_young_adult	Textism_or_not	Raw_frequency
MSN	adolescent	textism	8398
MSN	adolescent	no textism	36653
MSN	young adult	textism	347
MSN	young adult	no textism	3709
SMS	adolescent	textism	133
SMS	adolescent	no textism	876
SMS	young adult	textism	1696
SMS	young adult	no textism	22094
Twitter	adolescent	textism	1298
Twitter	adolescent	no textism	21670
Twitter	young adult	textism	4255
Twitter	young adult	no textism	95041
WhatsApp	adolescent	textism	6317
WhatsApp	adolescent	no textism	49548
WhatsApp	young adult	textism	10206
WhatsApp	young adult	no textism	129928

Since the number of instances in the corpus that are *not* omissions cannot be computed (in theory, any number of omissions can exist; irrespective of the total number of words per subcorpus), instead of a loglinear analysis, a chi-square test was conducted on the standardised frequencies of the omissions.

3 RESULTS

The following tables show the findings of the corpus study: Table 5 presents the normalised frequencies and Table 6 the results of the statistical tests.

Table 5: Normalised frequencies of the linguistic features (per 10,000 words).

Linguistic features	Instant messaging: MSN		Text messaging: SMS		Microblogging: Twitter		Instant messaging: WhatsApp	
	12-17 yrs	18-23 yrs	12-17 yrs	18-23 yrs	12-17 yrs	18-23 yrs	12-17 yrs	18-23 yrs
	norm. freq.	norm. freq.	norm. freq.	norm. freq.	norm. freq.	norm. freq.	norm. freq.	norm. freq.
textisms	1864.11	855.52	1318.14	712.90	565.13	428.52	1130.76	728.30
misspellings	24.42	27.12	19.82	6.31	16.11	10.57	24.70	13.27
typos	39.73	22.19	79.29	40.77	29.17	16.42	137.65	57.59
emoticons	690.55	236.69	198.22	356.45	216.39	196.38	83.95	101.97
symbols	16.87	4.93	267.59	237.49	24.82	20.14	39.38	22.41
omissions	518.75	315.58	356.79	479.61	390.54	423.98	620.60	493.17
borrowings	131.41	71.50	148.66	76.92	149.77	114.81	194.76	144.72
interjections	559.81	332.84	317.15	253.05	179.38	114.41	485.10	304.14

Table 6: Results of the statistical tests of the linguistic features.

Linguistic features	Interaction medium × age group × linguistic feature (DF = 3)		Interaction medium × linguistic feature (DF = 3)		Interaction age group × linguistic feature (DF = 1)	
	χ^2	Sig	Partial χ^2	Sig	Partial χ^2	Sig
textisms	97.48	***	3574.71	***	1121.06	***
misspellings	5.61	n.s.	17.84	***	30.02	***
typos	5.68	n.s.	676.18	***	305.62	***
emoticons	174.43	***	3711.52	***	12.14	***
symbols	9.41	*	1461.77	***	36.74	***
omissions	75.14	***	-		-	
borrowings	5.67	n.s.	173.01	***	91.36	***
interjections	3.71	n.s.	1692.75	***	457.63	***

N.s.: non-significant, $p > .05$; significant * $p < .05$, *** $p < .001$. DF: degrees of freedom.

3.1 Orthography

3.1.1 Textisms

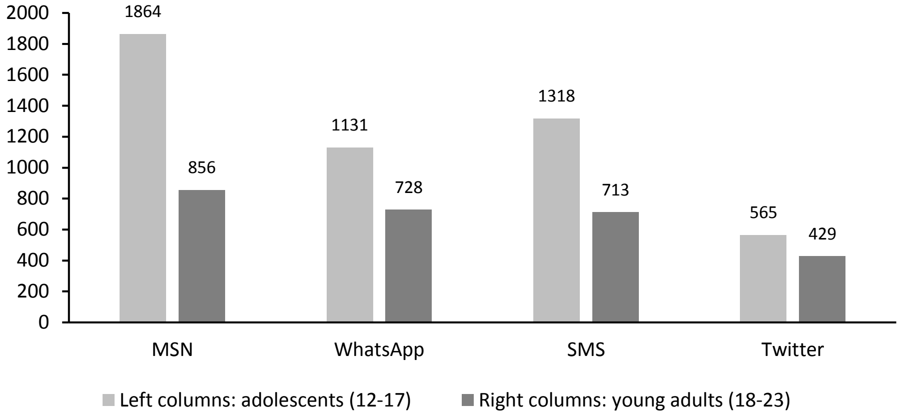


Figure 1: Normalised frequencies of textisms.

The statistical test reported in Table 6 shows that the three-way interaction medium × age group × textisms was significant (χ^2 (3) = 97.48, p < .001). Analysis of the normalised frequencies demonstrates that textisms were used more by adolescents than young adults in all media, but that this difference was dependent on medium: it was greatest in MSN chats, in which textisms occurred most, and smallest in tweets, in which they occurred least.

3.1.2 Misspellings

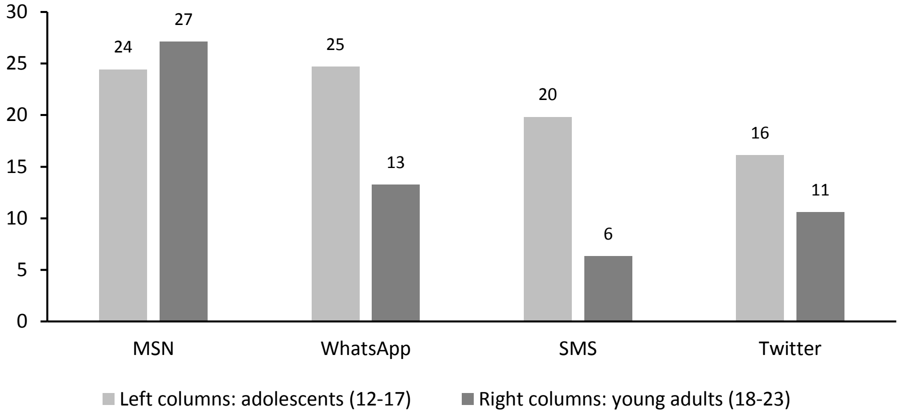


Figure 2: Normalised frequencies of misspellings.

It is apparent from the analysis that the two-way interactions medium × misspellings and age group × misspellings were significant (partial χ^2 (3) = 17.84, p < .001, partial χ^2 (1) = 30.02, p < .001). Misspellings occurred more in MSN chats than in the other three media. They were produced more by adolescents than young adults, except in MSN.

3.1.3 Typos

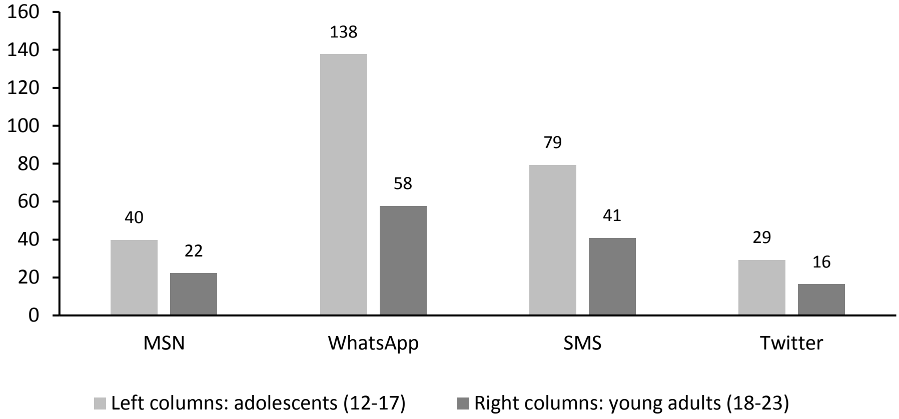


Figure 3: Normalised frequencies of typos.

The statistical tests show that both two-way interactions, namely medium × typos and age group × typos, were significant (partial χ^2 (3) = 676.18, p < .001, partial χ^2 (1) = 305.62, p < .001). More typos occurred in WhatsApp messages and then SMS text messages, than in the other two media. Adolescents made more typing errors than young adults in all four media.

3.2 Typography

3.2.1 Emoticons

Statistical tests reveal that the three-way interaction medium × age group × emoticons was significant (χ^2 (3) = 174.43, p < .001). In MSN chats, in which emoticons were most frequent, adolescents used many more of these than young adults. The situation was reversed for SMS text messages, in which it was young adults who used more emoticons. The frequencies of emoticons in WhatsApp, in which emoticons were used least, and on Twitter were close together for the two age groups.

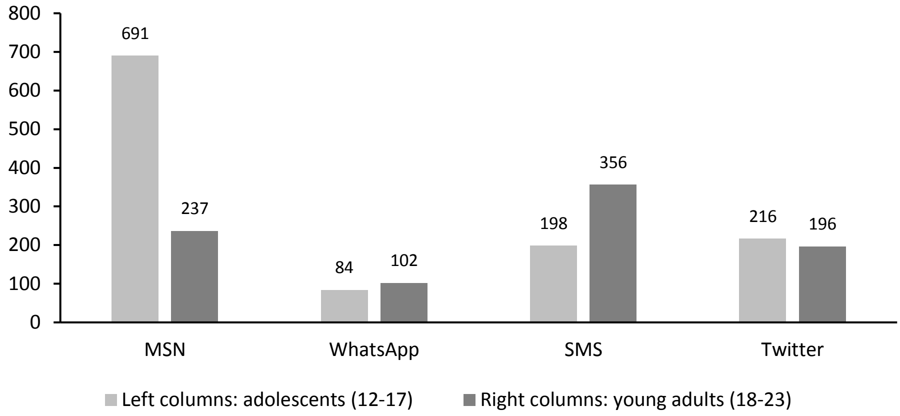


Figure 4: Normalised frequencies of emoticons.

3.2.2 Symbols

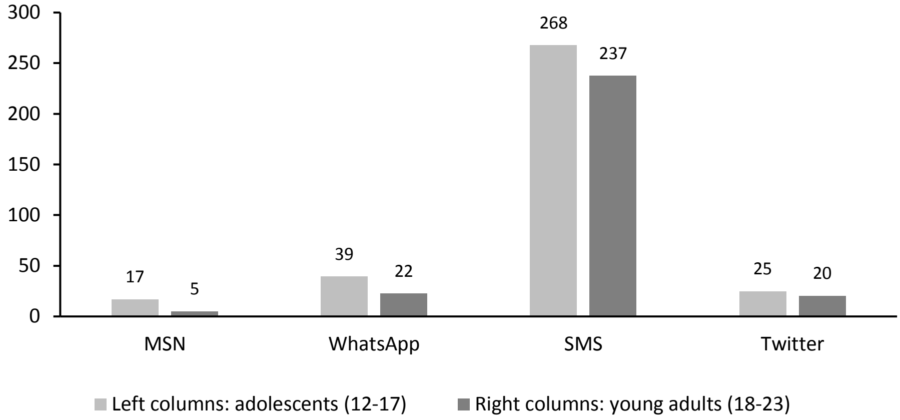


Figure 5: Normalised frequencies of symbols.

Statistical testing shows that the three-way interaction medium × age group × symbols was significant (χ^2 (3) = 9.41, $p < .05$). Symbols were used much more in SMS text messages than in the other three media, and they were used somewhat more by adolescents than young adults across all media.

3.3 Syntax

3.3.1 Omissions

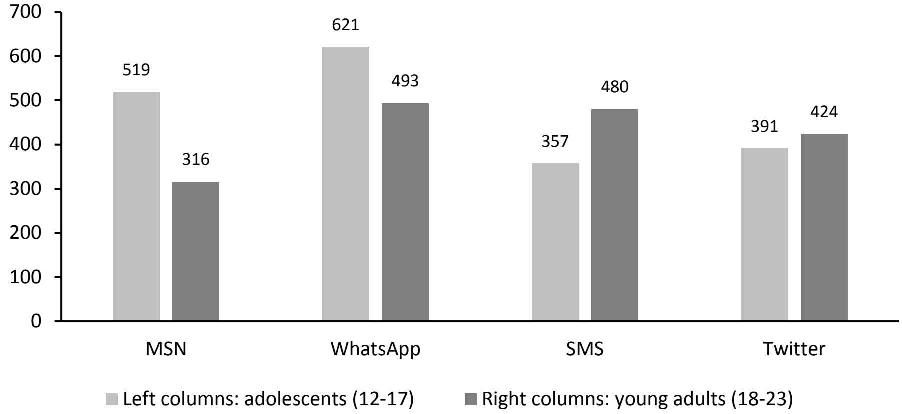


Figure 6: Normalised frequencies of omissions.

The three-way interaction medium × age group × omissions turned out to be significant (χ^2 (3) = 75.14, *p* < .001). Adolescents used more omissions than young adults in WhatsApp messages and MSN chats, while young adults used more in SMS text messages and tweets.

3.4 Lexis

3.4.1 Borrowings

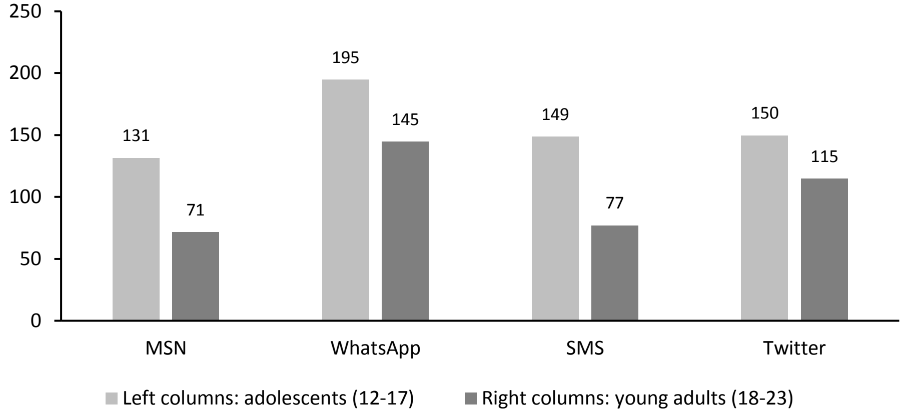


Figure 7: Normalised frequencies of borrowings.

Statistical testing reveals that the two-way interactions medium × borrowings and age group × borrowings were significant (partial χ^2 (3) = 173.01, $p < .001$, partial χ^2 (1) = 91.36, $p < .001$). Adolescents used more borrowed words, phrases, sentences, or textisms than young adults in the four media. Borrowings occurred most in WhatsApp, then on Twitter, and less frequently in SMS and MSN.

3.4.2 Interjections

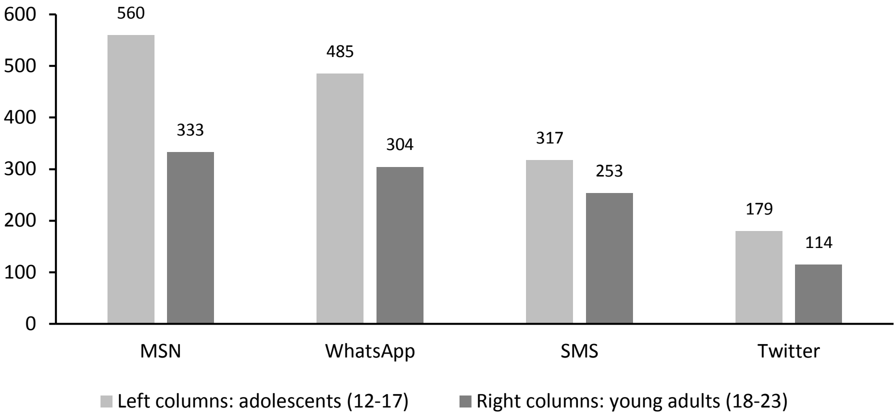


Figure 8: Normalised frequencies of interjections.

Both two-way interactions medium × interjections and age group × interjections proved to be significant (partial χ^2 (3) = 1692.75, $p < .001$, partial χ^2 (1) = 457.63, $p < .001$). Interjections were used more by adolescents than young adults in all four media. They occurred most in MSN chats and least in tweets.

4 DISCUSSION

The results for the linguistic features that were analysed in this corpus study together form the linguistic profiles of four new media and two age groups. These profiles ensue from the user characteristic age, and the various characteristics of the media examined.

4.1 Age

The results show that age plays a distinct role in the use of CMC language. This is consistent with findings by Hilte et al. (2016), who studied a corpus of Flemish computer-mediated messages and concluded that, in comparison to older youths (between 17 and 20 years old), adolescents (aged 13-16) more frequently used linguistic features of expressiveness deviating from the standard language. This was found, among other things, for reduplication of letters and punctuation, excessive use of capitalisation, emoticons, certain symbols (typographic kisses and hugs), and certain interjections (the onomatopoeic rendering of laughter) – each of these have been confirmed by the present study, with the exception of kisses, which in the present corpus were used more by young adults. Likewise, De Decker (2015), who also conducted a corpus study of Flemish CMC, observed that features such as 'flooding' (reduplication of letters), 'grapheme reductions' (phonetic abbreviations), and 'leetspeak' (incl. alphanumeric homophones and visual respellings) were used more by 13-to-16-year-olds than by 17-to-20-year-olds, as was the case in the present study. Adolescents were also found to diverge more from the standard language spelling in the Flemish written CMC studied by Peersman et al. (2016). The overall greater linguistic deviance of adolescents in CMC may be explained as follows. Teenagers, especially in puberty, are generally more non-conformist and innovative in their linguistic behaviour than adults (Eckert 1997, Androutsopoulos 2005). The most rebellious language behaviour is said to occur around the ages of 15-16, when youths feel the greatest pressure to rebel against the norms set by society, a period known as the adolescent peak (Holmes 1992). Young adults, on the other hand, feel a greater need to comply with the rules of the standard language, which has overt prestige in society. They start to feel social pressure not to appear immature, and so use Standard Dutch to conform to societal norms.

This explains why adolescents made significantly more use of textisms, typos, and symbols in all four media, and of misspellings in three media (all except MSN). In contrast, the young adults made a greater effort not to diverge from the standard language with regard to orthography and typography. The adolescents also used significantly more emoticons in MSN chats, whereas the young adults used more in SMS text messages. There appears to be no straightforward explanation for the lower frequency of emoticons in text messages by adolescents; it is possible that one or some of the contributors of text messages used very few emoticons, so an analysis of individual differences between the contributors could perhaps clarify this, especially given the rather low number of contributors of SMS text messages in the younger age group. Adolescents also diverged more from the standard language in terms of lexis: in all four media, they used relatively more borrowings, which are not (yet) part of Standard Dutch, and interjections, which are characteristic of informal spoken language, but not for written standard language.

The results for the omissions were more complicated. The frequency of omissions was much higher with adolescents in MSN and WhatsApp, while it was higher with young adults in SMS and on Twitter. This is likely to be the result of a complex interaction between this linguistic feature with the variables age group and medium, as discussed below.

4.2 Medium

The medium used is found to have a large impact on CMC language use. In fact, it appears to have a greater effect than age group for all aspects except for misspellings (partial chi-squares of 17.84 vs. 30.02), for which age group had a greater impact. The partial chi-squares were higher for medium than age group for all other linguistic features – textisms (3574.71 vs. 1121.06), typos (676.18 vs. 305.62), emoticons (3711.52 vs. 12.14), symbols (1461.7 vs. 36.74), borrowings (173.01 vs. 91.36), and interjections (1692.75 vs. 457.63). This is in line with results reported by De Decker (2015) and Hilte et al. (2016), which show that medium was a significant determinant of the frequency of 'chatspeak' features and expressive markers in Flemish youths' CMC, even more so than age. Multiple medium characteristics play a part here (see Table 1), namely limitations in message size, (a)synchronicity, visibility, interactivity, and technology. These characteristics can either encourage or discourage deviations from the standard language.

The first characteristic concerns limitations in message size. SMS text messages and tweets are limited in number of characters, as opposed to MSN chats and WhatsApp messages. The message size limit in SMS (up to 160 characters) and on Twitter (a maximum of 140) requires considerable succinctness in communication. This explains the higher frequency of omissions in SMS text messages and tweets by young adults. Young adults apparently attempt to fill their text messages and tweets with as much information as possible without exceeding the message size limit, which they can achieve by means of omissions: leaving out nonessential elements, often function words. The lower frequency of omissions in adolescents' SMS text messages and tweets, in comparison with those sent by young adults, suggests that the latter more carefully formulate their utterances to be as concise as possible. The absence of a message size limit in MSN Messenger and WhatsApp provides young adults with the space needed to conform more to the norms of the (written) standard language with regard to syntactic completeness. This characteristic also partly explains the lower frequency of interjections in SMS and on Twitter, as the character limitations in these media mean that nonessential words, such as interjections, are elided. The lack of such a limit in MSN chats and WhatsApp, by contrast, offers plenty of space for the use of interjections.

Another difference between the new media lies in synchronicity, i.e. the simultaneity of communication. Instant messaging is a (near-)synchronous medium: the communication takes place in practically real-time, which puts users under more pressure to respond quickly. The speed inherent in instant messaging is conducive to deviations from the standard language, because the high pace of communication provides little time for spelling or grammar checks. SMS and Twitter are asynchronous, so more time passes between the exchange of messages. These media offer time to edit messages and reflect upon one's words. This explains the high frequency of misspellings in MSN chats and WhatsApp messages, and of textisms in MSN chats, in comparison to the other media. It also helps to explain the high frequency of interjections in MSN and WhatsApp: the near-synchronous communication in instant messaging makes these written media resemble a spoken conversation, in which interjections are common (although, of course, the conditions for verbalisation and mutual awareness in written CMC are not the same as those in spoken language). The asynchronous communication in SMS and on Twitter endows these media with more of the characteristics of written language. Synchronicity is also related to omissions. The higher frequency of omissions in adolescents' MSN chats and WhatsApp messages, in comparison to their SMS text messages and tweets, is inconsistent with the aforementioned limit on message size in SMS and on Twitter. This finding can be attributed to the synchronicity of instant messaging, which causes users to communicate in ways similar to informal speech – with many sentence fragments and omitted words. Young adults use this synchronicity slightly less eagerly: in MSN and WhatsApp, they also imitate an informal conversation, but take somewhat more time than adolescents to write syntactically more complete sentences; they are not pressed for time, because there are no limits on the message size.

New media also differ in terms of visibility and interactivity, two characteristics that are strongly linked. Communication in MSN chats, SMS text messages, and WhatsApp messages is private and typically one-to-one (interaction between two people), and so visible for a small number of selected interlocutors, whereas communication on Twitter is usually public and one-to-many, so it can be read by a greater number of people. Tweets are often more aimed at informing a wider audience rather than sending personal messages. The public character of tweets discourages users to diverge from the standard language norms, in contrast with the privacy of the other three media. This explains the low frequency of textisms, misspellings, and typos in tweets. The high frequency of symbols in SMS, notably of hearts (*<3*) and kisses (esp. *X* and *x*) to conclude SMS text messages, reflects the personal character of this medium. In addition, this characteristic explains the high frequency of emoticons in especially the MSN chats written by adolescents and SMS text messages by young adults. This results from the one-to-one (or sometimes some-to-some, in MSN) private communication taking place via these media, in which emoticons are regularly used to convey the writer's feelings and to avoid misunderstandings about the sentiment behind an utterance, as opposed to the generally one-to-many public communication of tweets, which require fewer emoticons because their content is often more neutral and less focused on emotions. The lowest frequency of emoticons in WhatsApp has a completely different cause: in this medium, the pragmatic functions of emoticons are also fulfilled by emoji.[10] Furthermore, the characteristic of visibility explains the high frequency of English borrowings in tweets in particular. The English language currently enjoys prestige among Dutch youths, and using English words is thus seen as 'hip' and 'cool' among this group. That is why they are frequently used in tweets, whose public nature allows a large audience to witness how 'cool' the writer is. Yet this does not explain the high frequency of borrowings in WhatsApp messages, which may, in fact, be caused by a temporal development: perhaps the use of English words has become even more popular between the times of collecting the SoNaR data and the WhatsApp data.

Finally, new media are used on different technological devices. MSN Messenger was a chat program for computers; text messages and WhatsApp messages are usually sent via mobile phones; while tweets are sent from either computers or mobile phones. These devices differ as to their keyboards and possibilities of using a predictive dictionary. The frequency of textisms in SMS text messages, tweets, and WhatsApp messages, and of misspellings in the former two media, may be lower because mobile phones, from which these messages are usually sent, often contain a predictive dictionary (which users can choose to utilize or not, to their own liking): when typing the first letter(s), the software 'guesses' the rest of the word. The words in the digital dictionaries that are used for this are spelt according to the standard language orthographic rules, which decreases the chance of textisms. However, such a predictive dictionary was not used with MSN chats. Moreover, the frequency of typos in SMS text messages may be higher than otherwise because of the small keypads on mobile phones, which increase the risk of typos.[11] A computer keyboard, as was used with MSN chats, has larger keys and thus presents a lower risk of typos. Typos also seem to be more affected by technology than synchronicity, seeing that the asynchronous communication of SMS does offer sufficient time for checking and correcting typos. Finally, the frequency of omissions in WhatsApp as compared to MSN – both

[10] Emoji could not be coded in the present study due to the file format in which WhatsApp messages were contributed to the corpus.

[11] Mobile phones can have an alphanumeric keyboard, with which three or four letters and a number are assigned to a single key, or a (possibly touchscreen) QWERTY keyboard, which is comparable to a computer keyboard, but much smaller. This is likely to affect the risk of typos, but unfortunately there was no information available about the devices with which the new media texts in the corpus were produced.

near-synchronous media which encourage omitting some elements to achieve a conversational writing style – can be explained by technological differences. The frequency of omissions is even higher in WhatsApp, because the small keyboards of mobile phones provide users with an extra incentive to omit parts of speech, whereas the large computer keyboards used for MSN did not.

5 CONCLUSION

It can be concluded from the results of this corpus study that, as expected, the language Dutch youths use when they communicate via social media indeed diverges from Standard Dutch on several writing dimensions, namely orthography, typography, syntax, and lexis. As for orthographic peculiarities, this CMC language is overall characterized by textisms (which include deviations in letters as well as in spacing, diacritics, punctuation, and capitalisation), misspellings, and typing errors. Typographic features are symbols and emoticons – as well as emoji in WhatsApp, but those concern visuals rather than typography. Regarding syntax, CMC language deviates from the written standard by its many omissions. Characteristic of the vocabulary of CMC language are borrowings, especially English ones, and interjections.

More importantly, this register analysis clearly shows the effects of medium and age group on the frequency with which certain linguistic features occur in computer-mediated messages. All interactions between medium (MSN, SMS, Twitter, and WhatsApp) and each of the linguistic features were highly statistically significant, due to an interplay of different medium characteristics. This was also the case for all interactions between age group (adolescents, young adults) and the linguistic features. Factors such as age and especially medium, whose impact was even greater, thus make sure that 'CMC language' is not a homogeneous language variant – rather, it encompasses various registers. The present study thus emphasizes the crucial importance of the variables age and medium for online language use, as attested in (Dutch) written computer-mediated communication, and once more confirms that youths' online writings offer a wealth of linguistic diversity.

6 LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

A drawback of this study is that the collection periods for different parts of the corpus were not the same. The SoNaR texts were collected between 2009 and 2011, thus quite some years ago. The WhatsApp messages are more recent, collected in 2015. It is not inconceivable that Dutch youths' CMC language has changed somewhat between these collection periods; after all, language is subject to change, and this is particularly true for youth languages, which are dynamic and constantly evolving. This means that some of the differences found between the WhatsApp data and the data from the other three media could possibly be attributed not just to the characteristics of the various media, but also (partly) to temporal developments. Analysis of more recent data would, therefore, be a welcome addition to the current study.

It would also be interesting to expand the analysis in terms of age groups, with the addition of digital texts written by children (for instance, aged 6-11 years). Yet due to practical and ethical considerations, collecting such private texts from young children could pose a real challenge. Besides expanding the corpus in age, it could also be enlarged in terms of medium. The study reported here has examined four well-known new media, while of course there are many more, and those that are popular among young people change very rapidly. Future research could thus analyse other media. It would be valuable to complement this register analysis with, for example, Facebook posts. In fact, these were already collected by the author between December 2015 and May 2016, so such an analysis would be a viable option for a future study.

Online language variability among new media could also be studied more in depth by including even more media characteristics into the research design, e.g. focusing on the software used to compose the messages, such as whether or not it includes predictive dictionaries, autocorrection, or spelling checkers. Additional user or situational characteristics, such as (the users' relationship with / profile of) the conversational partner and the communicative purpose of the interaction, would also be exciting ways to expand the analysis.

As a concluding suggestion, one more possibility for future corpus-lingusitics studies into CMC would be to include an extra independent variable, besides age and medium, with an obvious choice being gender. Other research suggests that there are differences between girls and boys in the use of several linguistic features of digital writing (e.g. Wolf 2000, Baron 2004, Parkins 2012, Hilte et al. 2016). This could be further explored for Dutch computer-mediated messages, to gain an even more nuanced picture of the registers that exist within CMC language.

Given that the language with which Dutch youths communicate via social media clearly diverges from Standard Dutch, chances are that this informal CMC language interferes with their more formal 'school language.' However, prior research does not provide a conclusive answer as to whether this is indeed the case. Therefore, this open issue will be investigated in future studies of the author's ongoing (doctoral) research project into the impact of CMC on literacy. As such, the present corpus study is only a first step in studying Dutch youngsters' written CMC. The next steps will dig deeper into the possible relation between Dutch youths' social media use and their writing skills. This will be examined in both a correlational study and an experimental study. The former to see if any evidence for a relationship can be found, the latter to explore the causality of this relationship (if it exists at all), and thus whether it is indeed CMC that affects literacy, and not vice versa. In this extended outlook, let me briefly outline the design of these two studies.

Youths who will participate in the correlational study will be tested at school, so in an educational setting. They will first write an essay – with the text genre of expository discussion – to measure their formal writing skills. Subsequently, they will fill in questionnaires about their social media use. The essays will be analysed for several measures of writing quality, namely lexical richness, syntactic complexity, formality, and writing productivity. It will then be examined whether participants' CMC use (in terms of frequency, variety, intensity, use of textisms, etc.), as self-reported in the surveys, correlates with the writing quality of their essays. This work will thus study whether participants' private online writing habits are related to the quality of the 'offline' texts they write at school.

The experimental study will use social media as the experimental prime. All school classes that participate will be divided into two groups: an experimental group, who will communicate via WhatsApp together during the priming phase, and a control group, who will spend that time on a control task, namely colouring mandalas. All participants will then write stories – with the genre of narrative storytelling – to test their productive writing skills, which will again be analysed for several measures of writing quality. Next, they will complete a grammaticality judgement task (GJT), to test their receptive grammar and spelling skills: they will be presented with sentences in which they have to spot and correct 'language errors,' i.e. deviations from Standard Dutch. It will then be measured whether the immediately preceding use of WhatsApp has a direct impact on the writing quality of the experimental groups' stories or on their performance on the GJTs.

Both studies will involve youths from different educational levels and age groups, to find out if these are mediating factors in the potential impact of Dutch youths' informal written CMC on their more formal writing skills. We hypothesize that writers of a younger age group or lower educational level could experience a greater extent of interference of social media on their school writings. Irrespective of what these future studies will find, it is nevertheless important to point out to all youngsters, no matter their age or education, that the informal digital language they use in computer-mediated messages and the standard language are different variants (registers) of Dutch – variants they ought to keep separate and employ effectively depending on the context.

Acknowledgments

This study was made possible by a grant of the Dutch Organisation for Scientific Research (NWO, *Nederlandse Organisatie voor Wetenschappelijk Onderzoek*), project number 322-70-006. Furthermore, thanks go to the anonymous reviewers, as well as my supervisors Wilbert Spooren and Ans van Kemenade, for their useful comments on earlier drafts of this paper.

References

Androutsopoulos, Jannis, 2005. Research on youth language. Ammon, Ulrich, Norbert Dittmar, Klaus J. Mattheier and Peter Trudgill (eds.): *Sociolinguistics: An International Handbook of the Science of Language and Society* 2. Berlin: Mouton de Gruyter. 1496–1505.

Androutsopoulos, Jannis, 2011. Language change and digital media: A review of conceptions and evidence. Kristiansen, Tore and Nikolas Coupland (eds.): *Standard Languages and Language Standards in a Changing Europe*. Oslo: Novus Press. 145–161.

Baron, Naomi, 2004. See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23/4. 397–423.

Bergs, Alexander, 2009. Just the same old story? The linguistics of text messaging and its cultural repercussions. Rowe, Charley and Eva L. Wyss (eds.): *Language and New Media: Linguistic, Cultural, and Technological Evolutions*. Cresskill, NJ: Hampton Press. 55–73.

Bogle, Vellah and Inger Hollebeek, 2013. *Taalvoutjes: het boek*. Utrecht: Van Dale.

Broeren, Karin, 2012. De 10 irritantste taalfouten op social media. *Ze.nl*. http://www.ze.nl/p/141435/de_10_irritantste_taalfouten_op_social_media. (Last accessed 29 June 2017.)

Carrington, Victoria, 2004. Texts and literacies of the Shi Jinrui. *British Journal of Sociology in Education* 25/2. 215–228.

Crystal, David, 2006. *Language and the Internet* (2nd ed.). Cambridge: Cambridge University Press.

Crystal, David, 2008. *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

Daniëls, Wim, 2009. *Sms & msn: hoest begonnuh?, hoe sgrijf jut?, ist errug?* Houten: Prisma.

Darics, Erika, 2013. Non-verbal signalling in digital discourse: The case of letter repetition. *Discourse, Context & Media* 2/3. 141–148.

De Decker, Benny and Reinhild Vandekerckhove, 2012. English in Flemish adolescents' computer-mediated discourse: A corpus-based study. *English World-Wide* 33/3. 321–352.

De Decker, Benny, 2015. Prototypische chatspeakkenmerken in Vlaamse tienerchattaal: de invloed van gender, leeftijd en medium. *Taal en Tongval* 67/1. 1–41.

Drouin, Michelle and Claire Davis, 2009. R u txting? Is the use of text speak hurting your literacy? *Journal of Literacy Research* 41/1. 46–67.

Eckert, Penelope, 1997. Age as a sociolinguistic variable. Coulmas, Florian (ed.): *The Handbook of Sociolinguistics*. Oxford: Blackwell. 151–167.

EricMM, 2015. Taalfouten: geen verloedering maar verandering. *Joop*. http://www.joop.nl/opinies/taalfouten-geen-verloedering-maar-verandering. (Last accessed 29 June 2017.)

Ferrara, Kathleen, Hans Brunner and Greg Whittemore, 1991. Interactive written discourse as an emergent register. *Written Communication* 8/1. 8–34.

Frehner, Carmen, 2008. *Email - SMS - MMS: The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Bern: Peter Lang.

Hård af Segerstad, Ylva, 2002. *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. Dissertation University of Gothenburg.

Herring, Susan, 2001. Computer-mediated discourse. Schiffrin, Deborah, Deborah Tannen and Heidi E. Hamilton (eds.): *Handbook of Discourse Analysis*. Oxford: Blackwell. 612–634.

Herring, Susan, 2012. Grammar and electronic communication. Chapelle, Carol A. (ed.): *Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. 1–11.

Hilte, Lisa, Reinhild Vandekerckhove and Walter Daelemans, 2016. Expressiveness in Flemish online teenage talk: A corpus-based analysis of social and medium-related linguistic variation. Fišer, Darja and Michael Beißwenger (eds.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Academic Publishing Division of the Faculty of Arts of the University of Ljubljana. 30–33.

Holmes, Janet, 1992. *An Introduction to Sociolinguistics*. London: Longman.

Jacobs, Gloria, 2008. People, purposes, and practices: Insights from cross-disciplinary research into instant messaging. Coiro, Julie, Michele Knobel, Colin Lankshear and Donald J. Leu (eds.): *Handbook of Research on New Literacies*. New York, NY: Routledge. 469–490.

Labov, William, 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste and Ineke Schuurman, 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. Spyns, Peter and Jan Odijk (eds.): *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme*. Heidelberg: Springer. 219–247.

Parkins, Róisín, 2012. Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5/1. 46–54.

Peersman, Claudia, Walter Daelemans, Reinhild Vandekerckhove, Bram Vandekerckhove and Leona Van Vaerenbergh, 2016. The effects of age, gender and region on non-standard linguistic variation in online social networks. http://arxiv.org/abs/1601.02431. (Last accessed 29 June 2017.)

Plester, Beverly, Clare Wood and Puja Joshi, 2009. Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. *British Journal of Developmental Psychology* 27/1. 145–161.

Proudfoot, Candice, 2011. *An Analysis of the Relationship between Writing Skills and 'Short Messaging Service' Language: A Self-Regulatory Perspective*. Dissertation Potchefstroom Campus, North-West University.

Robin F, 2014, August 8. De jongste taalfouten. *Onze Taal*. https://onzetaal.nl/nieuws-en-dossiers/weblog/de-jongste-taalfouten. (Last accessed 29 June 2017.)

Shaw, Philip, 2008. Spelling, accent and identity in computer-mediated communication. *English Today* 24/2. 42–49.

Silva, Cláudia, 2011. Writing in Portuguese chats :). A new wrtng systm? *Written Language & Literacy* 14/1. 143–156.

Stroop, Jan, 2010. *Hun hebben de taal verkwanseld: over Poldernederlands, 'fout' Nederlands en ABN*. Amsterdam: Athenaeum-Polak & Van Gennep.

TN, 2014, June 30. De jongste taalfouten. *Onze Taal*. https://onzetaal.nl/nieuws-en-dossiers/weblog/de-jongste-taalfouten. (Last accessed 29 June 2017.)

Truijens, Aleid, 2009, July 21. Straattaal: Algemeen Cool Nederlands. *De Volkskrant*. http://www.volkskrant.nl/binnenland/straattaal-algemeen-cool-nederlands~a339199. (Last accessed 29 June 2017.)

Verheijen, Lieke, 2013. The effects of text messaging and instant messaging on literacy. *English Studies* 94/5. 582–602.

Verheijen, Lieke, 2016. De macht van nieuwe media: hoe Nederlandse jongeren communiceren in sms'jes, chats en tweets. Van de Mieroop, Dorien, Lieven Buysse, Roel Coesemans and Paul Gillaerts (eds.): *De macht van de taal: Taalbeheersingsonderzoek in Nederland en Vlaanderen*. Leuven / Den Haag: Acco. 275–293.

Verheijen, Lieke and Wessel Stoop, 2016. Collecting Facebook posts and WhatsApp chats: Corpus compilation of private social media messages. Sojka, Petr, Aleš Horák, Ivan Kopeček and Karel Pala (eds.): *Text, Speech and Dialogue: 19th International Conference, TSD 2016, LNAI 9924*. Cham: Springer. 249–258.

Werry, Christopher, 1996. Linguistic and interactional features of Internet Relay Chat. Herring, Susan (ed.): *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam: Benjamins. 47–63.

Winzker, Kristy, Frenette Southwood and Kate Huddlestone, 2009. Investigating the impact of SMS speak on the written work of English first language and English second language high school learners. *Per Linguam* 25/2. 1–16.

Wolf, Alecia, 2000. Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior* 3/5. 827–833.

Wood, Clare, Nenagh Kemp and Beverly Plester, 2013. *Text Messaging and Literacy: The Evidence*. London: Routledge.

W00t00w, 2015, July 20. 'Bijna driekwart Nederlandse tieners heeft Instagram en Snapchat op telefoon'. *Tweakers*. https://tweakers.net/nieuws/104306/bijna-driekwart-nederlandse-tieners-heeft-instagram-en-snapchat-op-telefoon.html. (Last accessed 29 June 2017.)

Gender and grammatical Frequencies in social media English from the Nordic countries

***Steven Coats,** University of Oulu*

Abstract

English has become firmly established as a primary vehicle for global communication, and is thus also increasingly used in online contexts for local communicative purposes, for example in the Nordic societies. This paper investigates the extent to which English is used on Twitter in the Nordic countries and builds on previous research by investigating the link between gender and grammatical or part-of-speech frequencies, a link which has hitherto been considered mainly in the context of data collected in L1 Anglophone contexts.
The Twitter Streaming API was used to create a corpus of English-language messages originating from the Nordic countries. Automatic methods were used to disambiguate author gender and apply part-of-speech tags, and the relative frequencies of grammatical types by gender were determined for each country. Principal components analysis shows that Nordic English-language discourse on Twitter diverges according to gender for a number of grammatical features. The analysis supports L1 findings pertaining to gendered differences in feature frequencies in English.

Keywords: Twitter, CMC, sociolinguistics, gender, corpus linguistics

1 INTRODUCTION

Recent shifts in communication behavior towards online social media platforms provide opportunities for the study of variation in English as it is used worldwide. While the status of English, as the world's principal lingua franca, continues to consolidate in many global contexts of use, it is hardly a monolithic entity: English as it is used in global computer-mediated communication (CMC) exhibits a great variety of features in orthography, lexis, grammar, and style, especially in non-L1 environments. Such diversity has been characterized by Blommaert (2012) as a "supervernacular".

CMC and social media such as Twitter have become important sites of interaction for many, and in recent years a number of studies have investigated various properties of Twitter language (for an overview of the communicative and discourse functions of Twitter language, see Page 2012, Zappavigna 2011, and Squires 2015). The ubiquity and volume of Twitter data, its public availability through a well-maintained set of APIs (*Application Programming Interfaces*), and the extensiveness of the associated tweet metadata fields allow for a rich variety of analyses. As a significant proportion of tweets are associated with metadata detailing the physical location of their authors, geographical analyses of language use and linguistic diversity have been a natural focus of research interest (e.g. Leetaru et al. 2013, Mocanu et al. 2014). Twitter data has also been used to investigate dialectological (Eisenstein et al. 2014) and sociolinguistic aspects of American English, including the relationship between gender and language variation (Bamann, Eisenstein and Schnoebelen 2014).

Differences between the genders in the relative frequency of lexical types or word classes have been investigated in a number of studies. A large, corpus-based study of lexical type frequencies based on writing samples submitted to a website found significant differences between males and females in the relative frequencies of pronouns, numbers, negators, articles, and prepositions, among other world classes (Newman et al. 2008). Corpus-based research using language data extracted from instant messaging or blog posts has also found that some differences in feature frequency can be associated with gender. For example, it has been found in online writing that females may use more personal pronouns, modal verbs, and emoticons, while males use more determiners such as articles or demonstrative pronouns and more numbers or numerals (Baron 2004, Herring and Paolillo 2006, Argamon et al. 2007). Similar findings have resulted from a large-scale investigation of word frequencies and gender on Twitter, although gender-based associations with particular features are typically less strong than associations based on local networks (Bamann, Eisenstein and Schnoebelen 2014). For the most part, however, analysis of type frequencies in English has been conducted on data from Anglophone contexts, mainly in the United States, and relatively little corpus-based research has looked into relative frequencies in non-L1 contexts.[1] Frequency-based analyses of variation in global Englishes as they are manifest in aggregate online media such as Twitter have not yet been undertaken on a large scale, although some studies exist.[2] Given the global nature of social media and the ever-increasing importance of English, variation in English in global contexts represents an important site of language variation and change.

Knowledge of English is extensive in the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland, nations with well-developed economies and high levels of educational attainment. With populations that are to a large degree bilingual in a national language and English, the Nordic countries are perhaps the societies in which English is most extensively used without being an official language: English is so prevalent in the Nordics that it has been suggested that the national languages are becoming linguistic systems with "restricted functional range" (Görlach 2002: 16). Although much research has addressed various aspects of English use in the Nordic countries (for Sweden, e.g., see Bolton and Meierkord 2013; for Finland see the extensive survey study of Leppänen et al. 2011), and some preliminary work on language use on Twitter by country has also provided data for the Nordics (Mocanu et al. 2013), linguistic diversity on social media in Northern Europe has not been investigated in detail. Likewise, although some work exists on grammatical feature frequencies in Nordic non-CMC genres (e.g. for Swedish in Allwood 1998), there are few studies of feature frequencies in English in non-L1 environments, and the relationship between author gender and feature frequency in CMC or social media language varieties such as Twitter has not yet been explored in Nordic contexts, whether in local languages or English.[3]

This study adopts an approach based in part on multidimensional analysis (Biber 1988, 1995). After establishing the extent to which English is used on Twitter in the Nordic national contexts, relative grammatical feature frequencies are calculated and the features most strongly associated with gender identified. Using principal components analysis, the underlying associations among feature frequencies, gender, and communicative function are established.

[1] See, however, Xiao 2009 for a corpus-based investigation of world English varieties as represented in the International Corpus of English.

[2] E.g. Coats (2016).

[3] For an analysis of feature frequencies in English as it is used in various Asian contexts see Xiao (2009). Baron (2004) analyses a small corpus of Instant Messenger data in English from American and Swedish university students.

2 METHODS

The methods used in the study include the collection of data from Twitter's Streaming API, the filtering of this data to remove tweets sent by bots or other non-human agents, the disambiguation of tweet author gender and assignation of tweets to gendered subcorpora, the assignation of exact location and language to each tweet, the tokenization of tweets, part-of-speech tagging of the English-language tweets, and the statistical analysis of the resulting subcorpora. Data collection, filtering, and statistical analysis were done in Python and in R.

2.1 Data collection

Data was collected in .json format from Twitter's Streaming API from 9 November 2016 until 18 February 2017 by utilizing the *Tweepy* library in Python (Roesslein 2015).[4] The data collection script saved only tweets with a populated *place* field.

2.2 Filtering for automatic tweets

A substantial proportion of messages on Twitter are automatically generated texts created by bots or scripts, some of which automatically generate English text. The *Foursquare* app, for example, can automatically tweet short English-language sentences about a user's GPS-determined location. In an effort to reduce the potential error that such messages could introduce into the analysis (such users may not necessarily author any English-language tweets), an initial filtering step selected from the metadata *source* field those sources that are likely to be used by human agents.[5]

2.3 Geolocation

When composing a tweet, users often select a *place* from a list automatically generated by Twitter. These place suggestions are based on a user's IP address, with the coordinates automatically assigned by Twitter as a bounding box of latitude-longitude coordinates in the tweet's metadata. Some users (those using smartphones or

[4] https://github.com/tweepy/tweepy.

[5] The sources selected were *Twitter Web Client*, *Twitter for iPhone*, *Twitter for Android*, *Twitter for iPad*, *Twitter for Windows Phone*, *Twitter for Android Tablet*, *Tweetbot for Mac*, and *Instagram*. Although there were over 1,500 sources in the initial data, these eight accounted for 91% of all the tweets collected from the Streaming API.

other GPS-enabled devices) additionally opt to broadcast exact latitude-longitude coordinates with each status update; these appear in the *geo* metadata field.

Each tweet in the data was assigned exact latitude-longitude coordinates: either the exact coordinates from the *geo* field, or (if no GPS coordinates were available), a set of latitude-longitude values calculated as the center of the bounding box circumscribing the *place* field. Although users can manually enter a *place* that does not correspond to their physical location, this does not seem to occur on a large scale. For tweets that contained both *place* and *geo* objects, the product-moment correlation of the coordinate values in the Nordic data was 0.989 (for longitude) and 0.960 (for latitude).[6]

Filtering for the *country_code* field selected only tweets with geo-coordinates within the territorial boundaries of the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland. Of the 310.7 million tweets collected globally in the initial dataset, 1.76m were from the Nordic countries.

Subcorpora were prepared for each country by filtering the data according to the *language* field: tweets in the principal national language(s), and tweets in English.[7] Tweets originating from outside the Nordic countries and in other languages were not further considered. The English-language data comprised in total 460,260 tweets and 6,360,835 tokens.

2.4 Gender disambiguation

Unlike some social media platforms, Twitter does not provide users with a profile field where gender is reported; nor are users required to otherwise supply gender information. In the absence of self-reported gender information, an automatic procedure for gender disambiguation based on values in the *author_name* field was employed. Disambiguation of tweet author gender based on gender-name associations has been employed for data from the United States (Rao et al. 2010; Mislove et al. 2011),[8] but, to the best of our knowledge, not for the Nordic countries.

[6] Some *place* values in the data were obviously not accurate, such as over 1,000 tweets with a *place* value for Bouvet Island, a small, uninhabited sub-Antarctic island. Twitter uses an internal database of *places* that includes places with ISO-3166 codes; these place names (and others) are then automatically suggested to users based on their IP address and keyboard input when they are selecting a *place* for a tweet. The *location* field in the Twitter user profile utilizes the same Twitter-internal database of locations from which users can select the appropriate one.

[7] Based on the value in the *language* field. For Norway, both *Nynorsk* and *Riksmål* were categorized as “Norwegian”. For Finland, corpora were also created for the country’s second official language, Swedish.

[8] Latent attribute inference using Twitter data manually tagged for gender is a popular topic in machine learning (cf. Pennacchiotti and Popescu 2011; Ciot, Sonderegger and Ruths 2013). The approach used here relies on the association between given name and author gender, rather than using machine learning to infer gender based on the content of messages whose authors’ gender has been manually tagged.

In order to assign tweets to male or female gender categories, lists of the most frequent given names in the Nordic countries were obtained from the national statistical offices. The *author_name* field for each tweet was then filtered via regular expressions for strings that either begin with or include as a discrete element the most common male and female given names in the corresponding Nordic country.[9] While extensive name information was available for Denmark, Sweden, and Finland, it was less available for Iceland and Norway. In total, 13,506 unique male and 15,497 unique female given names from the lists were matched with the value of the *author_name* attribute for each unique user in the dataset. Users matching both male and female names were discarded. The method assigned gender to 61.5% of Nordic tweets (25% of Iceland, 57% of Norway, 60% of Denmark, 63% of Sweden, and 70% of Finland tweets).[10]

2.5 Additional text filtering

Before tokenization and part-of-speech tagging was undertaken, HTML escape characters in the *text* field were replaced with the corresponding characters. The following subcorpora were created for further analysis: First, from the gender-disambiguated data, for each country a subcorpus of tweets in all languages, in order to gauge the relative representation of different languages in the Nordics. Second, for each Nordic country a male subcorpus and a female subcorpus consisting of English-language messages geo-located to those countries whose *author_name* values matched the corresponding list of frequent male and female given names.

2.6 Tokenization and part-of-speech tagging

The Carnegie-Mellon University Twitter Tagger (Gimpel et al. 2011, Owoputi et al. 2013) was used to tokenize the gendered English-language subcorpora and apply part-of-speech tags using a subset of the Penn Treebank tagset (Marcus, Marcinkiewicz and Santorini 1993), with additional tags for the Twitter-specific features *username*, *hashtag*, and *retweet*. The tool was trained on Twitter data and is somewhat tolerant of the non-standard orthography typical of Twitter messages.

[9] http://www.statice.is, http://www.ssb.no/befolkning, http://www.scb.se/sv_/Hitta-statistik, and the open data portal for Finland https://www.avoindata.fi.

[10] The differences are due in part to the somewhat different name frequency information obtained from the national statistical offices. For example, only 402 given names were obtained from Iceland, but 1741 from Norway, 5,382 from Denmark, 25,226 from Sweden, and 7,899 from Finland. For a dataset of American tweets disambiguated for gender using name data from the U.S. Census Bureau, Mislove et al. report 64.5% gender disambiguation and a similar overrepresentation of males (2011: 556). The reason for the male overrepresentation in the data is unknown: Males may be more active on Twitter, or for whatever reason, may be more likely to use their legal name in the *author_name* field.

3 ANALYSIS AND DISCUSSION

The linguistic profiles of the national subcorpora were determined, and the relationship between gender and grammatical features in English-language messages assessed using Student's t-tests of population means. Principal components analysis was used to investigate underlying variability and so gauge the extent to which males and females from the Nordic countries may utilize different communicative styles in English on Twitter.

3.1 Language profile

English is extensively used in Twitter user messages originating from the Nordic countries. Table 1 shows the proportions of tweets in the national language(s), English, and other languages for tweets that were assigned gender based on the *author_name* values.[11]

Table 1: Percent tweets by country and language.

	Nat. Lang.	English	Other
Iceland	74.4	13.7	11.9
Norway	43.5	27.1	29.3
Denmark	38.3	41.5	20.2
Sweden	57.5	23.3	19.2
Finland	63.2	22.6	14.2

Use of English on Twitter is most extensive in Denmark, followed by Norway, Sweden, Finland, and Iceland. For the combined male and female data, the proportion of tweets in English by province is shown in Figure 1.[12] Although clear patterns of English use within the individual Nordic countries are not evident, there is a trend towards higher rates of English use in capital regions and more urbanized areas: For example, the territories of the national capitals

[11] For Finland, the percentage shown includes messages in the national languages of Finnish and Swedish (Finnish = 62.0% of tweets, Swedish = 1.2%). "Other" includes tweets classified as in other languages, as well as (typically short) tweets whose language could not be automatically detected.

[12] As of early 2017, the Twitter-internal library of places which are prompted to users when they compose tweets does not contain any province or city names for Iceland. Only the place "Iceland" can be given. As such, tweets from Iceland with a *place* value but without exact GPS coordinates are located in the center of the latitude-longitude bounding box around the country. For this data, this falls within the province of Norðurland vestra, which in Figure 1 has an English density of 12.4%. Because relatively few of the gendered tweets contain GPS coordinates (for Iceland 5.7%) and far more tweets have *place* coordinates, the overall percentage of English tweets in the gendered data from Iceland is 13.7%.

of Oslo, Copenhagen, Stockholm, and Helsinki show a higher proportion of tweets in English than do their respective countries overall. In a sociolinguistic context, such a pattern may demonstrate the fact that residents of capitals and larger cities typically have above-average levels of income and educational attainment, and that English may serve as a high-prestige language associated with internationality.

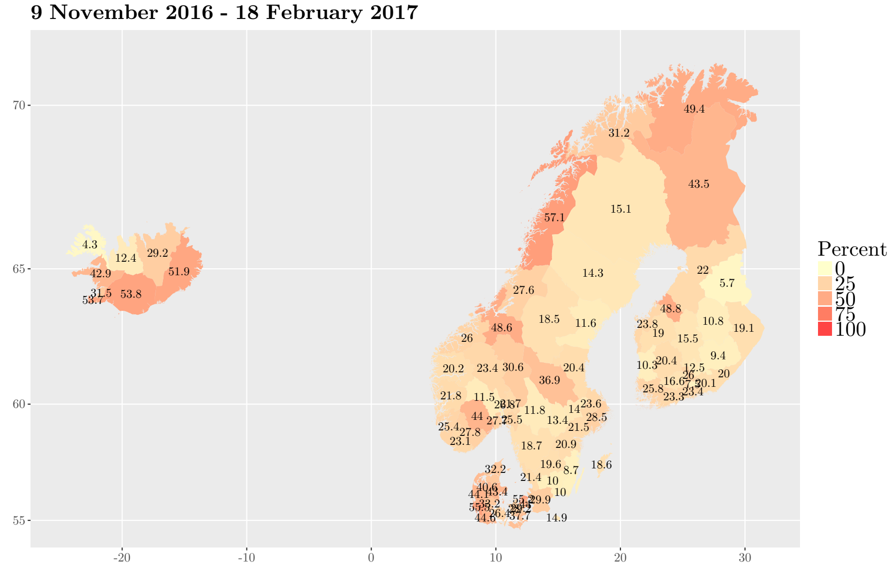


Figure 1: Percent of gendered tweets in English.

Males use the national language on Twitter more than females do in all five Nordic countries; females use English more in all countries except for Iceland (Table 2).

Table 2: Percentage of tweets by country, gender and language.

		Nat. Lang.	English	Other
Iceland	males	74.6	14.0	11.4
	females	74.0	13.3	12.7
Norway	males	46.0	24.1	29.9
	females	38.9	32.8	28.3
Denmark	males	45.8	37.6	16.6
	females	27.5	47.2	25.3
Sweden	males	58.8	22.9	18.3
	females	55.4	24.0	20.6
Finland	males	64.2	21.4	14.4
	females	61.4	24.5	14.1

The difference is most pronounced for Denmark and Norway, and less pronounced for Sweden, Finland, and Iceland. The differences in English use by gender were significant at $p < 0.05$ for all countries but Iceland (Fisher's Exact Test).[13]

3.2 Relationships among grammatical features, country and gender

Thirty-eight of the PoS tags were applied at least once in all of the ten gendered subcorpora. For each subcorpus, the relative frequency of each tag per 1,000 tokens was calculated (Table 3).

Table 3: Frequencies of grammatical features per 1,000 tokens.

	Iceland		Norway		Denmark		Sweden		Finland	
	m	f	m	f	m	f	m	f	m	f
Left bracket (()	1.03	1.22	1.59	1.03	1.85	1.18	1.64	1.41	2.07	1.16
Right bracket ())	1.09	1.03	1.47	0.88	1.74	1.16	1.59	1.34	2.25	1.07
Comma	16.87	12.41	21.81	14.66	19.72	15.91	24.25	16.68	20.25	16.97
Other punctuation (: ; ... + - = < > [ ])	19.47	30.56	20.77	17.18	26.02	20	17.89	19.14	27.6	20.87
Sentence-ending punctuation (. ? !)	57.56	49.09	55.96	49.41	54.12	44.31	66.26	54.88	56.75	52.06
Quotation marks (»)	8.77	5.92	7.85	6.56	7.29	6.74	8.83	8.54	9.26	7.22
Coordinating conjunction	17.9	17.59	18.19	19.27	19.57	20.34	20.82	21.26	19.41	21.8
Number	13.3	10.72	14.29	9.52	13.21	10.21	13.71	11.49	15.77	11.44
Determiner	65.97	62.44	61.75	67.12	60.24	53.43	63.68	60.34	54.53	53.84
Existential *there*	0.42	0.38	0.48	0.34	0.43	0.34	0.45	0.39	0.62	0.52
Foreign word	0.06	0.09	0.03	0	0.03	0.02	0.04	0.02	0.06	0.03
Hashtag	36.28	59.71	39.56	34.39	36.98	38.74	32.69	34.59	61.26	59.39
Preposition or subordinating conjunction	73.23	72.79	76.78	55.47	78.25	65.39	76.68	70.42	75.73	69.39
Adjective	50.85	42.13	48.07	65.93	50.99	50.4	53.19	52.86	52.75	52.72
Comparative adjective	1.75	1.5	1.73	1.18	1.83	1.4	1.83	1.52	1.82	1.77

[13] Iceland: p = 0.188, odds ratio = 0.94; Norway: p < 2.2e–16, odds ratio = 1.54; Denmark: p < 2.2e–16, odds ratio = 1.48; Sweden: p = 1.05e–16, odds ratio = 1.06; Finland: p < 2.2e–16, odds ratio = 1.19.

	Iceland		Norway		Denmark		Sweden		Finland	
	m	f	m	f	m	f	m	f	m	f
Superlative adjective	3.57	3.01	2.4	1.72	2.27	2.4	2.5	2.59	2.46	2.77
Modal verb	11.19	8.65	9.77	7.27	10.93	10.32	11.88	9.98	8.92	9.41
Noun, singular or mass	118.51	109.55	109.15	119.79	114.41	99.38	109.84	108.27	112.37	105.37
Proper noun	74.5	64.23	80.85	85.14	76.04	55.15	64.91	64.46	74.76	56.95
Plural noun	29.08	25.3	28.04	35.65	29.33	23.45	33.04	27.66	31.34	27.73
Personal pronoun	59.26	60.28	50.62	53.61	55.41	80.04	63.16	72.71	44.03	68.76
Possessive pronoun	14.21	16.36	10.83	13.13	12.13	15.98	11.89	17.87	9.86	14.46
Adverb	42.27	35.55	39.61	37.5	43.39	48.44	48.44	47.17	39.45	49.53
Comparative adverb	2.12	1.5	1.4	1.06	1.61	1.18	1.58	1.4	1.39	1.41
Phrasal particle	4.41	4.61	4.3	4.1	4.17	4.04	4.23	4.09	3.26	3.28
Retweet	0.06	0.09	0.3	0.13	0.09	0.1	0.06	0.2	0.07	0.22
to	15.72	17.02	16.95	14.01	17.15	17.11	18.06	17.16	18.92	17.86
Interjection/ emoticon/ emoji	30.29	60.65	36.51	70.44	35.08	63.5	25.08	46.45	28.34	43.34
URL	34.95	47.49	29.03	29.91	31.45	29.91	28.34	31.61	37.1	33.97
Username (preceded by @)	55.15	41	79.08	54.51	58.55	75.72	49.35	45.81	59.27	53.44
Verb, base form	40.15	38.65	36.05	31.81	38.89	38.94	40.64	42.69	34.28	38.75
Verb, past tense	17.96	15.23	17.53	20.12	16.64	19.13	18.24	17.59	16.08	18.01
Verb, gerund or present particle	18.68	17.59	16.92	16.58	18.36	18.3	16.8	18.48	18.36	18.89
Verb, past participle	5.5	7.24	6.89	4.67	7.79	6.03	8.18	6.79	7.2	6.25
Verb, non-3rd person singular present	26.79	26.52	23.36	34.09	24.67	32.94	28.3	31.42	21.11	28.93
Verb, 3rd person singular present	20.5	19	19.72	13.51	19.73	17.61	20.62	19.18	21.15	19.29
Wh-determiner	0.67	0.38	0.58	0.41	0.62	0.53	0.84	0.7	0.69	0.71
Wh-pronoun	4.54	4.89	4.26	2.92	3.57	4.02	4.34	4.24	3.88	4.23
Wh-adverb	5.32	7.43	5.47	4.96	5.33	6.19	6.09	6.57	5.5	6.11

While the distributions of feature frequencies for frequent features such as pronouns or verbal forms approach normality, infrequent features such as Wh-determiners are not normally distributed in the data. Thus, to determine whether differences in feature use by gender exist, Mann-Whitney U tests were conducted for each feature on the basis of the mean standardized values for males and for females in the gendered subcorpora. Of the 39 features, eleven exhibited significant ($p < 0.05$) differences in use between males and females: Right brackets, commas, sentence-ending punctuation, quotation marks, numbers/ numerals, prepositions or subordinating conjunctions, comparative adjectives, and 3rd-person singular present verb forms were significantly more likely to be utilized by males, while possessive pronouns, interjections/emoticons/emoji, and non-3rd-person singular present verb forms were significantly more likely to be used by females (Table 4).

Table 4: Grammatical features by gender.

	Feature	Gender	p-value		Feature	Gender	p-value
1	Left bracket (()	m	0.151	21	Personal pronoun	f	0.095
2	**Right bracket ())**	**m**	**0.032**	22	**Possessive pronoun**	**f**	**0.016**
3	**Comma**	**m**	**0.016**	23	Adverb	f	1.000
4	Other punctuation (: ; … + - = < > [ ])	m	0.841	24	Comparative adverb	m	0.151
5	**Sentence-ending punctuation (. ? !)**	**m**	**0.016**	25	Phrasal particle	m	0.548
6	**Quotation marks (»)**	**m**	**0.032**	26	Retweet	f	0.151
7	Coordinating conjunction	f	0.548	27	*to*	m	0.690
8	**Number**	**m**	**0.008**	28	**Interjection/ emoticon/emoji**	**f**	**0.008**
9	Determiner	m	0.690	29	URL	f	0.690
10	Existential *there*	m	0.095	30	Username (preceded by @)	m	0.222
11	Foreign word	m	0.310	31	Verb, base form	f	1.000
12	Hashtag	f	1.000	32	Verb, past tense	f	0.421
13	**Preposition or subordinating conjunction**	**m**	**0.008**	33	Verb, gerund or present particle	f	1.000
14	Adjective	f	1.000	34	Verb, past participle	m	0.222
15	**Comparative adjective**	**m**	**0.032**	35	**Verb, non-3rd person singular present**	**f**	**0.032**
16	Superlative adjective	m	0.841	36	**Verb, 3rd person singular present**	**m**	**0.008**
17	Modal verb	m	0.151	37	Wh-determiner	m	0.421
18	Noun, singular or mass	m	0.222	38	Wh-pronoun	m	0.841
19	Proper noun	m	0.151	39	Wh-adverb	f	0.151
20	Plural noun	m	0.151				

Significant differences by gender at p < 0.05 for features in bold (Mann-Whitney U test)

A t-test of population means conducted on the same data gave similar results (Figure 2).

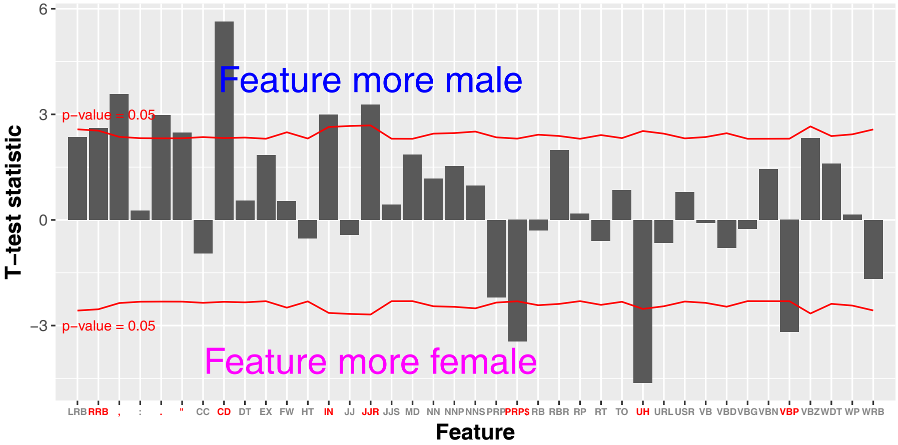


Figure 2: Grammatical features by gender (t-test).

Gendered differences were also considered by country and feature on the basis of the aggregate feature frequencies per unique user in the data. While differences in sample size make the results of t-tests for infrequent and non-normally-distributed features somewhat unreliable, particularly for Iceland due to the small number of users in the sample, many of the differences in feature frequencies between males and females were found for most or all of the Nordic countries.

3.3 Principal components analysis

In order to explore the underlying patterning of the variance in the data, a principal components analysis was conducted on a covariance matrix of the normalized frequencies of the 39 variables for the ten English subcorpora (the male and female subcorpora for each of the five Nordic countries). The first two components capture 58.21% of the variance in the data. The strongest loadings (> 0.2) on the first two components are shown in Table 5.

Table 5: Loadings > 0.2 on first two principal components.

Feature	PC 1	PC 2
Interjection/emoticon/emoji	0.76	-0.27
Personal pronoun	0.36	0.32
Proper noun	-0.22	-0.54
Sentence-ending punctuation	-0.25	
Preposition	-0.27	0.20
Hashtag		0.41
Noun		-0.26
Adjective		-0.25
Determiner		-0.20

The strongest positive loadings on the first principal component are for two features with interpersonal interaction and stance orientation functions: Interjections/emoticons/emoji and the use of personal pronouns. Negative loadings are associated with features that typically relate to the presentation of information (proper nouns) and the organization of discourse (sentence-ending punctuation and prepositions).

The second principal component also shows a positive loading for personal pronouns and a negative loading (somewhat greater in magnitude than for the first component) on proper nouns, but positive loadings for prepositions and hashtags and negative loadings for nouns, adjectives, and determiners. Tokens tagged as interjections have a negative loading on the second principal component.

Both principal components seem to index interactive discourse, but with somewhat different focuses. It may be the case that the first principal component captures affect expression and stance orientation (for example, in tweets expressing affective content that include emoticons or emojis), while the second principal component may capture interactions that make reference to discourse external to the tweet messages themselves, such as through the use of hashtags.

The positions of the gendered subcorpora along the first two principal components are shown in Figure 3. The analysis shows clear functional separation between males and females along the first principal component: The male subcorpora all have negative values, while the female subcorpora have positive values. Gender separation along the second principal component is less distinct. Although the female subcorpora from Iceland, Denmark, Sweden and Finland exhibit higher values than the male subcorpora, the Norwegian female subcorpus is an outlier, with a negative value much lower than any those for the male subcorpora. An examination of the data reveals that the values for Norwegian females are strongly influenced by the extremely high Twitter activity of a single author whose posts tend to consist mainly of sequences of hashtags.

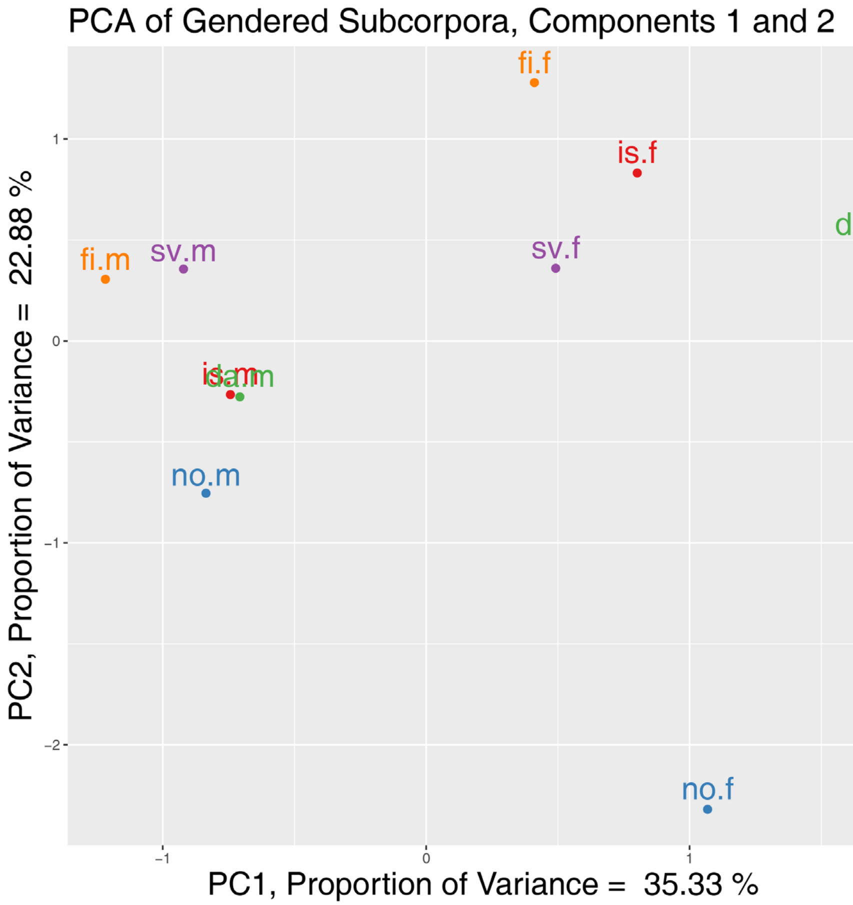


Figure 3: Loadings on components 1 and 2 of PCA for English subcorpora.

The distance between male and female subcorpora for the same country are also notable, and the Euclidean distance for the first two principal components is comparable for the individual Nordic countries. The genders are closer in Sweden and Finland and somewhat further apart in Iceland, Denmark, and Norway.

Component scores for the gendered subcorpora were calculated by summing the scaled frequencies (expressed in terms of standard deviation distance from the mean value for all ten subcorpora) of those components with weights > 0.2 on the first two components (see Biber 1988: 93—97).

Table 6: Component Scores for PC 1 and PC2.

		PC 1	PC 2
Iceland	male	6.19	11.28
	female	6.83	12.48
Norway	male	6.37	11.02
	female	6.57	12.25
Denmark	male	6.42	11.37
	female	6.89	11.51
Sweden	male	6.42	10.86
	female	7.00	11.81
Finland	male	5.74	11.33
	female	6.35	11.93

Here as well, a modest but clear functional separation is observable in the differences between male and female scores.

4 CONCLUSION

Corpora consisting of messages in English posted online collected from social media sites such as Twitter can shed light on the ways in which English continues to develop and diversify globally, especially in contexts where it has not traditionally been a language of daily communication. Data that has been appended metadata tags for location and disambiguated for author gender can provide insight into global English varieties and the relationships between language and gender in different geographical and social contexts.

While it is not surprising that English is extensively used on a global internet platform such as Twitter, the present research confirms high rates of use of English on Twitter in the Nordic countries attested in previous research. Overall, people in Denmark and Norway send more tweets in English than do those in Iceland, Sweden and Finland, and females more than males. It may be the case that the proportion of messages from the Nordic countries written in English on Twitter is increasing over time: For example, Mocanu et al. (2013) report rates of use for English in the Nordics in GPS-enabled tweets collected from 2010—2012. They find Iceland has 45%, Norway 24.6%, Denmark 40%, Sweden 18.1%, and Finland 27.1% English tweets.[14] This study finds similar values (slightly higher for Norway, Denmark and Sweden; slightly lower for Iceland and Finland), but considers not only GPS-tagged tweets (i.e. those with a populated *geo* field) but also those with a

[14] http://www.twitterofbabel.org/

place value. Considering the fact that GPS-tagged tweets are typically sent on smartphones by users who are, on average, younger than the overall population and tend to use more English (see Pavalanathan and Eisenstein 2015), the data from the present study suggests and increase in English use in the Nordics over the past six years.

The results of the gender analysis in the present work complement those from previous corpus studies on English-language data collected from CMC or Twitter in Anglophone societies such as the United States: Females tend to use features such as personal pronouns, possessive pronouns or affect markers more often than males, whereas males use features such as punctuation, numbers/numerals, and nouns more than do females (Bamann, Eisenstein and Schnoebelen 2014). The same general pattern can be found in the present data set for English used on Twitter in the Nordic countries by persons with common Nordic names.

Multidimensional approaches based on factor analysis or principal components analysis have shown that differences in aggregate grammatical feature frequencies for national varieties of English can be interpreted in terms of communicative or discourse-functional dimensions (Biber 1988; 1995; Xiao 2009). The Nordic Twitter data used in this study was induced to reflect author gender, and the results show differentiation by gender along a first principal component, explaining a large proportion of variance in the data. The loadings on this component correspond to grammatical features whose discourse or communicative functions may contrast interactive stance orientation and affective content with informational and discourse organization functions – a finding comparable to the proposed "involved versus informational production" dimension found by Biber in a corpus of print media texts (1988: 107).

Although most work on differences in feature frequencies by gender has been conducted on L1 English data, there is some evidence for differential use of word classes by gender in other languages as well.[15] This study shows that gender-based differences in feature frequency in Twitter data from the Nordics matches up well with differences found in CMC and non-CMC data from Anglophone and non-Anglophone contexts.

It has been suggested that the small differences in aggregate Anglophone and non-Anglophone feature frequencies between males and females may reflect different orientations towards the use of communicative or discourse functions for the negotiation of affect maintenance or solidarity (Holmes 1998). Exploratory data analysis suggests that functional separation of English-language feature frequencies by gender can be observed for Nordic Twitter corpora with induced author gender. This tentative confirmation of some of the trends observed in CMC and Twitter data from L1 Anglophone contexts raises interesting questions as to

[15] For French, see Schenk-van Witsen (1981). For French, Turkish, Indonesian and Japanese, see Ciot, Sonderegger and Ruths (2013).

the possible causes: Have cultural attitudes found in Anglophone contexts such as the United States been transmitted through the internet and other media to Northern Europe and become manifest in the patterning of grammatical features by Nordic people using English? Or is it the case that there may be underlying differences in interaction and communication style between the genders that are rooted not in cultural specifics, but aspects of human biology?

One interesting prospect for future investigation could thus be to investigate the extent to which the gender differentiation in grammatical type frequencies found in English-language data are also present in language data in the Nordic languages. Another possibility for future research, suggested by the presence of metadata fields in tweets that indicate direct responses to others, would be to combine aggregate feature frequency information by gender with user network information in order to gauge the relative contribution of each to differences in language. As English continues to evolve in diverse geographical as well as ever-more specialized technological contexts of CMC, the investigation of the relationship between language use and factors of demographic identity such as gender will continue to provide insights into our shared experience.

References

Allwood, Jens, 1998: Some frequency based differences between spoken and written Swedish. *Proceedings from the XVI:th Scandinavian conference of linguistics*. Turku, Finland. Department of Linguistics, University of Turku. http://sskkii.gu.se/jens/publications/docs076-100/084.pdf. (Last accessed 1 March 2017.)

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker and Jonathan Schler, 2007: Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, 12/9. http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2003/1878. (Last accessed 1 March 2017.)

Bamann, David, Jacob Eisenstein and Tyler Schnoebelen, 2014: Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18/2. 135–160. http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/full. (Last accessed 1 March 2017.)

Baron, Naomi S., 2004: See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology*, 23/4. 397–423.

Biber, Douglas, 1988: *Variation across speech and writing*. Cambridge University Press: Cambridge, UK.

Biber, Douglas, 1995: *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press: Cambridge, UK.

Blommaert, Jan, 2012: Supervernaculars and their dialects. *Dutch Journal of Applied Linguistics*, 1/1. 1–14.

Bolton, Kingsley and Christiane Meierkord, 2013: English in contemporary Sweden: Perceptions, policies, and narrated practices. *Journal of Sociolinguistics* 17. 93–117.

Ciot, Morgane, Morgan Sonderegger and Derek Ruths, 2013: Gender inference of Twitter users in non-English contexts. *Proceedings of the 2013 conference on empirical methods in natural language processing*. Stroudsburg, PA: Association for Computational Linguistics. 1136–1145. http://www.aclweb.org/anthology/D13-1114. (Last accessed 1 March 2017.)

Coats, Steven, 2016: Grammatical feature frequencies of English on Twitter in Finland. Squires, Lauren (ed.): *English in Computer-mediated Communication: Variation, Representation, and Change.* Berlin: De Gruyter. 179–210. https://doi.org/10.1515/9783110490817-009. (Last accessed 1 March 2017.)

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing, 2014: Diffusion of lexical change in social media. *PLoS ONE* 9/1. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0113114. (Last accessed 1 March 2017.)

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan and Noah A. Smith, 2011: Part-of-speech tagging for Twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual meeting of the association for computational linguistics: human language technologies*. Stroudsburg, PA: Association for Computational Linguistics. 42–47. www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf. (Last accessed 1 March 2017.)

Görlach, Manfred, 2002: *Still more Englishes*. Amsterdam: John Benjamins.

Gustafson-Capková, Sofia and Britt Hartmann, 2008: *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf. (Last accessed 1 March 2017.)

Herring, Susan and John Paolillo, 2006: Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10/4. 439–459.

Holmes, Janet, 1998: Women's talk: The question of sociolinguistic universals. *Australian Journal of Communications* 20. 125–149.

Leetaru, Kalev H., Shaowen Wang, Guofeng Cao, Anand Padmanabhan and Eric Shook, 2013: Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18/5–6. http://firstmonday.org/article/view/4366/3654. (Last accessed 1 March 2017.)

Leppänen, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Räisänen, Mikko Laitinen, Heidi Koskela, Salla Lähdesmäki and Henna Jousmäki, 2011: National Survey on the English Language in Finland: Uses, meanings and attitudes. *Studies in Variation, Contacts and Change in English* 5. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/05/evarieng-vol5.pdf. (Last accessed 1 March 2017.)

Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini, 1993: Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19/2. 313–330. http://dl.acm.org/citation.cfm?id=972475. (Last accessed 1 March 2017.)

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J. Niels Rosenquist, 2011: Understanding the demographics of Twitter users. *Proceedings of the fifth international AAAI conference on weblogs and social media*. Menlo Park, CA: AAAI. 554–557. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234. (Last accessed 1 March 2017.)

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang and Alessandro Vespignani, 2013: The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE* 8/4. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981. (Last accessed 1 March 2017)

Newman, Matthew L., Carla J. Groom, Lori D. Handelman and James W. Pennebaker, 2008: Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45/3. 211–236. http://dx.doi.org/10.1080/01638530802073712. (Last accessed 1 March 2017)

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneid and Noah A. Smith, 2013: Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of NAACL-HLT*. Stroudsburg, PA: Association for Computational Linguistics. 380–390. http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.naacl13.pdf. (Last accessed 1 March 2017.)

Page, Ruth, 2012: The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication* 6/2. 181–201.

Pavalanathan, Umashanthi and Jacob Eisenstein, 2015: Confounds and consequences in geotagged Twitter data. http://arxiv.org/pdf/1506.02275v2.pdf. (Last accessed 1 March 2017,)

Pennacchiotti, Marco and Ana-Maria Popescu, 2011: A machine learning approach to Twitter user classification. *Proceedings of the fifth international AAAI conference on weblogs and social media*. Menlo Park, CA: Association for the Advancement of Artificial Intelligence. 281–288. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2886/3262. (Last accessed 1 March 2017.)

Rao, Delip, David Yarowsky, Abhishek Shreevats and Manaswi Gupta, 2010: Classifying latent user attributes in Twitter. *Proceedings of the 2nd international workshop on search and mining user-generated contents*. New York, NY: Association for Computing Machinery. 37–44. http://dl.acm.org/citation.cfm?doid=1871985.1871993. (Last accessed 1 March 2017.)

Roesslein, Josh, 2015. *Tweepy*. Python programming language module. http://www.tweepy.org. (Last accessed 1 March 2017)

Schenk-van Witsen, Rosalien, 1981. Les différences sexuelles dans le français parlé: Une étude-pilote des différences lexicales entre hommes et femmes. *Langage et Societé*, 17/1. 59–78. http://www.persee.fr/doc/lsoc_0181-4095_1981_num_17_1_1328. (Last accessed 1 March 2017.)

Squires, Lauren, 2015: Twitter: Design, discourse, and implications of public text. Georgakopoulou, Alexandra and Tereza Spilioti (eds.): *The Routledge Handbook of Language and Digital Communication*. London: Routledge. 239–256.

Vandergriff, Ilona, 2013: Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics* 51. 1–12. http://www.sciencedirect.com/science/article/pii/S037821661300057X. (Last accessed 1 March 2017.)

Xiao, Richard, 2009: Multidimensional analysis and the study of world Englishes. *World Englishes* 28/4. 421–450.

Zappavigna, Michele, 2011: Ambient affiliation: A linguistic perspective on Twitter. *New Media and Society* 13/5. 788–806.

Part 3
Conversation and conflict in CMC

Conversations on Twitter

Tatjana Scheffler, University of Potsdam

Abstract

In this paper, we analyse the linguistic structure of a corpus of German conversations on Twitter. Near real-time conversations conducted on social media are interesting from a linguistic viewpoint, because they show features of informal, spoken dialog while being transmitted asynchronously and in the written mode. The current study focuses on models of dialog structure developed for spoken conversations and their applicability to conversations on Twitter. We show that many well-known dialog phenomena can be observed in Twitter conversations, such as the use of particles, questions, turn-taking, informal lexical choice, corrections and fillers. At the same time, speakers on social media also frequently avail themselves of more formal, written-like options, and some spoken-like features take on new meanings in social media. Our approach allows for sub-dividing the conversations into three different types based on their structure, since a single medium such as Twitter combines several subgenres, such as chats among friends, surveys, customer-service dialogs, and so on. We distinguish broadcasts from linear conversations and group discussions.

Keywords: dialog, Twitter, social media, conversation structure, German

1 INTRODUCTION

In this paper we investigate German Twitter conversations. We identify properties of the structure of Twitter conversations and look specifically for phenomena typical of informal spoken conversations. We find that many features of spoken conversations are found equally in our Twitter corpus. However, there are also some differences that open interesting avenues for future work, such as a novel way of marking clarification requests, and idiosyncrasies in the use of discourse particles.

It is a defining feature of social media that they allow for interaction among their users. As opposed to traditional written (news) media, text is not only produced by a few and consumed by many, but instead linguistic data is produced and consumed near-simultaneously by many speakers.[1] Even though all "social" media enable conversations in this way, different channels can be distinguished by their interactive properties, as detailed in Table 1. Of the existing media with a mainly textual basis, Twitter is among the most conversational in nature. This paper studies the conversation structure of German Twitter data, in order to pin down the commonalities and differences of such computer-mediated conversations with spoken dialogs.

The paper makes three contributions. First, in Section 3, we detail our method for extracting conversations from Twitter and give an overview of the resulting corpus, a dataset of over 2.5 million threads (each between two and several hundred tweets). In Section 4, we analyse the dialog structure of the extracted Twitter threads and show structural measures to identify different types of conversations: broadcasts, group discussions, and linear conversations. In Section 5, we address several linguistic phenomena that are said to be typical of spoken conversations, in order to get a closer view of the linguistic properties of Twitter conversations. The careful comparison of "spoken" phenomena occurring in different social media allows us to tease apart the effects of the mode (spoken vs. written), interactional vs. informational style (Storrer 2013), informal vs. formal relations between speaker and hearer, binary interaction vs. multilog, etc. We find that some features of spontaneous interaction, for example questions, including clarification questions, occur frequently in the Twitter dialogs. On the other hand, while some modal particles are more frequent in the Twitter conversations than in monological text, this is not as pronounced overall. We argue that different social media with their specific configurations allow us to further study which property of a linguistic context licenses which types of expression.

[1] Though social media content is produced in writing, in this paper we use the terms 'speaker' and 'hearer' loosely to refer to the producers and addressees of utterances.

In order to enable comparison across different types of media, we focus here on linguistic phenomena that differentiate between spoken conversations and written text, and we exclude novel features specific to social media channels, such as emoticons, inflectives, across the board capitalization, etc. Though those social media innovations are important objects of linguistic study, we are more interested in the following research questions: Which characteristics typical of free spoken interactions carry over to social media conversations (on Twitter)? Which differences in frequency, use and meaning do we find between the modes, and how can this be explained?

2 BACKGROUND

In this paper, we study Twitter conversations from the perspective of the conceptual orality continuum (Koch and Oesterreicher 1985), comparing the medium to typical spoken or written data. In particular, we analyse to what extent the *dialog structure* of social media (Twitter) corresponds to what is known about spoken conversations. In this section, we address both lines of previous research in turn.

2.1 Characteristics of Spoken Dialogs

Herbert H. Clark and colleagues have established a view of conversations as a specific kind of linguistic communication in linguistics and psychology (Clark and Schaefer 1987, Clark and Schaefer 1989). From this perspective, conversations are not merely sentences uttered by different people in turn, but must be viewed as joint actions (like a hand-shake) of several participants (simultaneously speakers and hearers). Previous research shows how speakers and hearers coordinate across a conversation to achieve their common communicative goals. In prototypical face-to-face conversations, all participants are furthermore on equal footing (as opposed to, say, a radio interview, where one participant leads the conversation) with regard to access to and position in the dialog. Conversations are situated in a physical context and unfold in real-time, typically in spoken form. They are characterized by phenomena representative of spontaneous speech, such as clarification requests, corrections, fillers, pauses, and the like. This line of research is based on the analysis of natural conversations, either in person or over the telephone.

This work shows that contributions in dialog must be *grounded*, i.e. acknowledged and accepted by the conversation participants, in order to advance the discourse. Thus, unlike in written monolog, each contribution in spoken conversations consists of two phases, a *presentation* and an *acceptance* phase, where the presentation is done by the speaker and the acceptance must be taken over by the hearer (Clark and Schaefer 1989). If there are no problems, the acceptance of a dialog contribution is signalled by the hearer. When problems of understanding occur, these are signalled by one of the conversation participants and clarification requests and/or corrections may follow. In the easiest case, the hearer in a dialog signals understanding by choosing an appropriate, relevant following contribution. Since what is a "relevant next contribution" has been conventionalized in many cases, we find that dialog contributions can be well characterized by *adjacency pairs* (Clark and Schaefer 1989: 271), which are pairs of speech acts that often occur together in dialogs. The first part of the adjacency pair is the initiating act (for example, a question), while the second item in the pair provides the expected relevant reply (e.g., an answer).

Since the *kinds* of contributions made in a dialog are so important to characterize the conversation, dialog researchers have focused on the notion of *dialog acts*, an extension of the idea of speech acts (Austin 1975), but adapted to cover all possible linguistic contributions in dialog. The dialog act carried out by an utterance is the communicative function of that utterance, independent of the actual semantic content. Examples of dialog acts are Inform, Thank or Promise. The dialog acts that can be found in conversation depend on the type of conversations, and many different dialog act taxonomies exist, several of which have been used for extensive annotation studies of dialog acts in naturally occurring spoken conversation (Core and Allen 1997, Bunt et al. 2010).

Finally, it was noted early on in the literature that, because of the setting discussed above, spoken conversations typically contain specific linguistic features that are largely missing from written text, such as corrections, fillers and discourse particles. When contributions are not successful, this can be detected and rectified relatively quickly in conversation. Speakers use specialized markers to indicate the detection of communicative problems (mis- and non-understanding) and corrections of their own speech or the interlocutor's contributions. Fillers and particles are used to contribute non-truth conditional content in speech, in addition and in parallel to the at-issue meaning of the individual contributions. These items are said to be largely absent in written language, due to editing, planning, and genre restrictions (Rudolph 1991).

2.2 Spoken versus written media and CMC

It is clear that social media in general fall somewhere in between the prototypical poles of spontaneous spoken conversation and formal written text (Koch and

Oesterreicher 1985). But research points to the fact that conceptual orality cannot be captured as just one parameter on a continuous line, and that various linguistic phenomena reflect different aspects of speech-like linguistic contributions. For example, register studies following Biber (1993) distinguish several dimensions on which conversations and newspaper text differ: the informational/interactive dimension, the non-/narrative dimension, and so on. Each text type can then be situated along each of these dimensions, and the various forms of social media do not necessarily all group together. It is therefore interesting to study different types of social media, because it may allow us to distinguish which aspects of the context linguistic phenomena are facilitated or constrained by: e.g., informal style, interactive situation, real-world situatedness, synchronicity, etc.

German computer mediated communication has been the focus of several previous studies. Here, we only mention a few that touch upon the issues mentioned above. Beißwenger (2007) compares chats to spoken conversations, discussing the question of medial vs. conceptual orality, turn-taking, as well as the extra-linguistic action of deleting a drafted post. Chats closely resemble Twitter conversations, in that they are near real-time computer-mediated interactions (though some differences remain). In related work, Storrer (2013) investigates the conceptual orality continuum with regard to several computer mediated text types, and claims that the distinction between interactional and presentational writing is central in this context. This dimension distinguishes, for example, published Wikipedia articles (presentational) from the corresponding discussion pages (interactional). She points out that language adapts to the intended audience and topic and identifies differences in contribution lengths, and the use of computer mediated communication (CMC) specific items such as action words and emoticons. Similarly, (Storrer 2014) points out that there are large differences in language use within a medium based on the interactional style and the distance between speaker and hearer. A CMC medium cannot be considered a monolithic genre. Other studies identify linguistic phenomena that are specific to CMC (in German), or distinguish texts in these media from those in others (i.e. traditional newspaper texts) (Beißwenger 2013). Bartz et al. (2013) introduce a typology of such phenomena (across-the-board capitalization, emoticons, etc.) for use in the annotation of German CMC corpora. However, apart from colloquialisms, these items are not the focus of the current study. Here, we concentrate not on novel linguistic phenomena specific to social media, but on those features of spoken discourse that may also be found in the discourse carried out in Twitter conversations.

In this paper, we consider specifically the question of to what extent the dialog models that were developed for spoken conversations are applicable to written conversations on Twitter. We chose Twitter because its setting is most similar to spoken conversations among the major social media. Table 1 summarizes the main context properties of the linguistic contributions on the major social media platforms. All computer-mediated communications are available in written form. But while blogs are certainly written with a reader in mind, the production of blog posts does not in itself require a reader to be successful. Writing a blog is thus an individual action of a speaker, and while certainly informal, typically not interactional in nature. In contrast, forums, Facebook posts and tweets are more interactive in that they (at least in many cases) require an acceptance phase in Clark and Schaefer's (1989) use of the term, and thus constitute a joint action. These media also typically allow more than two participants in a conversation. There is a difference between blogs and Facebook on the one hand, and forums and Twitter on the other, in that the latter are common platforms where users interact, whereas in the former the platform (blog, Facebook page) belongs to one privileged user and the others are merely invited to "comment" on this page, yielding a power differential.

Table 1: Interactive properties of a range of social media.

Property	Spoken	Blogs	Forums	Facebook	Twitter
mode	spoken	written	written	written	written
action	joint	individual	joint	joint	joint
speakers	2+	mainly one	many	many	many
ownership	common	single	common	single	common
partic. status	equal	unequal	equal	unequal	equal
timing	synchronous	asynch.	asynch.	near-synch.	near-synch.
planning	little	much	medium	little	little
situatedness	situated	online	online	online	online

Further, the technical set-up and the way the media are consumed cause a difference in the timing of contributions and the amount of planning that goes into them. Spoken conversations happen in real time, speakers and hearers are synchronously active. As a result, there is very little time for planning utterances beforehand, and thus they are spontaneous in style. Even though writers on Facebook and Twitter are in principle able to access utterances later on, since they are written and remain on the platform, most conversations happen in near-real time. Individual utterances become unavailable quickly as they are "swamped out" of the timeline by new status updates from other users, especially on Twitter. In contrast, interactions on blogs and forums are centred around a topic of common interest, and span much longer time periods (as interlocutors return to the blog/forum to discuss topics of interest). It follows that these media allow more time for planning and editing contributions, with less pressure on timely responses. Finally, all social media differ from face-to-face conversations in that the latter are situated in a physical context that is the basis of grounding, and which can be referenced in the contributions. Instead, all social media are somewhat removed from any physical or often even previous social context of the interlocutors (the exception being private Facebook walls, where the conversation participants are usually known to each other). This can have effects on the linguistic means that must be chosen to make reference to people and events, and on the management of so-called common ground (Stalnaker 1978).

3 CONSTRUCTING A CORPUS OF TWITTER CONVERSATIONS

The overall communicative settings detailed in Table 1 show that, among the considered social media, Twitter is closest to conversational speech because it consists (at least in part) of conversations in near-real time, between two or more participants, who come together on an equal footing to jointly fulfil a communicative function. There are two main differences between spoken conversations and those on Twitter: the first is the spoken vs. written mode, and the second is the fact that face-to-face conversations are situated in a physical and social context, so that speakers can make reference to prior knowledge of the hearers or to objects and events that are easily inferable or apparent in the physical surroundings.

Twitter is a medium that allows users to post short "status messages". Its contributors are private citizens, public institutions, and businesses, as well as bots that automatically post informational content, advertising, or jokes and memes. Since we are interested in the linguistic features exhibited on social media, with a focus on dialog, we would like to specifically extract tweets that are written by individuals (excluding for example press statements by organizations and companies as much as possible, as well as all tweets by bots), and that are part of larger conversations.

Unfortunately, Twitter's API[2] does not make the extraction of entire conversations possible, and thus there has been limited computational linguistic research into Twitter conversations. In some cases, researchers have determined a set of users of interest and extracted all tweets by these, as well as by all their contacts (Ritter et al. 2010). This enables the reconstruction of conversations, including these seed users and some analyses. In this approach, the selection of users is crucial, and may restrict the general validity of any results. In contrast, we follow the approach proposed by Scheffler (2014) to construct a language-specific general Twitter corpus with a high recall, and then reconstruct all conversations contained in this general corpus. Since the Twitter API severely rate limits the

[2] https://dev.twitter.com/overview/api

number of tweets that can be extracted, this approach is only applicable to languages beyond the top five or so on Twitter: English, Spanish, Indonesian, Malay, and Japanese (Mocanu et al. 2013).

In the chosen approach, a stop word list of frequently occurring words in a language (in our case, German) is used to extract all tweets that contain these terms, using the Twitter API's *filter* keyword. The corpus examined in this work was created in April, 2013, using a precompiled stop word list for German with few manual corrections. The tweets are then filtered using the high-quality language identification module *langid*[3] (Lui and Baldwin 2012).[4]

The resulting dataset is estimated to contain > 90% of the German tweets sent during the time period. The conversation threads are reconstructed by following each tweet's *in-reply-to*-link in reverse (connecting a tweet to the one it was a reply to). This sorts all tweets into conversation threads. It must be noted that some threads may be incomplete for different reasons: (i) Tweets sent after the collection period are missing, even if they are in reply to existing conversations, because they were not included in the original dataset. (ii) A missing tweet somewhere within a conversation will lead to an erroneous split of the conversation into two subthreads. A tweet may be missing if it is not German, does not contain any of the stop words (e.g., is only a link), or was missed due to rate limiting by Twitter. In some cases, it is clear that a tweet is missing from the corpus because a subsequent tweet refers to it (by an *in-reply-to*-link). For those cases, we have attempted to re-fill the initial corpus by searching for these tweets specifically. This is a slow process due to rate limiting and not always successful, because users or tweets may have been deleted in the meantime.

The corpus was collected using the method described above from April 1–30, 2013, and is referred to as the "April13" corpus in the remainder of this work (Scheffler 2014). It contains 24,179,189 tweets from which we extracted 2,657,004 conversation threads (dialogs), consisting of 7,790,794 tweets, excluding the singletons. In this paper, we only consider conversations of at least length 2, i.e., that contain at least one reply in addition to the original tweet (we will call this the "TwitterDialogs," which is a new subcorpus studied for the first time in this paper). This restriction on conversations has the additional benefit of being a reliable filter for spam or automatic content. Typical bot tweets never receive any replies. To illustrate this effect, Table 2 shows the most frequent hashtags in

[3] https://github.com/saffsd/langid.py

[4] We have also created an improved stop word list for Twitter corpus extraction for German in collaboration with Nikolas Zoeller, FH Potsdam: We started with the 400 most frequent words in the large internet corpus deWaC , and manually removed a few obviously non-distinctively German words ('war', 'die'). We recorded all tweets retrieved using this list for two days (> 5 mio. tweets) and computed the ratio of German to non-German tweets using *langid* (confidence threshold: 0.85). A total of 27 words with a German/all-ratio < 0.2 were removed, to yield the final stop word list of 361 words. The list is available at https://github.com/TScheffler/TwitterCorpora.

the original April13 corpus compared with the most frequent hashtags in TwitterDialogs. The general corpus is dominated by automatic posts from mobile games (*#androidgames*, *#iphone*, etc.) and from other bots (*#pegelmv*, *#ostsee* originate with one bot posting water levels in the Baltic Sea). In contrast, the top ten hashtags used in dialogs reflect a few Twitter-specific items (*#ff* for "Follow Friday" recommendations, questions marked by *#followerpower*), but otherwise indicate important topics for discussions in the period and place when the data was collected: *#bvb* and *#fcb* denote popular soccer teams, *#piraten*, *#afd* and *#spd* are German political parties, *#tatort* is a popular TV crime show, and *#s21* and *#piratinnenkon* refer to prominent events during the collection time (a court investigation and a conference, respectively).

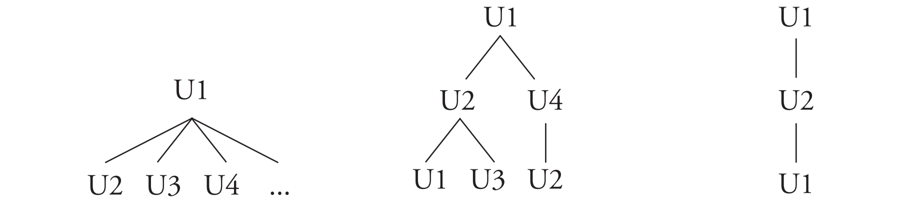
Table 2: Most frequent hashtags in the April13 and TwitterDialogs corpora.

April13	TwitterDialogs
#gameinsight	#ff
#android	#piraten
#androidgames	#bvb
#ipadgames	#afd
#ipad	#tatort
#pegelmv	#fcb
#ostsee	#spd
#iphone	#followerpower
#iphonegames	#s21
#news	#piratinnenkon

4 DIALOG STRUCTURE IN TWITTER

The resulting corpus includes (almost) all German Twitter threads during the sample month, but a closer look reveals that these are of different types. Visualizing the tree structure of these multilogs helps understand this. The tree structure of a conversation can be characterized by its size (the total number of tweets in the conversation), depth (defined as the length of the longest path from the root to a leaf, thus describing the longest conversation strand), and the number of users that take part in it. In some threads, one initial tweet receives hundreds of parallel answers, but no actual discussion ensues. This yields a conversation tree that is wide but whose depth is limited, possibly only to 2. We call those types of threads 'broadcasts,' since they often start with a statement by a (Twitter) celebrity which receives many responses from different people (see Figure 1(a)).

Note that this type of "conversation" cannot exist in face-to-face spoken dialog, since no contribution can receive hundreds of parallel replies. Linguistically, most broadcasts are very simple. An excerpt of a typical 'broadcast' thread is given in example (1). In this thread, 181 users reply to the 'Good morning, Germany' greeting by the actor Zach Braff, who has over 1.7 million followers.



(a) Broadcast; depth=2. (b) Group discussion; d=3. (c) Linear conversation; d=3.

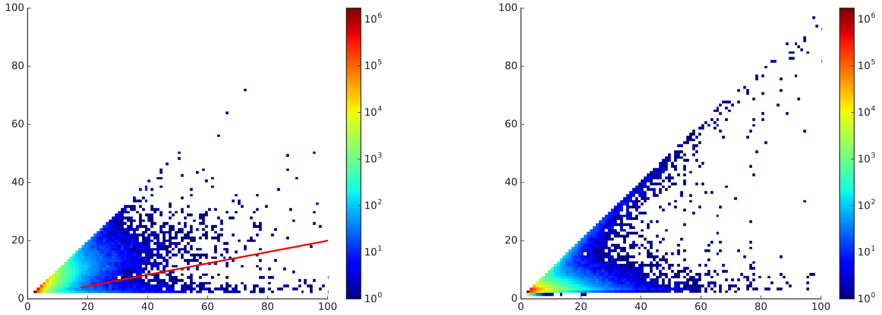
Figure 1: Three different kinds of tree structure for threads.

(1) Thread, size=182; maximum depth=2
@zachbraff: Guten Morgen Deutschland.
U2: @zachbraff oh ja, das ist gut!
U3: @zachbraff Guten Morgen, Zach Braff! Wie geht es Ihnen an diesem wunderschönen Tag?
U4: @zachbraff Guten Morgen mein süßes Schnitzel
U5: @zachbraff Guten Morgen Zach.
...[5]

Figure 2 shows 2D histograms of the size vs. depth and size vs. number of participants for all conversations in the corpus. In Figure 2, broadcast threads are along the x axis below the red line in plot (a), and along the diagonal in plot (b), which shows the number of distinct users that participated in each thread. Broadcast-type threads can have the properties of face-to-face conversations (such as question-answer pairs), but are unlike any spoken conversations in the number of participants (up to several hundred), and their short depth.

The second kind of threads on Twitter we call 'conversations.' If they are longer than 2 turns, their depth also increases, indicating that initial replies receive replies of their own, just like in spoken conversations. At the extreme (the diagonal in Figure 2(a)), the depth of the thread equals its size, so that the conversation consists entirely of a back-and-forth interchange between very few participants. In this case, the tree structure of the conversation is a linear chain, see Figure 1(c). Example (2) shows the start of an example linear conversation thread.

[5] @zachbraff: Good morning, Germany. U2: @zachbraff oh yeah, this is good! U3: @zachbraff Good morning, Zach Braff! How are you doing on this beautiful day? U4: @zachbraff Good morning my sweet dumpling. @zachbraff Good morning Zach.



(a) Size vs. depth of conversations (b) Size vs. number of users in conversations.

Figure 2: Multilog structure in Twitter conversations (excluding a few longer threads).

(2) Thread, size=103; maximum depth=28
U1: Kollers Klartext in den SN: "Es zahlt: Der Mittelstand". http://t.co/Tpu3fGH4Wx schade, dass er nicht häufiger twittert @U2
U3: @U1 @U2 Die Abschaffung der Kapitalertragssteuer erscheint mir aber weder zweckmäßig noch den Mittelstand entlastend.
U1: @U3 nicht?
…[6]

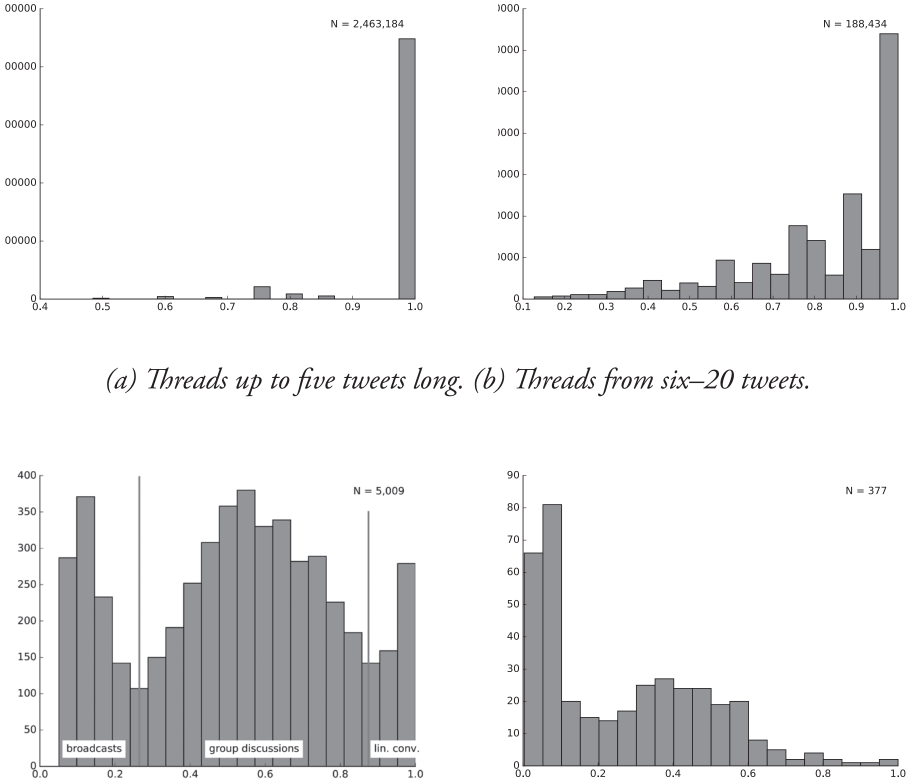
The diverse structure of threads becomes apparent when one analyses the angle of the vector pointing to the (x,y)-coordinates of each thread in the range of 0 to 1 from the size-axis to the diagonal. The equation is given in (3).

(3) $$z(x) = \frac{4}{\pi} \arctan \left(\frac{\mathrm{depth}(x)}{\mathrm{size}(x)} \right)$$

Figure 3 shows histograms of the factor z. It is clear from Subfigure (a) that shorter threads are overwhelmingly linear conversations. Very large threads are likely to be broadcasts with many replies but no depth (Subfigures (c) and (d)). Finally, threads with a medium angle (in the middle of the histograms) are likely to be group discussions, conversations with a relatively large size and medium depth, so they contain some branching structures (see Figure 1(b) for illustration). This diversity in the structure and nature of Twitter threads has implications for linguistic

[6] *U1: Koller says in SN: "The middle class has to pay" [link] Too bad that he doesn't tweet more @U2 — U3: @U1 @U2 Removing the capital gains tax doesn't seem useful or good for the middle class to me. — U1: @U3 it doesn't? — …*

analysis, since for example group discussions should be expected to be quite different from broadcasts in some respects. The red lines separating the broadcasts from the group discussions and linear conversations have been selected visually, but in future work the separation should be set algorithmically.



(a) Threads up to five tweets long. (b) Threads from six–20 tweets.

(c) Threads from 21–50 tweets. (d) Threads over 50 tweets long.

Figure 3: Histograms of factor z relating size and depth for threads. N is the total number of threads pictured in each graph.

5 LINGUISTIC PROPERTIES OF TWITTER DISCOURSES

In the following, we will consider some linguistic properties of Twitter conversations in turn, in order to determine their similarity and differences with spoken conversations.

5.1 Dialog Acts

In studying spoken conversations, dialog acts are often used to characterize their linguistic structure, topic composition, and type. For example, information exchanges contain many questions and answers, whereas argumentative exchanges include more agreements, disagreements, and so on. In earlier works (Zarisheva and Scheffler 2015, Scheffler and Zarisheva 2016) we annotated a set of 172 Twitter conversations (1,213 tweets) with 57 dialog acts from an adapted DIT++ schema (Bunt et al. 2010). The ten most frequent dialog acts found in Twitter conversations are shown in Table 3, along with the ten most frequent acts in the Switchboard telephone conversation corpus (Stolcke et al. 2000). The Twitter dialogs (we analysed a mix of long and short conversations) resemble spoken conversations in the way that declarative acts (Statement in the DAMSL schema, Inform and Information Providing in the Twitter schema) are by far the most frequent. Agreements and different types of questions also frequently occur in both kinds of conversations. However, spontaneous speech is characterized by Backchannels, Abandoned utterances and Non-Verbal material, which does not occur frequently in Twitter. Instead, the short length of most Twitter dialogs can be seen from the fact that Open[ing]s and TopicIntroductions can be found in the top ten dialog acts. In addition, the overall higher frequency of questions, agreements, and disagreements suggests a larger portion of informational and argumentative exchanges in the Twitter dialogs.

Table 3: Dialog acts in the Switchboard telephone corpus and Twitter conversations.

Switchboard		Twitter	
36%	Statement	25%	Inform
19%	Backchannel	11%	InformAnswer
13%	Opinion	9%	Agreement
6%	Abandoned	8%	SetQuestion
5%	Agreement	6%	Disagreement
2%	Appreciation	6%	PropQuestion
2%	Yes-No-Question	5%	Information-Providing
2%	Non-Verbal	3%	Correction
1%	Yes-Answers	3%	TopicIntroduction
1%	Conventional-Closing	3%	Open

5.2 Questions

The dialog act analysis shows that questions are very common in Twitter conversations. Questions are an important marker of an interactional style (Storrer 2013), and are very rare in most written texts. All types of questions make up 18% of the utterances in the Twitter dialog act corpus. In contrast, the German newspaper commentary corpus PCC (Stede and Neumann 2014) contains only 75 questions in 2,900 sentences (2.6%).

There are a number of reasons for using questions on Twitter. While many questions are uttered to fill information gaps or ask for opinions, another typical use in conversation is for clarification, in order to initiate repair of communication problems. In German Twitter discussions, clarification questions are frequently marked by multiple question marks. (Purver et al. 2001) distinguish seven types of clarification questions. In an annotation study of 194 clarification questions from our corpus,[7] we found instances of all types except the rare gaps and gap fillers, which seem to depend on spoken interaction. Table 4 shows the prevalence of different types of clarification questions in Twitter conversations vs. the spoken conversations from the British National Corpus analysed in (Purver et al. 2001), with examples from our Twitter corpus. The linguistic means for marking clarification questions on Twitter resemble those used in spoken dialogs. Conventional phrases such as 'what?'/'really?' are frequently used, as are different types of reprise questions. Certain types of clarification questions that address a specific detail of the previous utterance (such as 'already?' as a reply to 'Should we pick you up?') do not fit any of the seven types of clarifications introduced in Purver et al. (2001). Finally, clarification questions on Twitter are sometimes marked solely with a range of question marks, without any further linguistic content. In speech, this may correspond to a confused facial expression and it could be seen as another (novel) conventional means of marking a clarification question on social media.

Even though the linguistic types of clarification questions found on Twitter resemble those in spoken conversation, their function is sometimes different. Since previous utterances are in the written mode and therefore persistent over time, clarification questions are not triggered by failure to hear/see what was said. Instead, questions like (6) are meant sarcastically or at a meta-level (= "Did you really mean to say what you just said?"). Many communication problems (and subsequent clarification questions) are due to the fact that it is hard to distinguish between sarcastic or ironic and literal utterances on Twitter. Many of the clarification questions thus tried to figure out whether the speaker meant what they said literally or was joking. Regular non-reprise clarification questions such as (7) can also be used for this purpose.

[7] Many thanks to Julia Gantzlin for annotating the data.

Table 4: Types of clarification questions in Twitter and spoken conversation.

Type	BNC	Twitter	Example (Twitter)
Reprise fragments	29.10%	22.60%	(4) was ihr tun könnt??? Mich aus der insolvenz retten mir 150 tausend Euro überweisen!!!! *what you can do??? Save me from bankruptcy wire me 150 thousand Euro*
Reprise sluices	12.80%	22.10%	(5) wieso heimlich??? Darf ruhig jeder wissen :D *why secretly??? Anybody can know it :D*
Reprise sentences	8.90%	1.00%	(6) die Erde ist rund??? Oh Oh das musste schon mal jemand zurück nehmen! *the Earth is round??? Uh oh someone had to take that back before!*
Non-reprise clarifications	13.30%	15.50%	(7) wie meinst du das? *how do you mean?*
Gaps	0.50%	0%	
Gap fillers	3.80%	0%	
Conventional	30.70%	30.90%	(8) hä??? Eher overgedressed *whaaa??? More like overdressed*
Question marks	–	4.00%	(9) ?????????????? ich komm hier jetzt gar nicht mehr mit.... *?????????????? I can't keep up here....*
Others	–	3.60%	(10) [sollen wir dich abholen? —] jetzt schon?? *[should we pick you up? —] already??*

5.3 Particles

According to the literature, German modal particles are a phenomenon that is mainly found in spoken language (Bross 2012). Though the use of particles has a colloquial feel, it is not immediately clear whether the use of modal particles depends on the spoken medium, colloquial style, or interactional vs. informational types of conversation. Here, we compare the occurrence of modal particles in the Twitter conversations with the German newspaper corpus PCC and the spoken-like (though edited) OpenSubtitles[8] corpus (Lison and Tiedemann 2016). We study the 17 common modal particles listed in König (1997). In the newspaper commentaries, these particles make up 3.2% of (non-punctuation) tokens. In the Twitter conversations, they are more common, accounting for 4.4% of tokens. This is true despite the fact that these conversations contain many additional Twitter-specific tokens, such as user names and URLs, that inflate the token count. Particles make up 2.9% of tokens in the subtitles corpus.

[8] http://www.opensubtitles.org/

The distribution of particles among the three corpora is shown in Figure 4, which shows the occurrence frequency relative to the number of (non-punctuation) tokens in the corpora. It can be seen that the particle 'ja' in particular is much more frequent in Twitter and OpenSubtitle conversations. This is due to the fact that this item is used as the answer particle '*yes*' as well as a modal particle. In addition, 'aber' (*however*), 'auch' (*also*), 'halt' (*just*), and 'schon' (*already*) are also more frequent on Twitter. Other particles, such as 'doch' (*however*), 'wohl' (*possibly*), and especially 'nun' (*now*) may in fact be more typical of written language and/or informational style than conversations. It seems, therefore, that a blanket statement to the effect that modal particles are generally more frequent in speech (or spoken-like social media) is unsupported based on this data. Different particles show very different profiles depending on the context of the communicative situation.

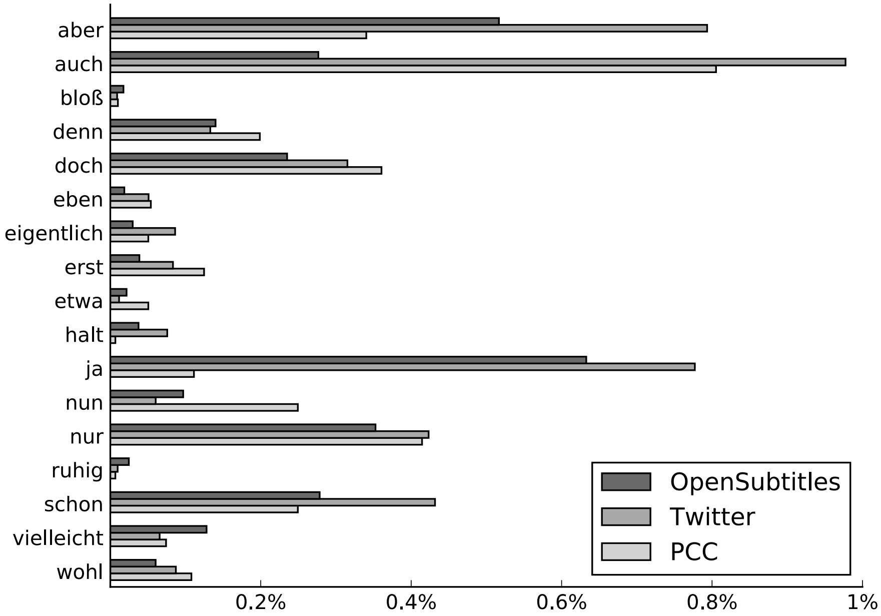


Figure 4: Frequency of modal particles in Twitter, scripted speech (OpenSubtitles) and written newspaper text (PCC).

5.4 Intensifiers

The use of intensifiers such as 'really' and 'very' is associated with informal and colloquial registers, in particular spoken conversations. Tagliamonte and Denis

(2008) analyse speech and IM text messages from Toronto teenagers and show that intensifiers also occur frequently in the text messages, though slightly less often than in speech. But they also note that the choice of intensifier depends on the medium. In text messaging, the teenagers prefer the innovative variant 'so' over formal 'very' and informal 'really,' whereas 'really' is the most frequent variant in speech.

Here, we look at the use of formal and informal intensifiers in the German Twitter conversations vs. newspaper texts. First, the expectation that intensifiers are more common in conversations carries over to the Twitter data. In the Twitter dialogs, 0.46% of all tokens are intensifiers. In the newspaper commentaries, intensifiers only amount to 0.14% of tokens. Next, we compare the use of formal vs. informal intensifiers given in (11) and (12), respectively. Formal intensifiers are relatively more frequent in the texts, accounting for 65% of all intensifiers. In Twitter conversations, the informal variants account for about the same number of intensifiers as the formal variants (50%; see Figure 5). But interestingly, the formal variants are still very common here as well. In future work this should be compared to spoken data, or that obtained from other social media.

(11) formal: wirklich ('really'), sehr ('very'), absolut ('absolutely')

(12) informal: echt ('really'), krass, extrem ('extremely'), ordentlich, total ('completely'), sau, voll, völlig ('completely')

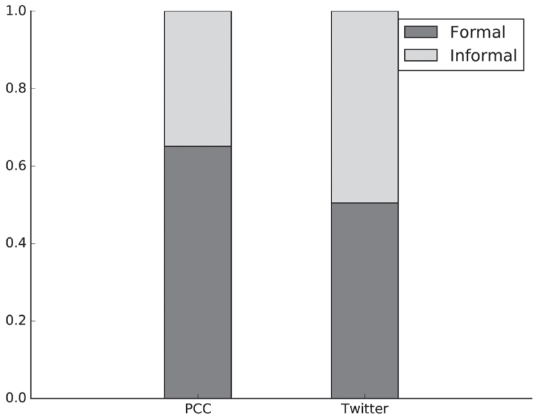


Figure 5: Ratio of formal and informal intensifiers in newspaper text vs. Twitter conversations.

6 CONCLUSION

In this paper, we have provided a view of one particular set of social media data, Twitter conversations. These conversations are computer-mediated and thus come in written form, but otherwise resemble spoken conversations in structural respects. The participants in Twitter conversations are not restricted in number and this can change throughout the conversation, just like in face-to-face interactions. The participants are furthermore relatively equal in standing, and make their utterances spontaneously and in a relatively short time span (though not synchronously, as in spoken conversations). Since successful communication is a joint action, speakers and hearers must coordinate to achieve their common communicative goals. This coordination process can be observed through adjacency pairs (or dialog act sequences) and other grounding phenomena, such as corrections and clarification questions.

The Twitter dialogs considered here exhibit all the linguistic markers typically attributed to face-to-face conversations, though some differences can be found. On the one hand, the most prominent dialog acts in Twitter conversations are informational, just like in speech. But due to the very short length of many Twitter threads, openings and topic introductions are also more frequent in the Twitter corpus. In addition, a subset of Twitter discussions is clearly argumentative, which leads to a slightly higher portion of agreements and disagreements. On the other hand, common phenomena of unplanned spontaneous speech, such as backchannels and fragments, are almost completely missing from Twitter conversations. Rehbein (2015) uses the example of filled pauses, and demonstrates that when such speech-specific phenomena are present on Twitter, they are used deliberately to carry extra-propositional meaning.

Based on the analyses shown here, computer mediated conversations can be interesting data sources for some linguistic phenomena that are specific to informal conversation, but difficult to study in spoken corpora. We have shown that, for example, questions are very frequent in the Twitter threads, but not in newspaper corpora. The case of clarification questions furthermore underlines the joint communicative action between speakers and hearers, as these instances highlight cases where communication breaks down because of mis- or non-understandings. Twitter users avail themselves of the same linguistic means to mark clarification questions, but they add an innovative variant thanks to the written mode, an indication of non-understanding with only a series of question marks.

Despite the similarities, it is not the case that Twitter conversations are just written versions of spoken dialogs. As expected, particles and intensifiers are found frequently in Twitter conversations as features of informal, colloquial language.

In this respect, the CMC conversations differ markedly from standard newspaper corpora in both the frequency and range of items that are used. But it is to be expected that the use of these linguistic items also differs from their use in speech corpora, as shown for English intensifiers by Tagliamonte and Denis (2008). Further work is thus needed to situate Twitter conversations (and other social media) on the 'conceptual orality' continuum and determine the mix of conservative and innovative features that can be observed.

Finally, we showed through an analysis of the dialog structure of Twitter conversations that even within this medium, different types of conversations must be distinguished. This distinction was made on structural grounds, not based on topic or linguistic features (which could make the definition circular). While most conversations are very short (typically, only one root plus a reply), longer conversations belong to three broad classes: 'Broadcasts' contain root tweets which get many replies (usually from different users) but do not lead to any further discussion; they are characterized by a short depth and are often linguistically less complex. 'Linear conversations' are private discussions among a very small number of users, which develop in a linear fashion, i.e. each answer is a reply to the last contribution. Finally, there is a number of conversations in between the two extremes, exhibiting some branching of the dialog tree. We called these 'group discussions'. All conversation data from Twitter is much less likely to contain bot generated content than a random set of tweets, which makes it very amenable to linguistic research.

In sum, Twitter conversations are made up of informal, interactive exchanges between speakers which allow us to tease apart the differences between highly edited, monological text and spontaneous, colloquial speech on several dimensions. This will enable more detailed studies of linguistic phenomena across different traditional and computer-mediated channels of communication.

References

Austin, John L., 1975: *How to Do Things with Words*. Oxford University Press.

Bartz, Thomas, Michael Beißwenger and Angelika Storrer, 2013: Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL* 28/1/. 157–98.

Beißwenger, Michael, 2007: *Sprachhandlungskoordination in Der Chat-Kommunikation*. Berlin: De Gruyter.

Beißwenger, Michael, 2013: Das Dortmunder Chat-Korpus: Ein Annotiertes Korpus Zur Sprachverwendung Und Sprachlichen Variation in Der Deutschsprachigen Chat-Kommunikation. *LINSE-Linguistik Server Essen*. 1–13.

Biber, Douglas, 1993: The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26/5-6. Springer. 331–45.

Bross, Fabian, 2012: German Modal Particles and the Common Ground. *Helikon: a Multidisciplinary Online Journal* 2. 182–209.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria and David Traum, 2010: Towards an ISO Standard for Dialogue Act Annotation. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. 2548–2555.

Clark, Herbert H. and Edward F. Schaefer, 1987: Collaborating on Contributions to Conversations. *Language and Cognitive Processes* 2/1. Taylor & Francis. 19–41.

Clark, Herbert H. and Edward F. Schaefer, 1989: Contributing to Discourse. *Cognitive Science* 13/2. Wiley Online Library. 259–94.

Core, Mark and James Allen, 1997: Coding Dialogs with the DAMSL Annotation Scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*. 28–35.

Koch, Peter and Wulf Oesterreicher, 1985: Sprache Der Nähe–Sprache Der Distanz. *Romanistisches Jahrbuch* 36/85/. 15–43.

König, Ekkehard, 1997: Zur Bedeutung von Modalpartikeln im Deutschen: Ein Neuansatz im Rahmen der Relevanztheorie. *Germanistische Linguistik* 136/1997. 57–75.

Lison, Pierre and Jörg Tiedemann, 2016: Opensubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. 923–929.

Lui, Marco and Timothy Baldwin, 2012: langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 25–30.

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang and Alessandro Vespignani, 2013: The Twitter of Babel: Mapping World Languages Through Microblogging Platforms. *PloS One* 8/4.

Purver, Matthew, Jonathan Ginzburg and Patrick Healey, 2001: On the Means for Clarification in Dialogue. *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue*. 116–25.

Rehbein, Ines, 2015: Filled Pauses in User-Generated Content Are Words with Extra-Propositional Meaning. *Proceedings of ExProM*. 12–21.

Ritter, Alan, Colin Cherry and Bill Dolan, 2010: Unsupervised Modeling of Twitter Conversations. *Proceedings of NAACL*. 172–180.

Rudolph, Elisabeth, 1991: Relationships Between Particle Occurrence and Text Type. *Multilingua* 10. 203–23.

Scheffler, Tatjana, 2014: A German Twitter Snapshot. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. 2284–2289.

Scheffler, Tatjana and Elina Zarisheva, 2016: Dialog Act Recognition for Twitter Conversations. *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. 31–38.

Stalnaker, Robert, 1978: Assertion. Cole, Peter (ed.): *Syntax and Semantics 9: Pragmatics*. New York: Academic Press.

Stede, Manfred and Arne Neumann, 2014: Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. 925–929.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema and Marie Meteer, 2000: Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26/3. Cambridge, MA: MIT Press. 339–73.

Storrer, Angelika, 2013: Sprachstil Und Sprachvariation in Sozialen Netzwerken. Frank-Job, Barbara, Alexander Mehler and Tilmann Sutter (eds.): *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Storrer, Angelika, 2014: Sprachverfall Durch Internetbasierte Kommunikation: Linguistische Erklärungsansätze – Empirische Befunde. *Sprachverfall?: Dynamik–Wandel–Variation (Jahrbuch des IDS)*. 171–96.

Tagliamonte, Sali A. and Derek Denis, 2008: Linguistic Ruin? LOL! Instant Messaging and Teen Language. *American Speech* 83/1. 3–34.

Zarisheva, Elina and Tatjana Scheffler, 2015: Dialog Act Annotation for Twitter Conversations. *Proceedings of SIGDial16*. 114–23. Prague, Czech Republic: Association for Computational Linguistics. http://aclweb.org/anthology/W15-4614. (Last accessed 29 June 2017.)

Exploring Wikipedia Talk Pages for Conflict Detection

Lydia-Mai Ho-Dac, *University of Toulouse, CNRS*
Veronika Laippala, *University of Turku*
Céline Poudat, *University of Nice Côte d'Azur*
Ludovic Tanguy, *University of Toulouse, CNRS*

Abstract

The present study concentrates on Wikipedia talk pages, which are online discussions where the authors discuss the composition and content of Wikipedia articles. These pages provide new data for describing and analysing collaborative writing processes, which often involve conflicts. Previously, many studies have explored Wikipedia conflicts, highlighting opposite editing patterns in relation to cooperation, conflicts or quality. Most of these studies belong to the field of social sciences, and linguistic analyses are not very common in this context. Therefore, the linguistic characteristics of Wikipedia conflicts in talk pages are still little described in the literature. In this context, our objective is to analyse relevant linguistic cues which may help identify and characterize conflicts on Wikipedia talk pages. To this end, we apply two automatic methods. The first one consists of the supervised automatic classification of conflicting vs. harmonic discussion threads. In the second we apply multidimensional analysis to the data to help profile the Wikipedia talk genre, enabling us to highlight key features and oppositions at a global level. The analyses are carried out on the WikiTalk corpus, a resource based on the French Wikipedia talk pages (160M words, 3M posts, 1M threads). The corpus includes a wide range of metadata, providing extra-linguistic characterization of the Wikipedia discussions.

Keywords: French Wikipedia talk pages, conflict detection, data-driven approaches

1 INTRODUCTION

The exponential development of the Internet has led to new communicative situations and genres. These new online genres, which are not yet fully characterized, are complex objects challenging the existing methodologies and analysis tools. In this context, the Wikipedia encyclopaedia project is one of the new textual objects that can be studied under the umbrella term Computer-Mediated Communication (CMC, see Herring et al. 2013). Wikipedia, which has now been available for more than 15 years, is an open and collaborative project, available in numerous languages. The success of this online encyclopaedia is indisputable, as evidenced by its huge size (5M articles in the English Wikipedia and 1.7M in the French Wikipedia, as of June 2016). In addition, Wikipedia is one of the 10 most consulted websites in the world.[1]

Over the last decade, Wikipedia has become a wealth of information which is increasingly used in the development of natural language processing (NLP) and text mining applications (Ferschke et al. 2013). It has also been the subject of many studies in social sciences. Indeed, since the quality of the encyclopaedia was first established by Giles (2005), a large number of studies have used Wikipedia to examine the coordination and collaboration processes that occur among people (Viegas et al. 2007, Brandes and Lerner 2007, Kittur and Kraut 2008, Stvilia et al. 2008), via the analysis of revisions and talk pages which provide evidence of collaborative editing, maintenance work, cooperation and conflict resolution (Kittur et al. 2007, Viégas et al. 2004).

Most of these studies do not focus on the linguistic and discursive aspects of Wikipedia pages, most likely because of the sprawling structure of the site (its multiplicity of pages and versions), which makes corpus building quite difficult. As a consequence, these works mostly rely on network analysis or on statistical features extracted from article revision histories. For instance, article reverts (when users restore a previous version) have proven to be significant features in the detection of conflicts (Viégas et al. 2004, Brandes and Lerner 2007, Kittur et al. 2007, Suh et al. 2007, Kittur and Kraut 2010, Miller 2012). Nevertheless, such features remain indirect markers of conflicts, as they may be interpreted differently, allowing no clear distinction between editorial conflicts and vandalism, for instance (Potthast et al. 2008, Yasseri et al. 2012, Adler et al. 2011). Other commonly used criteria include article and talk page length, number of revisions in article and talk pages, number of anonymous edits/users, character or word insertion or deletion between users, article labels, and so on.

Such criteria serve as the basis for the automatic detection of quality articles (Wilkinson and Huberman 2007), pages that are the focus of conflicts (Kittur et

[1] https://www.alexa.com

al. 2007, Vuong et al. 2008, Sumi et al. 2011), or topic categories which are more likely to generate conflicts, such as religion and philosophy, according to Kittur et al. (2009).

Although these studies have provided interesting insights on the evolution of Wikipedia's organization and collaborative editing, the linguistic characteristics of Wikipedia pages remain under-explored. In particular, talk pages are particularly interesting to observe as they are at the heart of the Wikipedia process. Each article is associated with a talk page, where most of the coordination work is done, and where potential conflicts are discussed and ultimately resolved in the best-case scenario (Viegas et al. 2007). Talk pages are the places where editors discuss the modifications to be made to an article, including sections to be rewritten or removed (Ferschke et al. 2012).

Wikipedia talks may be considered as a new discussion sub-genre. Wikipedia editorial talk pages are indeed quite specific: (i) they are directly related to the article they are associated with, and they share a common focus, i.e. article editing and improvement; (ii) they contain open asynchronous discussions that anyone may edit. In this respect they might be compared to forum discussions, except that they rely on a specific Wiki technology which has direct consequences on the macrostructure: in spite of clear recommendations concerning the form of the postings (level of the answer, mandatory signature and date, etc.), talk pages are often hybrids, combining dialogues whose structure may not be obvious (as Wikipedians may, for instance, edit previous postings), and checklist elements; (iii) they share common features referring in particular to editing actions, conflict management and Wikipedia procedures (e.g. NPOV, i.e. Neutral Point of View, relevance, source, quality, and so on).

Conflicts are particularly interesting to observe on Wikipedia, since they can be considered as frontiers between collaboration and discussion. Antagonistic edits of the article structure and content may indeed lead to disagreements, and this is quite common when co-editing, before participants agree on a more stable version of the article. Disagreements may turn to conflicts when the editing process and/or the discussion process are deadlocked, which leads to an automated report. In such cases, pages are tagged with specific labels signalling that a conflict is ongoing on the article or talk pages (e.g. NPOV or relevance disputes, "Calm talk" template). There are many examples of pages with such labels, such as *Abortion in Iran*, *Bengali cuisine*, and *Religion and sexuality*, to cite just a few. If a conflict grows in intensity and verbal abuse occurs, then the article and talk page may be blocked and some users may be banned; for instance if they write "toxic" comments by making personal attacks.[2] From Wikipedia's

[2] One of the policy of WP is to avoid any kind of personal attacks (see https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks).

point of view, conflicts must be regulated as they impact productivity, as noted in Wulczyn et al. (2016:2), "the Wikimedia foundation found that 54% those who had experienced online harassment expressed decreased participation in the project where they experienced the harassment".[3] Wulczyn et al. (2016) aimed to develop tools to identify toxic comments, and their first experiment on Wikipedia talk pages resulted in "Wikipedia DeTox",[4] an automatic detector of toxic comments. This automatic device is currently adapted to other CMC under the name "Perspective API," which provides the following definition of "toxic": "a rude, disrespectful or unreasonable comment that is likely to make you leave the discussion".[5] The relationship between toxicity, or verbal violence, and conflict is obvious, although verbal violence and toxicity are generally detected at the post level (Wulczyn et al. 2016), whereas conflicts are better observed and detected at the thread level, with threads corresponding to the sections of talk pages in this context.

The aim of the present study is thus twofold: (i) We would first like to explore the differences between the threads belonging to talk pages reported to be sources of conflict by Wikipedians, and the threads belonging to talk pages where no problems have been reported. Are the first set of threads clearly distinct from the second? With this in mind, we will perform an automatic classification on the WikiTalk corpus. (ii) At a descriptive level, we would like to contribute to the linguistic description of the discussions on Wikipedia talk pages, which have been little explored using linguistic criteria. Indeed, few linguistic studies have been conducted on French Wikipedia – see Denis et al. (2012) on the detection of conflicting threads and Poudat and Loiseau (2007) on the exploration of Wikipedia categories. In order to have a broader view of the linguistic characteristics of the French Wikipedia talk pages, we will propose a first profiling of the genre, using a mutidimensional analysis enabling us to highlight key features and oppositions at a global level. Threads that are the focus of conflicts will then be characterized within this global generic profile.

2 THE WIKITALK CORPUS

The WikiTalk corpus is composed of talk pages extracted from the French Wikipedia dump dated May 12th 2015, which contains 3.5M talk pages. Only 365,612 pages were kept in the released WikiTalk Corpus. Indeed, 57% of the talk pages were user pages and we chose to remove these, as they may not be

[3] These findings are reported in a report called "Harassment Survey" made available by the Wikipedia Foundation at the url https://commons.wikimedia.org/w/index.php?title=File%3AHarassment_Survey_2015_-_Results_Report.pdf.

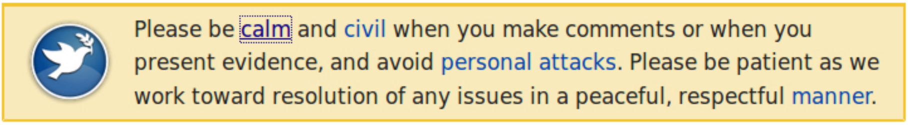
[4] https://tools.wmflabs.org/detox/

[5] http://www.perspectiveapi.com/

editorial discussions. Moreover, only 24% of the remaining talk pages contained more than two words.[6] The 365,612 remaining talk pages were associated with metadata, segmented into threads (i.e. headed sections) and posts (i.e. comments) and formatted according to the TEI-P5 guidelines.

Three kinds of metadata were automatically extracted to categorize and describe the discussions:

1. "*discipline*" indicates the associated thematic portals,
2. "*avancement*" (progress) corresponds to the article's quality scale based on Wikipedian assessments,[7]
3. "*interaction*" gives information about possible conflicts in the discussion. Such information may be manually inserted by Wikipedians via the template {{Calm talk}} which adds a dedicated banner to the top of the talk page (see Figure 1).[8]



Please be calm and civil when you make comments or when you present evidence, and avoid personal attacks. Please be patient as we work toward resolution of any issues in a peaceful, respectful manner.

Figure 1: The {{Calm talk}} banner.

These metadata are encoded in the teiHeader in the <classDecl> element:

```
<category type="discipline">
   <catDesc>Politique</catDesc>
   <catDesc>France</catDesc>
</category>
<category type="avancement">
   <catDesc>Featured</catDesc>
</category>
<category type="interaction">
   <catDesc>{{calm}}</catDesc>
</category>
```

Automatic thread and post segmentation is based on the wikicode with the help of local grammars. Thread segmentation is achieved using the headings signalled in the wikicode by the pattern /==.*?==/. On the other hand, post segmentation is performed using both the signature manually inserted by the writer (such as:

[6] 1,013,791 (68%) talk pages were blank and 116,432 (8%) consisted in redirections to another talk page.

[7] https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

[8] https://en.wikipedia.org/wiki/Template:Calm

Viking59 10 mai 2009 à 17:16 (CEST), and the presence of a change in the interactional level indicated by the number of semi-colons (:) at the beginning of the post. Figure 2 illustrates the encoding of the wikicode into the TEI-P5 norm according to the following transformations: <div> for threads, <head> for topic titles, <post> and the three attributes: @who, @when and @interactionalLevel for posts.

Wikicode

```
== Jeux ==
Sinon, ce serait bien de retravailler la section […]
Fredscare 18 avril 2007 à 17:00 (CEST)
:J'ai retravailler la section […] Bouchette63 6 avril 2008 à
02:10 (CEST)
::J'ai vidé la section […] PV250X 15 avril 2009 à 20:39
(CEST)

==Situation actuelle (2005 à aujourd'hui)==
Bonjour, […]
```

TEI-P5 encoding

```
[…]
<div id=“3” level=“1”>
<head>Jeux</head>
<post id=“5” who=“Fredscare” when=“18-04-2007-17:00”
interactionalLevel=“0”>
   <p id=“1”>Sinon, ce serait bien de retravailler la section
[…]</p>
</post>
<post id=“5” who=“Bouchette63” when=“06-04-2008-02:10”
interactionalLevel=“1”>
   <p id=“1”>J'ai retravailler la section […]</p>
</post>
<post id=“5” who=“PV250X” when=“15-04-2009-20:39”
interactionalLevel=“2”>
   <p id=“1”>J'ai vidé la section […]</p>
</post>
</div>
<div id=“4” level=“1”>
<head>Situation actuelle (2005 à aujourd’hui)</head>
<post who=“anonyme” bot=“no” when=“unknown”
interactionalLevel=“0”>
   <p id=“1”>Bonjour, […]</p>
[…]
```

Figure 2: From Wikicode to TEI-P5 encoding (extract from the “Sega” talk page).

Eight of the extracted talk pages, amounting to 413 posts and 47,284 tokens, were manually inspected to evaluate the extraction process. The results show that 23 posts were not extracted at all, and 33 posts were wrongly delimited, among which 25 merged several posts in one. As a result, the extraction process has an estimated precision of 0.92 and a recall of 0.95. Post attribute values (@who, @when and @interactionalLevel) were only checked for one talk page, but indicated 100% accuracy. Table 1 gives a quantitative overview of the WikiTalk corpus.[9]

Table 1: Quantitative overview of the WikiTalk corpus.

#talk pages	#threads	#posts	#words
365,612	1,023,841	2,406,514	161,833,298

3 CLASSIFICATION OF CONFLICTING VS. NEUTRAL DISCUSSIONS

Are threads belonging to talk pages associated with conflicts significantly different from those belonging to harmonic or neutral pages? To answer this question, we carried out a data-driven comparison of the global linguistic characteristics of two classes of discussions, distinguished according to an experimental classification of "conflicting" vs. "neutral" talks. The selection criteria used for distinguishing between these two classes are based on alerts and reporting issued by Wikipedians.

3.1 Experimental DataSet for thread classification

An automatic classification of the WikiTalk corpus has already been tested for distinguishing Wikipedia talk pages from Wikipedia articles and other CMC, such as online forums (Ho-Dac and Laippala 2017). The results showed that these three text genres could be automatically detected on the basis of a simple bag of words. Unfortunately, we could not adopt the method proposed in Ho-Dac and Laippala (2017) for the following two reasons. First, in contrast with Ho-Dac and Laippala (2017), where talk pages, Wikipedia articles and online forum were clearly identified genres and large amounts of training data were easily available, there is no training data available for conflict detection, as no large-scale corpora with discussions annotated as conflicting or not exist. Secondly, as opposed to

[9] Soon available at http://redac.univ-_tlse2.fr/

Ho-Dac and Laippala (2017), where the analysis could be done over entire talk pages and Wikipedia articles, the thread level seems more suitable for detecting conflicts, as thus is used in this work.

As stated above, the development of a supervised machine learning system that would automatically classify threads requires a large amount of threads categorized as conflicting vs. neutral. In order to provide training data and because there is very little information at the thread level, we opted for an experimental classification of "conflicting" vs. "harmonic/neutral" talk pages, and then used this to assess the hypothesis that threads belonging to "conflicting" talk pages would be significantly different from those belonging to "harmonic/neutral" pages. The selection criteria used for distinguishing between these two classes are based on alerts and reporting issued by Wikipedians.

We considered that talk pages were conflicting when they were associated with metadata signalling the presence of a conflict, that is:

- <category type="interaction"> in teiHeader indicates that the "calm talk" template was inserted;
- a parallel talk page was created for discussing the article's neutrality;[10]
- the talk page is not a main page but a parallel talk page created for discussing the article's neutrality.

In contrast, talk pages associated with featured articles[11] were considered to be "neutral," based on the assumption that the acknowledged quality of these articles means that there is a consensus amongst the contributors. Criteria for *a priori* "neutral" talks are as follows:

- <category type="avancement"> in teiHeader indicates that the associated article was assessed to be "Featured" or "A-class";
- a parallel talk page was created for deciding if the article deserves the "featured" or "A-class" status.

The resulting data set collected from the WikiTalk corpus based on these criteria is described in Table 2. Note that all the talk pages which contained less than 100 words were excluded.

[10] This possibility seems specific to the French Wikipedia.

[11] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

Table 2: Experimental dataset for the classifier: conflict vs. neutral discussions.

Selection criteria	#talk pages
More than 100 words in the talk page	152,931
Conflict discussions (11 M words)	**2,028**
Calm talk template in the header	39
Existence of a parallel NPOV talk page	1,782
Talk page is a "neutrality" talk page	207
Neutral discussions (8.8 M words)	**4,569**
A-class article mentioned in the header	1,099
Existence of a parallel talk page about A-ranking	3,470

3.2 Thread classification on the experimental DataSet

We trained a text classification model using the Vowpal Wabbit linear classifier (Agarwal et al. 2011), and tested it on a sub-part of the threads that were experimentally classified (henceforth "Experimental DataSet"), and also on the threads that were manually annotated (henceforth "Annotated DataSet").

Four feature sets were tested: words, lemmas, character 5-grams and syntactic N-grams. While the first three sets are the one used in the traditional lexical approach, as in, for example, Scott and Tribble (2006), which proposes using keyword analysis to reflect thematic and stylistic features. Classification based on syntactic N-grams is less common (Kanerva et al. 2014, Goldberg et al. 2013). The syntactic N-grams we used are delexicalized *bi-arcs* composed of two syntax dependencies between tokens, with the actual lexical information deleted, but with all other information on the syntactic dependency, Part-of-Speech and other morphological features, as illustrated in Figure 3.

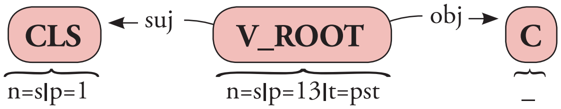


Figure 3: A delexicalized syntactic bi-arc describing a clitic+verb+conjunction as in the clause 'I find that'.

Syntactic analysis and lemmatisation were provided by the Talismane toolkit (Urieli 2013). The classification method based on syntactic N-grams enables a more robust analysis based on text characteristics that does not depend on the text topic, but instead attempts to generalize the level of description beyond individual lexical topics to typical structures (Laippala et al. 2015).

The first classification experiment is performed using the stochastic gradient method with two-thirds of the Experimental DataSet used for training and the remaining for testing. Table 3 gives the precision (P) and recall (R) for detecting the "conflict" category by using the two feature sets on 46,690 threads.

Table 3: Comparison of different lexical vs. syntactic approaches for the automatic classification of conflicting threads and posts.

Features	P	R	F-measure
Words	0.83	0.65	0.74
Lemmas	0.84	0.60	0.72
Character 5-grams	0.82	0.72	0.77
Syntactic Bi-arcs	0.55	0.48	0.52
# threads	**46,690**		

The results show that character-based and lexical feature sets have good performance, while bi-arcs consisting of only syntax are not very useful. The best results are achieved by using lemmas. The 40 most distinctive lemmas for the conflicts, as estimated by the classifier, can be divided to two groups:

- words referring to the writing process, highlighting current sources of editorial conflicts, as well as (dis)agreement cues: *style, to hope, respect, version, way of writing, restructuring, reformulation, neutralisation, clumsy, uncoherent, respect, mistake, controversy, debate, ok;*
- words referring to the article topics: *rwanda, dictatorship, mandarin, quebec, islam, buddhism.*

These distinctive lemmas give a clear picture of the characteristics of the threads that the classifier identifies as conflicting. Importantly, we can assume that the first group of lemmas referring to the writing process may be common to all conflicts, regardless of the discussion topic. Considering our general aim of identifying conflicts in general, this is crucial. A closer look on the threads classified incorrectly or with a high probability is, however, necessary in future work in order to better understand the basis of the classification. The features which were selected are informative, but not necessarily explanatory of the ways in which conflicts arise or get resolved.

3.3 Thread classification on the annotated DataSet

The classifier model we obtained was then assessed on an Annotated DataSet, gathering the 215 threads of two talk pages. The two talk pages associated with the articles *Psychoanalysis* and *Bogdanoff brothers* were manually annotated using a binary variable, signalling the presence or absence of an ongoing conflicts in the thread (Poudat et al. 2016). As Table 4 shows, around one thread out of every two was deemed to be conflicting.

Table 4: Annotated DataSet : conflicting annotated threads in two talk pages.

Talk page's topic	# threads	# conflicts	%
Bogdanoff brothers	75	37	49.3
Psychoanalysis	140	74	52.9
Total	215	111	51.6

Table 5 below gives the results of the classification of the annotated DataSet with the model trained on the experimental DataSet. The results indicate that the classifiers trained on the data deemed to be conflicting vs. neutral based on the metadata do not work for the manually annotated conflicts.

Table 5: Classifier results on the annotated DataSet.

Features	P	R	F-measure
Words	0.47	0.53	0.50
Lemmas	0.45	0.47	0.46
Character 5-grams	0.46	0.57	0.52
Syntactic Bi-arcs	0.53	0.45	0.49

As the classifier results on the experimental DataSet reported in Section 3 were decent, this difference indicates that the manually identified conflicts and the threads we assumed as conflicting based on the metadata differ.

In other words, conflict threads may need further linguistic analysis and manual evaluation to be properly detected, as Wikipedia metadata are obviously inadequate and insufficient for this purpose.

The next sections address these questions by proposing a range of new features for profiling threads in a bottom-up approach (Section 4), and presenting an ongoing project of manual conflict annotation in the WikiTalk corpus (Section 5).

4 A BOTTOM-UP APPROACH TO DISCUSSION PROFILING

The automatic classification method was supplemented by a second approach which uses exploratory data analysis techniques based on linguistic and structural features. Our objective is to highlight the structure and the profile of talk pages and threads in a bottom-up approach, without a specific focus on conflict. This method was applied to the whole dataset, i.e. 365,612 talk pages and 1,023,841 threads, using the *R FactoMineR* package dedicated to multivariate exploratory data analysis.[12] Four sets of features were calculated for each talk page and thread, named ***Global***, ***Thema***, ***Interact*** and ***DiscRel***.

4.1 Linguistic and structural features for profiling threads

The *Global* features correspond to general non-linguistic characteristics automatically extracted from the thread and talk page. Table 6 describes the eight *Global* features taken into account in this study.

Table 6: Global features for describing threads.

Label	Description
#words_log	Number of words in the thread (logarithm)
#threads	Number of threads in the page containing the thread
#posts	Number of posts in the thread
max_depth	Maximum depth, i.e., the highest interactional/hierarchical level of a post in the thread
#users_thread	Number of different participants in the thread by considering all anonymous (i.e., unregistered) users as a single participant
%anonymous	Percentage of anonymous posts in the thread, either unsigned or signed by an unregistered user
A-class	Binary feature indicating if the talk page (and by extension the thread) is linked to an A-class article
Keep_calm	Binary feature indicating if the talk page (and by extension the thread) has been tagged with a "calm talk" template

The ***Thema*** features give details of the main topics of the talk pages, based on the portal sections of the associated article. The French Wikipedia comprises 11 portals:[13] Art, Geography, History, Leisure, Medicine, Politics, Religion, Science, Society, Sport and Technology. Geography is the most important portal in the context of this study (119,359 talk pages). Figure 4 gives an overview of the amount of talk pages per portal, although it should be noted that an article (and its associated talk page) may belong to several portals.

[12] http://factominer.free.fr/index.html

[13] https://fr.wikipedia.org/wiki/Portail:Accueil

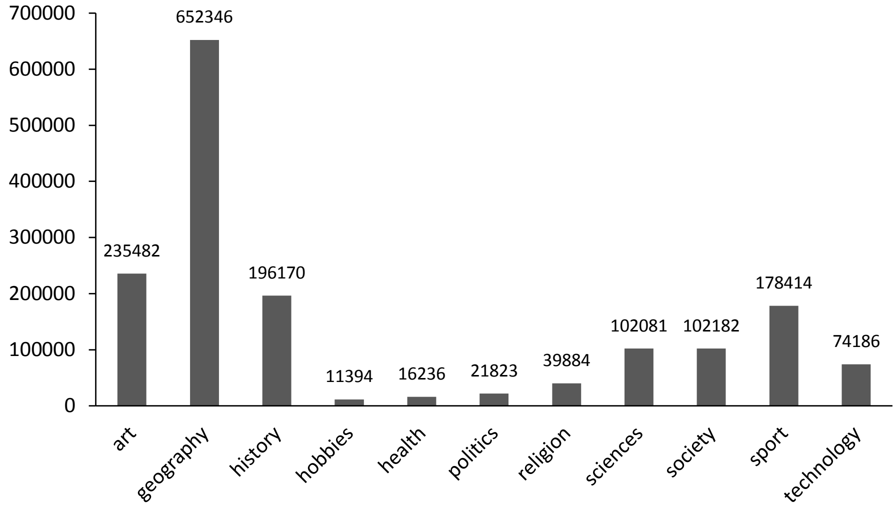


Figure 4: Amount of talk pages per portal.

More than 56% of the articles are categorized in at least two portals (44% in exactly two, with a maximum of six portals for a single article). We thus defined 11 binary features, one for each portal.

The *Interact* features correspond to the relative frequency of a range of basic interaction cues, related to agreement, disagreement and politeness. The counting was performed at the thread level, and 11 different types of cues were automatically identified with simple regular expressions (see Table 7).

Table 7: Interact features for describing threads.

Label	Description
politeness	*thanks, hello, goodbye, hi, sincerely, cheers, please, would you,* etc.
agreement	*OK, agree, yes, no, actually,* etc.
question	question mark (*?*)
je	1st singular person pronouns + the adverb *personally*
tu	2nd sing. pers. pronouns, informal "*you*"
vous	2nd plur. pers. and formal "*you*" pronouns
nous	1st plur. pers. Pronouns
on	Informal "*we*" (indefinite 3rd sing. pers. pronoun)
WP	Explicit reference to the Wikipedia project ("*Wikipedia*" or "*WP*")
pour	Sentence-initial *For* or *I'm for*
contre	Sentence-initial *Against* or *I'm against*

Table 8 gives the number of cues and the proportion of threads in which these *Interact* features were automatically detected. Agreement cues, questions and first singular person mentions occur in more than 25% of the total threads. The rarest features are the formal "we," "pro" and "against." These two latter features are actually very specific to threads dedicated to voting "for" or "against" editorial acts (e.g., article removal or article A-class ranking).

Table 8: Number and proportion of threads with Interact features.

Interact features	#cues	#threads with	%threads with
politeness	317,532	159,924	15.9
agreement	659,291	270,233	26.9
question	751,878	271,237	27.0
je	946,736	386,833	38.5
tu	400,052	106,427	10.6
vous	886,460	217,715	21.7
nous	120,560	79,328	7.9
on	630,616	201,656	20.1
WP	241,510	153,260	15.2
pour	142,785	85,871	8.5
contre	6,987	4,513	0.4
Total		1,005,592	100.0

The last type of feature, called ***DiscRel***, gives an idea of the rhetorical structures occurring in a thread. Using LexConn (Roze et al. 2012), "a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey," we projected these 328 connectives on each thread and measured the cumulative frequency for each discourse relation as defined in LexConn. Twenty-two discourse relations are defined in the LexConn database. When a connective is polysemous, all possible relations were considered. As for *Interact* features, the frequency was normalized on the number of words in the thread.

Table 9 gives the number and proportion of threads and connectives associated with each discourse relation (relation names are those used in the LexConn resource). The two columns labelled "Connectives" provide the number of connectives detected for each relation and proportion it covers among all the discourse relations. The two columns labelled "Threads with" indicate the number and proportion of the threads in which at least one connective expressing the relation occurs.

Table 9: Number and proportion of threads and connectives associated with each discourse relation.

Discourse Relations	Connectives		Threads with	
	#	%	#	%
alternation	583,585	4.9	**317,971**	**31.6**
background	512,690	4.3	189,967	18.9
commentary	25,581	0.2	21,740	2.2
concession	647,056	5.5	248,271	24.7
condition	1,483,308	12.5	496,852	49.4
consequence	162,213	1.4	123,036	12.2
continuation	1,462,713	12.4	469,608	46.7
contrast	528,004	4.5	240,919	24.0
detachment	32,297	0.3	27,487	2.7
elaboration	151,878	1.3	99,880	9.9
evidence	55,707	0.5	43,146	4.3
explanation	1,358,509	11.5	483,269	48.1
flashback	159,759	1.4	102,979	10.2
goal	749,597	6.3	**381,776**	**38.0**
narration	288,718	2.4	151,711	15.1
opposition	1,100,550	9.3	**330,437**	**32.9**
parallel	489,105	4.1	215,176	21.4
rephrasing	158,407	1.3	102,922	10.2
result	657,081	5.6	255,064	25.4
summary	17,858	0.2	15,636	1.6
time	905,059	7.6	**447,176**	**44.5**
unknown	301,741	2.6	157,851	15.7
Total	**11,831,416**	**100.0**	**1,005,592**	**100.0**

Table 9 shows strong variations and extremely frequent relations. Two groups of relation may be distinguished:

- The Condition, Continuation and Explanation relations, which each represent about 12% of all discourse relations, and appear in almost 50% of the total threads (49.4%, 46.7%, 48.1% respectively);
- The Alternation, Goal, Opposition and Time relations, which each represent a smaller percentage of all discourse relations (from 4.9% to 9.3%), but are also detected in a large proportion of the total threads (from 31.6% to 44.5%).

The occurrence of the first group of relations should be linked to the number of words in the thread (the more words, the more of these relations).

4.2 Exploring the threads with PCA

In order to observe how these different features interact with each other, and to help us identify the different thread profiles, we performed a standard multidimensional statistical analysis, and thus a Principal Components Analysis (PCA) was applied on the 1,023,841 threads. As we focus on the linguistic aspects of the discussions, we used the *Interact* and *Discrel* sets of cues as active variables to highlight the structure of the corpus and its main dimensions. The other features were projected afterward as illustrative variables in the reduce-dimension vector space resulting from the PCA.

This first two dimensions explain more than 20% of the total variance, the third one analysed here adding another 5%. Figures 5 and 6 show the first two factor maps, illustrating the main correlations among the features.

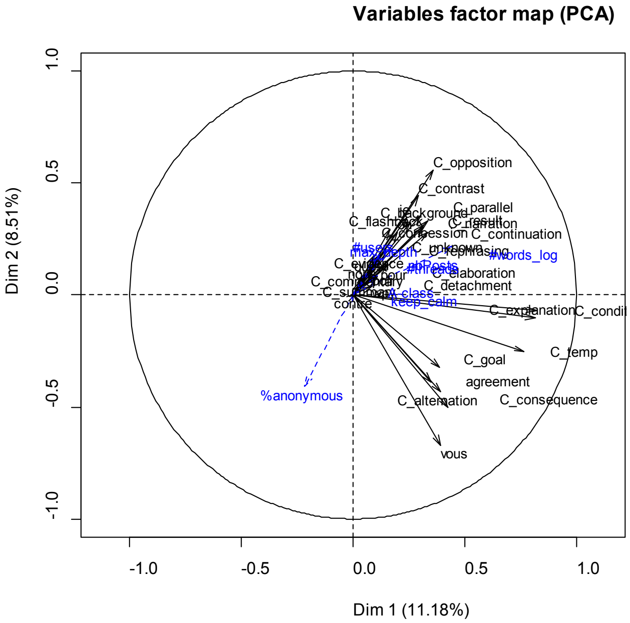


Figure 5: First factor map (dimensions 1 and 2) resulting from the PCA performed by taking into account the linguistic features. Additional features are shown in blue.

The first dimension, explaining around 12% of the total variance, is related to the size of the text units: the more words the threads contains, the more users participate, and the more features there are. As a consequence, the most frequent features (e.g. *Je*, *Vous*, Continuation, Condition and Explanation relations) are also the most significant.

We should also mention that the proportion of anonymous posts is higher for short threads. Let us also note that portals are not associated with significant linguistic cues.

The second and third dimensions are more clearly associated with linguistic features. The second dimension explains more than 8% of the total variance and opposes:

- threads with agreement cues (ok, agree, of course, yes, no, etc.), formal you and a significant presence of consequence, alternation and goal discourse relations (at the bottom of Figure 5); and
- threads containing a substantial amount of I ("je"), formal we/indefinite pronoun ("on") and connectives related to opposition and contrast (at the top of Figure 5).

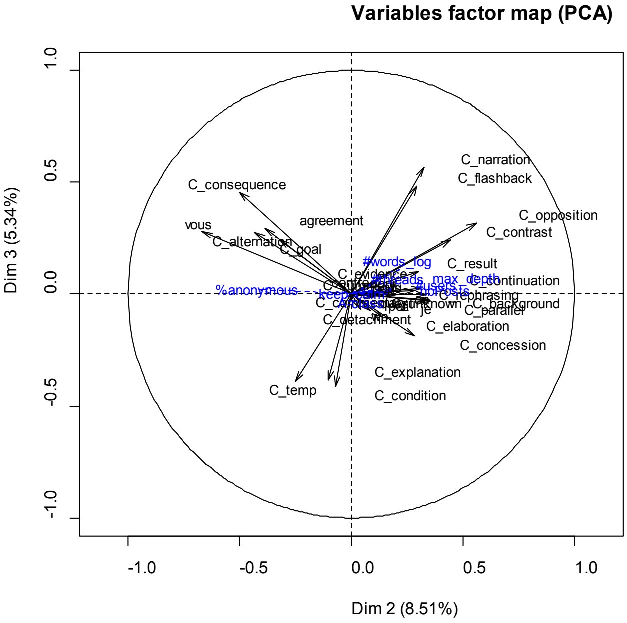


Figure 6: Second factor map (dimensions 2 and 3) resulting from the PCA performed by taking into account the linguistic features. Additional features are shown in blue.

The third dimension, which explains more than 5% of the total variance, opposes threads characterized by a significant presence of narrative relations (at the top of Figure 6), and threads including connectives expressing condition and explanation relations.

A closer look at the threads which are situated at the borders of dimensions 1, 2 and 3 provides a better understanding of the structure of the data, and the profiles of the threads they may relate to. The most extreme threads that dimension 1 opposes are very short ones that are usually made of anonymous posts. Actually, these threads may be described as very poor in terms of interaction, such as in example (1), a thread extracted from the talk page for "Protoplaste".

(1) ***techniques de l'obtention des protoplastes (technical criteria to obtain protoplasts)***

en cours (in progress)

On the other hand, we also found threads containing much more connectives and linguistic cues. Among these, dimension 2 may oppose threads characterized by a significant use of agreement markers as in example (2), to threads resorting to *I* ("je")*,* informal *we ("on")* and connectives expressing opposition, such as in example (3).[14]

(2) ***D'accord*** *pour rapporter les "controverses" scientifiques, mais sans négliger le style cf Wikipédia:Style encyclopédique. (****I agree*** *to report scientific "controversies" but without neglecting the encyclopedic style, see Wikipédia :Style) Les anglais me semblent plus pragmatiques de n'avoir traité que de l'"affaire". Pour résumer restons : neutre, impersonnel, clair, précis, compréhensible, non académique et moins "people". Bien à* ***vous*** *(kind regards).*

(3) ***Par contre, je*** *doute qu'****on*** *puisse "ignorer" l'existence de ce rapport et qu'au minimum, le contenu qui a été diffusé par d'autres media soit admissible mais* ***j'****attends l'avis d'autres wikipédiens à ce sujet. (****However, I*** *doubt that* ***anyone*** *may "ignore" the existence of this report and I think that the material disseminated through the media is admissible but* ***I*** *await the opinions of other Wikipedians on this question.)*

This closer look at threads positioned on the extremities of the factors provides another view of the data, but does not permit us to identify precise and interpretable profiles of conflict threads. The next step is the projection of the annotated conflict threads through the three-dimensional vector space resulting from the PCA.

[14] Example 2 and 3 are extracted from the talk page about the Bogdanoff brothers.

4.3. Annotated conflict threads through the factor map

Figure 7 gives the location of the 215 annotated threads of the Annotated DataSet (Section 3.3) through the factor map resulting from the PCA. It seems that the best dimension for describing conflict threads is dimension 2. Conflict threads (red crosses) appear to be mainly situated on the positive side of this dimension. According to the PCA, these conflicting threads may be defined as those with more *I* ("je"), informal *we ("on")* and connectives expressing opposition and contrast discourse relations, and fewer agreement cues and formal "you."

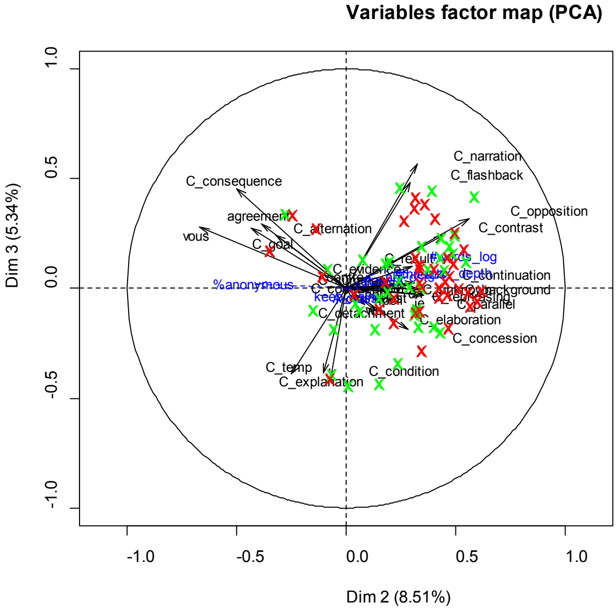


Figure 7: Second factor map (dimensions 2 and 3) with annotated threads located in the PCA and shown by red crosses for conflicting and green crosses for non-conflicting.

Example (4) illustrates one such profile, with the heading and the beginning of three posts of a thread annotated as conflicting in the talk page about Psychoanalysis ("Psychanalyse") (all the significant features are in bold)

(4) *<head> Citation et citations (Lacan et ses exégètes) </head>*

*<post>**je** propose des sources hors du champ de la critique psychanalytique pour exclure les débats LLNDLP ou Onfray etc (**I** propose sources outside the field of the criticism of psychoanalysis to exclude debates on LLNDLP or Onfray etc.) [...]</post>*

> *<post>Apparemment **on** oubli les politesse(s) avec Vous G de gonja…, **j'**invite chacun à jeter un oeil à ceci : (Politeness is not a virtue with you G. de gonja…, **I** encourage everyone to have a look at this) […]</post>*
>
> *<post> 'None' * **Je** ne vois pas bien ce que le commentaire de G de Gonjasufi apporte : personne n'a jamais nié que Lacan ait employé le terme. (I don't really see what G de Gonjasufi's comment provides) **En revanche,** ce que nous disons c'est qu'il ne s'agit pas d'une qualification de la psychanalyse dans son (**In contrast**, what we are saying is that it is not a disqualification of psychoanalysis as a whole) […]</post>*

5 CONCLUSION

We have proposed different ways to explore Wikipedia talk pages in this paper, motivated by the notion that CMC genres are indeed complex objects that challenge our traditional methods, and thus we assume that such objects require different levels of investigation. The profiling step still needs further analysis, but is already quite promising.

The results of the automatic classification show that the features taken into account and the parameters used for detecting conflict talk pages are still fairly inaccurate. In addition, our definition of a conflict discussion should be more specific. Data mining methods and first results in thread profiling give us some leads that must be followed up in this regard, and we are currently exploring relevant features to describe the thread level. We will notably use other categories to characterize talk pages and threads, combining, for instance, the article labels signalling conflicts, the talk page labels and the talk page type. On the linguistic level, the list of connectives and the discourse relation they express must be refined in order to distinguish discourse markers from conjunctions, and to get a better manage handle on polysemy (as for example, 17 connectives are associated with contrast in LexConn, including the very polysemous uses of "but" and "while").

In addition, other interaction features must be taken into account, including, for example, thread headings, timeline and context features. We are also concentrating on the first and the last posts of the threads, which generally play a key role in conflicts arising and being resolved. As such, we are currently annotating speech acts and politeness cues in these posts. Another avenue of investigation concerns the relation between disagreement and conflict: disagreement is quite common on Wikipedia, and although many conflicts arise from a disagreement, all disagreements do not naturally lead to conflict. What are the specificities of such disagreements / such conflicts? One of the main differences between disagreements and conflicts is certainly the presence of verbal violence, and we are currently exploring this question. In any case, it seems obvious that the most pressing need for identifying thread types is to provide a dataset of annotated threads according to interaction, politeness and conflict.

References

Adler, Thomas B., Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, Andrew G. West, 2011: Wikipedia vandalism detection: Combining natural language, metadata , reputation features. *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg. 277–288.

Agarwal, Alekh, Olivier Chappelle, Miroslav Dudik, John Langford, 2011: A reliable effective terascale linear learning system. *JMLR* 15. 1111-1133.

Brandes, Ulrik and Jürgen Lerner, 2007: Revision and co-revision in Wikipedia: Detecting clusters of interest. *Proceedings of International Workshop Bridging the Gap Between Semantic Web and Web 2.0*. Innsbruck, Austria.

Denis, Alexandre, Matthieu Quignard, Dominique Fréard, Françoise Détienne, Michael Baker and Flore Barcellini, 2012: Détection de conflits dans les communautés épistémiques en ligne. *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*. 351–358.

Ferschke, Oliver, Iryna Gurevych and Yevgen Chebotar, 2012: Behind the article: Recognizing dialog acts in wikipedia talk pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 777–786.

Ferschke, Oliver, Johannes Daxenberger and Iryna Gurevych, 2013: A survey of NLP methods and resources for analyzing the collaborative writing process in Wikipedia. Gurevych, Iryna and Jungi Kim (eds.): *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer.

Giles, Jim 2005: Internet encyclopaedias go head to head. *Nature* 438/7070. 900–901.

Goldberg, Yoav and Orwant, Jon, 2013: A dataset of syntactic-n grams over time from a very large corpus of English books. *Proceedings of the Second Joint Conference on Lexical and Computational Semantics* (*SEM).

Herring, Susan, Dieter Stein and Tuija Virtanen 2013: *Pragmatics of computer-mediated communication* 9. Berlin: De Gruyter.

Ho-Dac, Lydia-Mai and Veronika Laippala, 2017: Le corpus WikiDisc, une ressource pour la caractérisation des discussions en ligne. Wigham, Ciara and Gudrun Ledegen (eds.): *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Collection Humanités Numériques. Paris : L'Harmattan. 107–124.

Kanerva, Jenna, Juhani Luotolahti, Veronika Laippala and Filip Ginter, 2014: Syntactic n-gram collection from a large-scale corpus of internet Finnish. *Proceedings of the Sixth International Conference Baltic HLT*.

Kittur, Aniket and Robert E. Kraut, 2008: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 37–46.

Kittur, Aniket and Robert E. Kraut, 2010: Beyond Wikipedia: coordination and conflict in online production groups. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 215–224.

Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton and Ed H. Chi, 2007: He says, she says: conflict and coordination in Wikipedia. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 453–462.

Kittur, Aniket, Ed H. Chi and Bongwon Suh, 2009: What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1509–1512.

Laippala, Veronika, Jenna Kanerva and Filip Ginter, 2015: Syntactic n-grams as key structures reflecting typical syntactic patterns of corpora in Finnish. *Procedia - Social and Behavioral Sciences*. 233–241.

Miller, Nathaniel, 2012: Characterizing conflict in Wikipedia. Mathematics, *Statistics , Computer Science Honors Projects* 25.

Potthast, Martin, Benno Stein and Robert Gerling, 2008: Automatic vandalism detection in Wikipedia. *Advances in Information Retrieval*. Springer. 663–668.

Poudat, Céline and Sylvain Loiseau, 2007: Représentation et caractérisation lexicale des sciences dans Wikipédia. *Revue française de linguistique appliquée* 12/2. 29–44.

Poudat, Céline, Laurent Vanni and Natalia Grabar, 2016: How to explore conflicts in French Wikipedia talk pages? *JADT*. 645–656.

Roze, Charlotte, Laurence Danlos and Philippe Muller, 2012: Lexconn: A French lexicon of discourse connectives. *Discours* 10. 1–15.

Scott, Mike and Christopher Tribble, 2006: *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia, PA, USA: John Benjamins Publishing Company.

Stvilia, Besiki, Michael B. Twidale, Linda C. Smith and Les Gasser, 2008: Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology* 59/6. 983–1001.

Suh, Bongwon, Ed H. Chi, Bryan A. Pendleton and Aniket Kittur, 2007: Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. *Visual Analytics Science and Technology* 2007. IEEE. 163–170.

Sumi, Róbert, Taha Yasseri, András Rung, András Kornai and János Kertész, 2011: Characterization and prediction of Wikipedia edit wars. *Proceedings of the ACM WebSci'11*. Koblenz, Germany. 1–3.

Urieli, Assaf, 2013: Analyse syntaxique robuste du français: concilier méthodes syntaxiques et connaissances linguistiques dans l'outil Talismane. Ph.D. thesis, Université de Toulouse – Jean Jaurès.

Viégas, Fernanda B., Martin Wattenberg and Kushal Dave, 2004: Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 575–582.

Viegas, Fernanda B., Wattenberg, Martin, Jesse Kriss and Frank van Ham, 2007: Talk Before You Type: Coordination in Wikipedia. *40th Annual Hawaii International Conference on System Sciences*. 78–78.

Vuong, Ba-Quy, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw and Kuiyu Chang, 2008: On ranking controversies in Wikipedia: Models and evaluation. *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM. 171–182.

Wilkinson, Dennis M. and Bernardo A. Huberman, 2007: Cooperation and Quality in Wikipedia. *Proceedings of the 2007 International Symposium on Wikis*. ACM. 157–164.

Wulczyn, Ellery, Nithum Thain and Lucas Dixon, 2017: Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee: 1391–1399.

Yasseri, Taha, Robert Sumi, András Rung, András Kornai, János Kertész, 2012: Dynamics of conflicts in Wikipedia. *PloS one* 7/6.

Part 4
Building and processing CMC resources

The development of DOTI (Data of oral teletandem interaction)

Solange Aranha, *Sao Paolo State University*

Paola Leone, *University of Salento*

Abstract

Teletandem[1] (Telles and Vassallo, 2006) is a Voice Over Internet Protocol (VoIP) communicative activity in which two speakers are involved, each of whom is an expert in one language and who wishes to learn the language of the interlocutor. "Virtual meetings" which last one hour are organized weekly; students speak half of the time in their L1, and the other half in the L2. Teletandem is also a growing field of research, and the related data, collected by video-recording the conversations between two participants, are an interesting resource for analysing communication and learning processes. In order to build a teletandem databank (DOTI – Data of Oral Teletandem Interactions), we collected data from Sao Paolo State University at São José do Rio Preto (Brazil: languages Portuguese/English), and from the University of Salento (Italy: languages Italian/English). DOTI is currently composed of about 700 hours of video data from, oral teletandem sessions. The current paper describes: i) the state of the art with regard to developing a databank with video recorded oral sessions, as well as chat conversations; ii) teletandem as an interaction space; iii) different learning scenarios and microtasks that might influence the type of data and, in turn, metadata, in this context.

Keywords: computer mediated interaction, databank, learner corpus, foreign language, learning scenario

[1] Teletandem will be used with capital letters when we refer to the project "Teletandem Brasil – Foreign Language for all" and in lower case letters when we refer to practice/context/session.

1 INTRODUCTION

An interesting field of research in Applied Linguistics is the analysis of the various contexts in which L2 learning occurs, and the impact of all related variables on the development of L2 competence. The use of Information and Communication Technology (ICT) has recently created new opportunities for language learning worldwide, and many telecollaborative projects within universities have emerged in academic areas. Teletandem (Vassallo and Telles 2006), a telecollaborative project on language learning at the university level, is based on a multimodal form of interaction, carried out by the use of Voice Over Internet Protocol (VoIP) and Internet Relay Chat tool, aimed at promoting students' reciprocal learning. Two participants enrol in the activity and speak his/her language of proficiency for half of the oral session period, and for the other half the language he/she is learning. Such practice is based upon tandem principles, proposed by Brammerts in the 1980's (cf. Brammerts 1996), namely autonomy, separation of language and reciprocity. Autonomy implies the possibility that each participant has of organizing his/her own learning experience. Separation of languages means that only one language can be used for the part of the session dedicated to that language.[2] Reciprocity involves respect for the other's learning needs and commitment to practice both languages.

So far, Teletandem has led to many telecollaborative projects within universities, promoting networking for research purposes and the exchange of experiences and best practices. Nowadays, teletandem practice, as proposed by Telles and Vassallo (2006), is carried out in many universities around the globe: (seven in Europe: e.g. the University of Roma Tre, and Southampton University; twelve in the USA: e.g. Georgetown University, the University of Georgia in Athens, and Miami University, and two in South America: the University of Mexico and Cali University).[3]

The significance of such practice for language learning is twofold: it enriches the interactional skills of the participants through incidental learning, and grants them the possibility of sharing meaningful experiences in a dialogical and narrative path, which makes room for emphasizing relevant cultural characteristics. The teletandem experience allows participants to advance their linguistic-communicative competence as well as to expand their curiosity, to promote new themes, to question prejudices, to jeopardize discourses, and to discuss the interactional styles that characterize their cultures. Furthermore, it somehow establishes what Linnell (2009) calls the "sociocultural ecology" of linguistic learning, because the values and specificities of various cultures are not mediated by pedagogical materials and techniques, as is the case in a traditional language class (Telles, Zakir and Funo 2015). Because of this trait, teletandem implies a new type of mobility, achievable thanks to the use

[2] Code-switching is, however, possible when it is aimed to facilitate conversations and messages (Leone 2009).

[3] Information gathered among participating universities and their partner institutions (see www.teletandembrasil.org).

of ICT, i.e. *virtual mobility*, which works as a new way of "migration," even if temporary, to another country (Leone 2016). On the academic level, virtual mobility also supports future exchange programs (e.g. ERASMUS+).

The positive impact of teletandem on language learning is a good basis for expecting a gradual, but sustained increase in its use in higher education. This trend calls for further empirical research and implies a high demand for video/audio data.

Teletandem data, collected and filed using standard protocols which allow for systematic research, is required by users of the Teletandem network. In order to achieve this, we developed an arrangement between UNESP/SJRP and Unisalento with the ambition of filing and organizing existing data (Italian/English and Portuguese/English) in a databank composed of chat texts and video-recorded oral teletandem sessions, named DOTI (Databank of Oral Teletandem Interactions) (Aranha and Leone 2016).

The current paper is organized as follows. Section 1 relates teletandem to other CMC genres, reviews the literature on metadata in CMC corpora, and ends with the research questions. Section 2 describes the research context and DOTI project. Section 3 illustrates the main concepts used when defining the metadata for L2 interactions in pedagogical contexts (e.g. interaction space, learning scenario). Section 4 describes how those concepts are combined into DOTI metadata. Section 5 then concludes the paper.

1.1 Teletandem in relation to other CMC genres

Communication is generally synchronous during teletandem sessions, and quasi-synchronous when chat is employed. The typology of teletandem communication is defined both as telecollaboration and online intercultural exchange, according to Lewis and O'Dowd (2016), who intertwine the terms into a single meaning.

Teletandem practice implies multimodal spoken communication, and thus the data are both visual and vocal.[4] It provides a context for autonomous language learning, and is employed in institutions and sometimes even integrated into language courses.

Since during teletandem sessions the participants talk while keeping in mind a double focus, the language used and the discussion themes (Apfelbaum 1993; Bange 1992; Leone 2014a), teletandem is defined as "conversation for learning" (Kasper 2004; Kasper and Younhee 2015). As a pedagogical context, when collecting data for research purposes, the features of the learning situation need to be described

[4] This type of data generates problems with regard to privacy, which have been dealt with by asking the participants to sign a consent form.

and the characteristics of groups and participants must be recorded. For example, we must take into account the organization of a teletandem activity (e.g. length of the program), the learning situation (e.g. the presence or absence of a task), students' sociodemographic profiles (e.g. gender, age), because research shows that these properties might affect how participants interact (see Rampazzo 2017).

The main features of teletandem sessions (i.e. being spoken multimodal communication carried out in a learning institution) differentiate this form of telecollaboration from other CMC exchanges, such as conferencing systems communication (conference systems with text, etc.), email, discussion forums, blogs, tweets, and audio-graphic systems. These are written exchanges and they are not seen as empowering the users' language and cultural skills and, in most cases, they are not so strongly linked with a learning institution.

To the best of our knowledge, few multimodal data have been used for databank and corpora building. For instance, Chanier et al. (2014) describe the Corpus de Communication Médiée par de Réseaux (COMERE)[5] which covers different genres. Chanier and Wigham (2016) describe procedures used with the Learning and Teaching Corpora (LETEC), based on previous experiences with the Mulce projetc. In both cases, learning environments are currently scheduled to be included in the related corpus. Other corpora described in the recently published volume Wigham and Ledegen (2017) do not include either data from computer mediated learning contexts, nor spoken interaction data. In the pedagogical domain, Mangenot and Soubrié (2010) discuss the development of a learning objects' databank as an open resource, highlighting the importance of "task" as a unit for describing teaching practices. The shortage of such data is probably due to the fact that such learning experiences are recent, and the transcription procedure is still time consuming, even if transcription software (e.g. Transana, ELAN) now supports this. Nevertheless, according to Chanier and Wigham (2016: 216):

> Studying online learning, in order to understand this specific type of situated human learning (Learner Computer Interactions (LCI)) and/or evaluate pedagogical scenarios or technological environments, requires accessibility to interaction data collected from the learning situation.

1.2 CMC and metadata

Metadata are "management tools" (Autayeu, Giunchiglia and Andrews 2010) which allow users to process and select relevant data. For browsing the web and looking for a journal article, for instance, we can write two or more words of the paper's title.

[5] See https://corpuscomere.wordpress.com/

The titles of papers and books, keywords or business catalogues' names are manually generated natural language metadata. Conversely, the date of a picture is automatically generated by the camera. Natural language or standardized metadata are listed in different datasets, each including "web directory category names, business catalogue category names, thesauri and subject headings" (Autayeu et al. 2010). Datasets can be more general or specific to a certain domain. For example, DMoz or Open Directory Project is quite general, very large and used as a directory for classifying all sites, including well-known search engines such as Google. The Dublin Core is "a vocabulary of fifteen properties for use in resource description".[6] On the other hand, the Text Encoding Initiative (TEI; Burnard and Bauman 2013) encodes metadata for machine-readable texts and is used in the field of humanities, social sciences and linguistics, while LOM (Learning Object Metadata) and SCORM (Sharable Content Object Reference Metadata) are applied in that of pedagogy.

Metadata are characterized by "atomic concepts," with Autayeu et al. (2010) noting that the query "Bank and personal detail of George Bush" is made of four atomic concepts: bank, personal, detail and "George Bush". "Atomic concepts" are thus used to create complex concepts (Autayeu et al. 2010).

Most standardized metadata need to be extended in order to encompass more recent computer mediated texts and learning experiences. For this reason, careful and focused illustration of different computer mediated learning environments and practices is needed to create a model stemming from the highlighted characteristics (Mangenot and Soubrié 2010).

1.3 Research questions

Much research has been carried out within the teletandem learning context (cf. www.teletandembrasil.org). The list of published works in this area emphasizes the coverage of multiple theoretical perspectives and presents a fertile field for understanding how telecollaboration may enhance participants' competences. If the wide inventory of pedagogical experiences and scientific studies has enriched the original project with new interpretations and perspectives, one current task is to better understand how the initial proposal by Telles and Vassallo (2006) has been actually carried out in various international contexts.

The present research is a first step in this direction and tries to meet the urgent need to reflect upon what has been done, starting from two academic contexts: UNESP (Sao Paolo State University) and Unisalento (Universidad del Salento). Such a simplified but comprehensive description is also used to describe the

[6] See http://dublincore.org/

amount of data which has already been generated and recorded, and which can be further collected and filed within the project.

Bearing the above in mind, the current study aims at answering the following questions:

1) How can teletandem exchanges be encoded in standardized metadata?
2) What are the common characteristics of the learning contexts as they are developed in the Brazilian and the Italian higher educational institutions examined in this work?
3) Which metadata allow the identification of online interactions with learning purposes?

We intend to follow a common course to establish metadata for describing DOTI, as well as take a first step towards the definition of a protocol for collecting further data, and transcribing existing data.[7] This work has two main aims: a) to enhance collaborative and shared research among Teletandem network members; and b) to expand and reinforce the network between professors and mediators.

2 RESEARCH CONTEXT

Teletandem practice may occur within a language course, as part of a university program, as seen in some groups at Sao Paolo State University at São José do Rio Preto (UNESP/SJRP), or may occur as an elective activity, thus voluntary, as in the University of Salento. The former is coined institutional integrated teletandem, and the latter institutional non-integrated teletandem, according to Aranha and Cavalari (2014). Depending on the agreement between the two partner institutions that carry out a teletandem program, the computer mediated oral sessions may or may not be followed by other learning activities or tasks.

In our universities, teletandem practice has been adopted with students from different majors who study various foreign languages. Their level of L2 linguistic competence varies, although this is not taken into account if a person wants to join the Teletandem project. The tasks that can occur in diverse learning scenarios are adjustable to distinct levels of competence. The characteristics of a teletandem course can be described following: a) a general framework for identifying the communication setting (Mangenot and Soubrié 2010); b) a general framework for outlining both pedagogical and learning practices based upon teletandem at UNESP-SJRP and at Unisalento.

[7] 10% of existing data (Portuguese/English) has already been transcribed by using Transana.

2.1 UNESP and Unisalento specific characteristics of Teletandem Oral Sessions and Mediation

Teletandem practice displays a learning scenario that carries a coherent and complex activity framework – a TOS (Teletandem oral session) and teletandem mediation session – which consists of different pedagogical and didactic collaborative events (Mangenot 2008, Foucher 2010) aimed at developing students' plurilingual and pluricultural competences (Candelier et al., 2012, Leone et al. 2015).

Teletandem pedagogical scenarios (TTPS) are coherent with the following principles: (i) collaboration: the tasks are intended to be developed collaboratively; ii) interaction: communicative exchanges and oral sessions favour the development of learning strategies and autonomy, and also increase inter-comprehension skills.

TTPSs have varied purposes, which can be synthetized into four points. The first has the intent of preparing students to participate actively in (computer mediated) oral interactions with a proficient speaker, and be aware of all the linguistic and cultural strategies that such a practice involves (Aranha and Leone 2016). The second aims at improving self-evaluation and awareness about one's own learning skills and abilities.

The objectives are, therefore, to make the participants more autonomous in their learning and then to develop their "learning how to learn skill." Or better yet, to be aware of how to study and improve/articulate knowledge and competences outside of formal contexts and without teaching guidance. The third point, the scenarios based on teletandem, has the purpose of promoting the use of digital technology to facilitate one's learning capacity efficiently and flexibly. In this sense, the participants may take advantage of the great potential of new technologies, considered as key-knowledge tools in the Recommendation of the European Parliament and of the Council of 2006.[8] Finally, through intercultural discourse, TTPSs give the participants opportunities to strengthen their positive attitudes towards other people, ideas, experiences and cultures.

Teletandem is characterized by two macrotasks: mediation sessions and teletandem oral sessions (TOS). At UNESP/SJRP, mediators are both professors involved in the Teletandem project and graduate students (Masters and Doctorates) who investigate telecollaboration practices. At Unisalento, mediators are language instructors and professors involved in the project.

The linguistic and cultural exchanges between mediators and students happen within a social cultural perspective and allow each and every individual to advocate cultural identities in a broad sense. During mediation sessions, participants interact and "do

[8] Recommendation of the European Parliament and of the Council of 18th December 2006 on key competences for lifelong learning http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006H0962

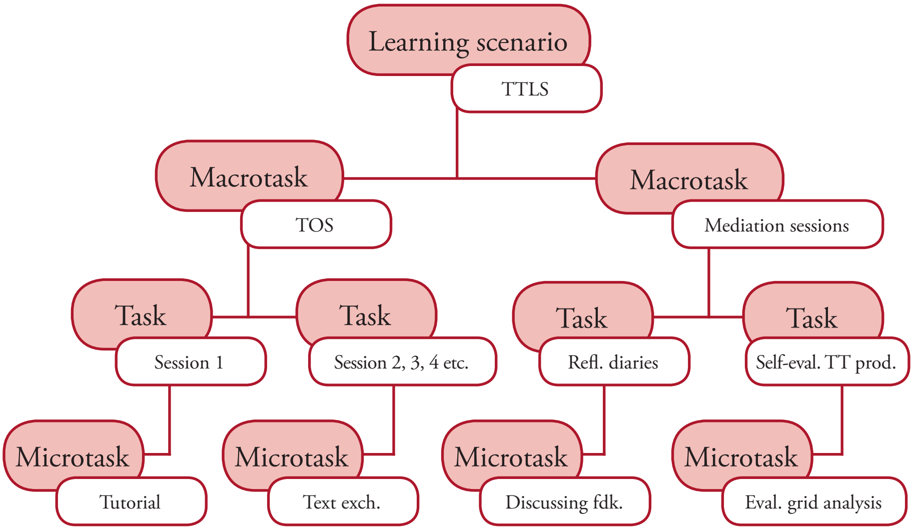


Figure 1: The organization of a pedagogical scenario based on Teletandem.

not simply shift between competing meanings to find the correct one, but, instead, navigate a constantly changing and emerging hermeneutic environment." (Feito 2007: 3). Teletandem principles and environments allow students to negotiate meaning, discuss points of view, envision new knowledge, and present cultural approaches and perspectives.

The length of learning scenarios based on the teletandem context is variable (from six to 15 sessions, of about one hour each, depending on the needs of each group). Online meetings are video recorded using Evaer[9] and consent forms are signed to protect the privacy of participants and allow the research process to be controlled (Aranha et al. 2015, Mackey and Gass 2005: 330). As shown in Fig. 1, for describing our learning scenarios we use the terms *tasks*, *macrotasks* and *microtasks* to present the complex environment in which teletandem practice takes place. As argued by Mangenot and Soubrié (2010), the concept of task is essential for developing descriptors of more recent teaching practice. *Macrotasks* are tasks with a larger scale and scope, involving teletandem sessions and mediation sessions. *Microtasks* are short duration tasks with reduced scopes, and these support overall task implementation.

In TTPS, technology is used both to carry out oral exchanges (via VOIP technology) and to develop access to documents and assigned activities within each scenario (e.g. Moodle, Google Docs). Technologies are also essential for some *microtasks*, such as collaborative writing, or recording and analysing recorded videos. TTPS offers, therefore, the use of technology for communication (Computer Mediated

[9] Evaer is an easy-to-use and low-cost software package for recording Skype calls. See http://www.evaer.com/

Communication: CMC) and for assisting students in their learning process (Computer Assisted Language Learning: CALL).

2.1.1 Pedagogical scenarios at UNESP/SJRP

At UNESP/SJRP, students who participate in the Teletandem Based Learning Scenario (TTPS) are from different linguistic levels and majors. In the integrated scenarios, students are majoring in language courses and aiming to be either be language teachers or translators. In non-integrated ones, undergraduates come from different courses, and their alleged level of proficiency is self-established, i.e. using the grid from the Common European Framework (CEFR), students place themselves in one level when they answer a questionnaire before the sessions begin. TOSs are fed by texts exchanged between partners and guided by some pedagogical tasks. Free conversation also occur.

2.1.2 Learning scenario at Unisalento

TTPS are institutionalized but non-integrated at Unisalento, and credits are awarded for participation and completion of tasks. Students who participate in learning scenarios based on teletandem attend Bachelor's and Master's degree courses, and specialize in one or more foreign languages (e.g. English, French, and Arabic). As at UNESP, language competence is self-established by students using CEFR grids for evaluation.

TOSs are (currently) characterized by free conversations or by discussions on specific topics (e.g. youth life-styles in the students' countries; Leone 2016).

3 METHODOLOGY

Teletandem practices in the two higher educational contexts had to be shared so that we defined the pedagogical characteristics of such learning practices, trying to uncover those which could allow the description of the whole process. For describing Teletandem sessions, the notion of "interaction space", as developed in Chanier et al. (2014), was used. Since teletandem is a pedagogic and communicative practice in which students and professors are both involved, the concept of learning scenario (Mangenot 2008, Foucher 2010) must also be present as an epistemological frame, useful for characterizing various sequences and events that determine it.

The components of different learning scenarios (e.g. characteristics of participants, number of sessions), microtasks (e.g. methodological procedures, verbal and non-verbal input), as well as the properties of the interaction space, within which the various forms of technology mediated communication are performed for completing the learning scenario, are all considered. Concerning the pedagogical implementation of Teletandem, we developed a didactic description that is a first step in the process of producing standardized metadata.

In the following subsections we examine the concepts of interaction space, learning scenario and task in more detail.

3.1 Interaction space

The notion of Interaction Space (IS) (Chanier et al. 2014) derives from TEI and aims at characterizing distinct genres within CMC (focused on written communication, such as Facebook posts), and is defined as an abstract concept "located in **time […]** where interactions between **a set of participants** occur within an **online location**".

As described by Chanier et al. (2014), IS entails concepts related to *Interaction Space* itself and to *CMC environment*. The first includes *participants*, i.e. a set of groups or individuals, *time frame*, i.e. the beginning and ending time, and *online location*. *CMC environment* gives access to online communication, and it can be *monomodal* or *multimodal*. *Modality* is "a specific way for realizing communication" (Chanier et al. 2014: 6), and it affords a specific *interaction type* (e.g., email). Modality can also be described in terms of "semiotic resource", that is the *mode* (i.e., text, speech and non-verbal) which realizes communication. Finally, *time* can be synchronous or asynchronous.

3.2 Pedagogical and learning scenario

For describing online learning situations, Chanier and Wigham (2016: 222) use the term pedagogical scenario. A pedagogical scenario describes:

a) the whole environment (such as a Learning Management System (LMS);
b) the various roles of participants (teachers, learners, experts and the role of each participant during the course);
c) each course activity and the role of each participant during this;
d) how activities are sequenced;
e) the resources that will be used and produced; and
f) the instructions that govern the learning activities.

A pedagogical scenario may consist of a learning scenario and a tutoring/supervision scenario. Using Chanier and Wigham's terminology, DOTI is, so far, composed only of one learning scenario (which we call macrotask), although the Teletandem project also considers a tutoring/supervision scenario, as described below.

3.3 Task as an essential concept

For us, a task is an essential unit for defining all the activities carried out in a learning scenario, and thus a task can be considered one of our "atomic concepts" (see par. 1.2). As above mentioned, the concept of a task is also essential to describe specific activities, such as microtasks.

Many definitions of a task appear in the literature, and all of them imply that any effective task integrated in formal educational programmes must be communicative, meaning-focused and linked to the real (i.e., beyond the classroom) use of that language (Skehan 1998). According to Gonzales-Llore and Ortega (2015), the primary focus of a task is on meaning. Even if there is a preplanned language learning goal, part of the learning must be incidental, and any particular language focus should be hidden from the learners, or 'implicit,' at least for a good part of the task module. Long (2015: 3470) emphasizes that classroom tasks[10] should be based on students' learning needs, definable by the activity they "need, or will need, to do in the L2," which Long terms "target task" (Long 2015: 3479). Gonzales-Llore and Ortega (2014: 5) mention holism as one definitional feature of a task in the context of technology-and-task integration:

> a task draws on real-world processes of language use, integrating form-function-meaning; this definitional feature goes to notions of 'authenticity' and 'real-world relationship'.

In our experience, we believe that autonomy in L2 learning is crucial for our students' future professions, and thus the main task of mediation sessions is based on a target task, which is "self-evaluating one's interaction skills and analysing the learning process." In fact, we aim at developing students' abilities to self-analyse their own learning process and the communicative use of "the lexis, collocations, pragmatics, skills, genre and registers" (Long 2015: 3466) necessary for reflecting on their own L2 production, learning process and needs.

Following the framework by Ellis (2003) and Gonzales-Llore and Ortega (2015), the two tasks of the learning scenarios in the two higher educational contexts examined in this work, i.e. diaries and self-evaluating interaction skills, can be described based on the following design features (see also Mangenot and Soubrié 2010):

[10] Although Long consider "classroom" tasks and our context is not within a classroom, we argue that the concept also applies to telecollaborative practices.

1. Goal (intended as the general purpose of the task). The task plan must offer a language-and-action experience, which means the task must entail (a) some communicative purpose (i.e. considering students' needs and wants) engineered by means of gap in information or some element that encourages language use that involves informational transfer; and (b) some outcome, resulting from task completion, including communicative outcomes (e. g. the production of an oral or written message, the accomplishment of a desired perlocutionary effect on interlocutors or on the world) and /or non-communicative outcomes (securing a flight booking, producing a plan, gathering knowledge, playing/winning a game, and so on). The goal is the development of autonomy in L2 learning.
2. Input, which may mean the verbal and non-verbal information provided for the task: websites, tutorials, previous learning experience, epiphanies, diaries, teletandem session video-recordings and the CEFR evaluation grids.
3. Conditions are how the information is provided. Normally students do not share the same information. For instance, in "self-evaluating interaction skills," each student does not know which video sequence his/her partners are going to show and comment on. In diaries, one-to-one feedback is given by the professor or mediator in charge and, although much information from the diaries is used for the group mediation meeting, much is personal and directed to one individual.
4. Procedure (e.g. group work vs. pair work; planning time vs. no planning time), at both UNESP/SJRP and Unisalento students work individually during the TOS learning scenario. Afterwards, at UNESP, they share their views in the reflexive diaries, which may be used by professors for classroom and mediation purposes. At Unisalento each student self-evaluates their production and discusses it with the mediator, as well as with thers.
5. Outcomes. Diaries are the products at UNESP/SJRP, while at Unisalento the focus is on an oral discussion of the experience supported by a presentation file. For both tasks, the process of the linguistic interaction and the cognitive activity generated by the task have a strong educational value.

4 DOTI characteristics and metadata

DOTI is composed of around 700 hours of teletandem oral sessions,[11] one of the learning scenarios described above. The majority of these sessions were collect-

[11] At UNESP, the texts produced within the macrotasks, tutorials, questionnaires and reflexive diaries are part of another databank.

ed from the Brazilian university, carried out in Portuguese/English. Unisalento provided fewer recorded oral sessions, with the TOSs being in Italian/English. However, the fact that this data is unbalanced in terms of number of hours for each pair of languages should not be seen as a weakness, as DOTI is ultimately intended as a multilingual databank.

Due to the attributes of DOTI, the databank will provide input to answer the following types of research questions: What are the differences between chat and video synchronous communication? What are the aspects that distinguish chat and oral communication in a learning environment and other contexts (e.g. among friends)? What are the distinctive features of teletandem oral session in relation to other types of oral communication between native and non-native speakers? What are the typical features of metalinguistic sequences in teletandem oral sessions and other virtual contexts (e.g. forums)? Which are the genres used for teletandem interactions in various modalities and microtasks? Which genres are typical of telecollaborative practice? Are genres related to learning scenarios?[12] How do the genres that occur within a teletandem context relate to cultural and linguistic learning?

In sections 4.1 and 4.2 we present metadata concerning the interaction space and the learning scenario. The former shows the general characteristics of teletandem oral sessions, and the latter presents a rough outline of pedagogical issues related to the formative path based on Teletandem.

4.1 Teletandem as an interaction space

In relation to the interaction space, Teletandem is characterized in terms of *participants*, *place/institution* and *time frame*. The participants of TOSs will be two students who want to learn the language of his/her partner; the institutions may be UNESP and UGA (University of Georgia), or UNISALENTO and other British or American universities; *place/institution* records the names of the institutions involved; *time frame* will include information on the semester/year, number of sessions and duration of each session.

In relation to *technology environment*, teletandem is multimodal (visual, oral and written), synchronous as opposed to asynchronous online communication (e.g. blogs). Moreover, the *language* used (e.g. English and Italian, Portuguese and English) will also be specified.

[12] Rampazzo's thesis (2016) shows that the Initial Teletandem Oral Session, as a genre, is dependent on the related learning scenarios.

4.2 Pedagogical scenario

The descriptors will be: *pedagogical scenario*, *macrotasks* (i.e. TOS and mediation sessions), *task* (e.g. learning diaries) and *microtask*, and thus the metadata subfields for the two universities examined here will be different. For example, at UNESP/SJRP teletandem is integrated in a course syllabus, while at Unisalento it is not. Because of this, integrated or non-integrated modalities are also taken into account. If it is non-integrated, then any credits that are awarded should also be included in the data.

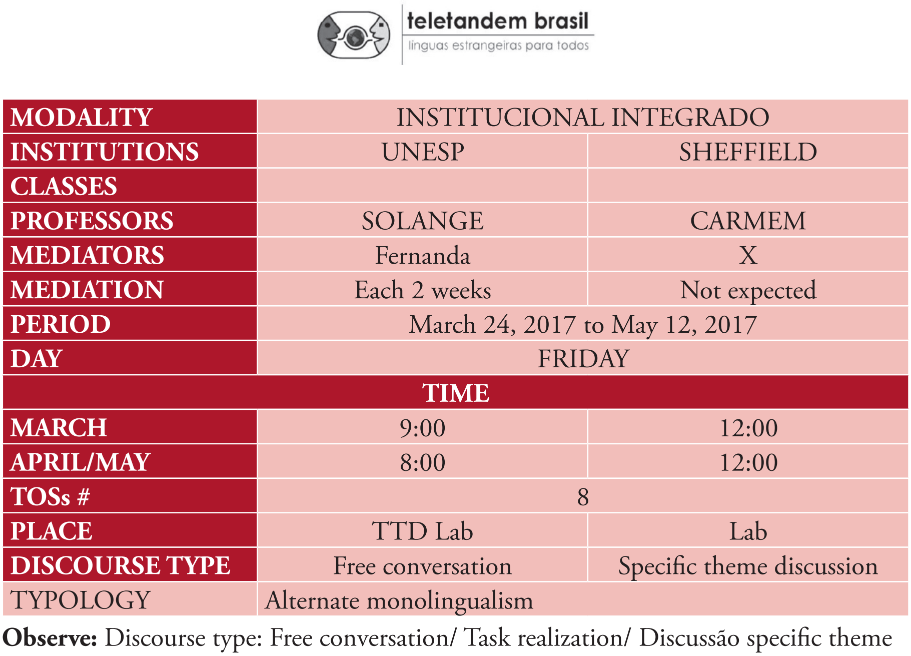
Concerning the learning scenario, all the information combined in the following fields and subfields are considered: the *university curriculum* – with an integrated or non-integrated modality; *time frame*, indicating when and for how long TTPS happened. *Pedagogical scenario* (Fig. 1) also entails the *number* of *macrotasks* and *typology* (e.g. teletandem sessions and mediation sessions). In the following section, we will focus on metadata concerning the Teletandem macrotasks, while mediation macrotasks are not considered since they are currently not part of DOTI.

Teletandem metadata clarifies characteristics related to the learning scenario and teletandem sessions. For the learning scenarios we created a template (Fig.2) that includes information on: 1) learning scenario modality (i.e. integrated, non-integrated); 2) institutions involved; 3) students' majors; 4) professors; 5) mediators; 6) periods of mediation; 7) length of teletandem activity; 9) number of interactions; and 10) place.

Teletandem sessions are described considering, first of all, the participants, based on their sociodemographic characteristics and university curricula. With regard to the CMC environment, Teletandem is multimodal. The mode, i.e., the semiotic source, is text (chat), speech and non-verbal. The interaction type is oral. Finally, time is synchronous (video-conference) and quasi-synchronous (chat).

In terms of TOS, we created another document that includes sociodemographic characteristics, including information about each participant, i.e. *Major*, *Gender* (F or M and Other), and *Alleged Language Competence level* in L2. The pedagogical characteristics of Teletandem are described in terms of *task* and *discourse type* (e.g. free conversation; discussions about a specific theme; development of a task).

The description of a task will include the goal, input, conditions, and the related procedures will also be described.



teletandem brasil
línguas estrangeiras para todos

MODALITY	INSTITUCIONAL INTEGRADO	
INSTITUTIONS	UNESP	SHEFFIELD
CLASSES		
PROFESSORS	SOLANGE	CARMEM
MEDIATORS	Fernanda	X
MEDIATION	Each 2 weeks	Not expected
PERIOD	March 24, 2017 to May 12, 2017	
DAY	FRIDAY	
TIME		
MARCH	9:00	12:00
APRIL/MAY	8:00	12:00
TOSs #	8	
PLACE	TTD Lab	Lab
DISCOURSE TYPE	Free conversation	Specific theme discussion
TYPOLOGY	Alternate monolingualism	

Observe: Discourse type: Free conversation/ Task realization/ Discussão specific theme

Figure 2: Document for describing a teletandem learning scenario at UNESP.

5 CONCLUSION

The study examined a specialized segment of computer mediated research, as collecting, organizing and sharing spoken oral data for language learning is an emerging field of research in CALL. More specifically, this study aimed to develop a databank, named DOTI, composed of approximately 700 hours of TOS, and presented several descriptors generated from two key concepts: Interaction Space and Learning Scenario. The former places DOTI within a broader context that includes resources and research on other forms of CMC (such as Facebook and Twitter). The latter is, instead, used to outline the distinctive features of the academic and educational contexts in which teletandem is practiced. When defining metadata, the concept of task, a unit for describing the learning scenario, proved to be significant. Moreover, the metadata used for the learning scenario need to be developed into more standardized forms.

Since every year new partnerships are formed, a growing body of experience can be used to define the agreements that occur between new partner institutions. This first step of this study at creating guidelines for developing the proposed databank will help other researchers to develop more reliable tools for future research. For this reason, the proposed metadata will also be used to establish a protocol of collecting and filing new data. The protocol will be used to: a) save time in collecting data by members of the network; b) share collecting and transcription methodologies; c) enhance the use of sound, scientific procedures. Once the databank is transformed into a LETEC (Learning and Teaching Corpora) corpus, the data can then be interrogated by multiple researchers and for various purposes.[13]

References

Apfelbaum, Birgit, 1993. *Erzählen im Tandem. Sprachlernaktivitäten und die Konstructio eines Diskursmusters in der Fremdsprache (Zielsprachen: Französisch und Deutsch).* Tübingen: Narr.

Aranha, Solange and Spatti Cavalari, 2014. A trajetória do projeto Teletandem Brasil: da modalidade institucional não-integrada à institucional integrada. *The ESPecialist* 35/2, 70–88.

Aranha, Solange and Paola Leone, 2016. DOTI: Databank of Oral Teletandem Interactions. Jager, Sake and Malgorzata Kurek (eds.): *New directions in telecollaborative research and practice: selected papers from the second conference on telecollaboration in higher education*. 327–332.

Aranha, Solange, Lidiane Luvizari-Murad and Augusto César Moreno, 2015. A criação de um banco de dados para pesquisas sobre aprendizagem via teletandem institucional integrado (TTDII). *Revista (Con) Textos Linguísticos* 9/12. 274–293.

Autayeu, Aliaksandr, Fausto Giunchiglia and Pierre Andrews, 2010. Understanding Natural Language Metadata. http://eprints.biblio.unitn.it/1836/1/026.pdf. (Last accessed 29 June 2017.)

Bakhtin, Mikhail, 1986. *Speech genres and other late essays. A selection of essays from the Russian original "Estetika slovesnogo tvorchestva" [1979].* Austin: University of Texas Press.

Bange, Pierre, 1992. A propos de la communication et de l'apprendissage en L2, notamment dans le forme institutionnelles. *Aile* 1. 53–55.

Brammerts, Helmut E., 1996. Tandem language learning via the internet and the International E-Mail tandem network. Little, David and Helmut E. Brammerts (eds.) *A guide to language learning in tandem via the internet.* CLCS Occasional Paper no. 47. Dublin: Trinity College Dublin. 9–22.

Burnard, Lou and Syd Bauman, 2013. TEI P5: Guidelines for electronic text encoding and interchange. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines. (Last accessed 29 June 2017.)

[13] We would like to mention that one of authors has been awarded a grant (FAPESP #2016/18705-9) that will help fund the organization proposed for developing DOTI. Moreover, on the Brazilian side, the various other microtasks (diaries, texts exchanged between partners, questionnaires) are also part of the databank.

Candelier, Michel, Antoinette Camilleri-Grima, Véronique Castellotti, Jean-Francois de Pietro, Ildikó Lörincz, Franz-Joseph Meissner, Anna Schröder-Sura and Artur Noguerol, 2012. *CARAP. Un Cadre De Référence pour les Approaches Plurielles des langues et des Cultures. Compétences et ressources.* CELV (Centre Européen pour le Langue Vivantes). http://apfmalte.com/uploads/CARAP.pdf. (Last accessed 10 August 2017.)

Cavalari, Suzi M. and Solange Aranha, in preparation. *Implications of Teletandem integration into foreign language programs: insights on the teacher-mediator's role.*

Cavalari, Suzi M. and Solange Aranha, 2016. Teletandem: integrating e-learning into the foreign language classroom. *Acta Scientiarum: Language and Culture* 38/4. 327–336.

Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, Djamé Seddah, 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics* 2/29. 1–30.

Chanier Thierry and Ciara Wigham, 2016. A scientific methodology for researching CALL interaction data: Multimodal LEarning and TEaching Corpora. Caws, Catherine and Marie-Joseé Hamel (eds.): *Language-Learner Computer Interactions: Theory, methodology and CALL applications*, Johns Benjamins.

Feito, José A., 2007. Allowing not-knowing in a dialogic discussion. *International Journal for the Scholarship of Teaching and Learning* 1/1. 1–11.

Foucher, Anne-Laure, 2010. *Didactique des Langues-Cultures et Tice : scénarios, taches, interactions.* Université Blaise Pascal - Clermont-Ferrand II.

Gonzales-Llore, Marta and Lourdes Ortega, 2015. *Technology-mediated TBLT. Researching Technology and Tasks.* Amsterdam/Philadelphia: John Benjamins.

Kasper, Gabriele, 2004. Participant orientations in German Conversation-for-Learning. *The Modern Language Journal* 88/4. 551–567.

Kasper, Gabriel and Kim Younhee, 2015. Conversation-for-Learning: Institutional Talk Beyond the Classroom. *The Handbook of Classroom Discourse and Interaction*. London: Wiley Blackwell. 390–408.

Leone, Paola, 2016. Collaborare per capirsi nel contesto di apprendimento teletandem. *Parlare insieme. Studi per Daniela Zorzi*. Bologna: Bononia University Press. 191–206

Leone, Paola, 2014a. Focus on form durante conversazioni esolingui via computer. *Varietà dei contesti di apprendimento linguistico*. Milano: Officinaventuno. 169–187

Leone, Paola, 2016. Migrazioni virtuali: teletandem per l'apprendimento di una L2. *Incontri*. 48–65.

Leone, Paola, 2009. Processi negoziali nel corso di scambi comunicativi mediati dal computer. *Oralità/scrittura. In memoria di Giorgio Raimondo Cardona*. Perugia: Guerra. 389–412

Leone, Paola and João Telles, 2016. The Teletandem network. *Online Intercultural Exchange: Policy, Pedagogy, Practice*. London: Routledge. 243–248.

Leone, Paola, Alessandro Bitonti, Donatella Resta and Bianca Sisinni, 2015. *Osservazione di classe, insegnamento linguistico e (tele)collaborazione.* Firenza: Franco Cesati.

Lewis, Tim and Robert O'Dowd (eds.), 2016. *Online Intercultural Exchange: Policy, Pedagogy, Practice.* London: Routledge.

Linell, Per, 2009. *Rethinking Language, Mind, and World Dialogically: Interactional and Contextual Theories of Human Sense-making.* Charlotte, NC: Information Age Publishing.

Long, Michael, 2015. *Understanding second language acquisition.* Oxford: Oxford University Press.

Luvizari-Murad, Lidiane H., 2011. *Aprendizagem de alemão e português via teletandem: um estudo com base na Teoria da Atividade.* Unpublished PhD Thesis.

Mackey, Alison and Susan M. Gass, 2005. *Second language research: methodology and design.* Mahwah, NJ: Lawrence Erlbaum.

Mangenot, François, 2008. La question du scénario de communication dans les interactions pédagogiques en ligne. *Jocair (Journées Communication et Apprentissage Instrumentés en Réseau*. 13–26.

Mangenot, François and Thierry Soubrié, 2010. Créer une banque de tâches Internet: quels descripteurs pour quelles utilisations? *La tâche comme point focal de l'apprendissage.* Clermont-Ferrand.

O'Dowd, Robert, 2013. The competences of the telecollaborative teacher. *The Language Learning Journal*. 194–207.

Rampazzo, Laura, 2017. *Gêneros textuais e telecolaboração: uma investigação da sessão oral teletandem inicial*. Dissertação (Mestrado em Estudos Linguísticos). Universidade Estadual Paulista "Júlio de Mesquita Filho", campus de São José do Rio Preto. São José do Rio Preto.

Skehan, Peter, 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.

Telles, João and Maria L. Vassallo, 2006. Foreign language learning in-tandem: Teletandem as an alternative proposal in CALLT. *The ESPecialist* 27/2. 189–212.

Telles, João A., Maisa de Alcântara Zakir and Ludmila B.A. Funo, 2015. Teletandem and culture-related episodes. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 31/2, 359–389.

Vygotsky, L.S, 1978. *Mind in Society.* Cambridge, MA: Harvard University Press.

Wigham, Ciara and Ledegen Gudrun (eds.), 2017. *Corpus de communication médiée e par les réseaux: construction, structuration, analyse*. L'Harmattan, Collection Humanités Numériques.

Part-of-speech tagging for corpora of computer-mediated communication: A case study on finding rare phenomena

Michael Beißwenger, *University of Duisburg-Essen*
Tobias Horsmann, *University of Duisburg-Essen*
Torsten Zesch, *University of Duisburg-Essen*

Abstract

The paper reports on experiments in the adaptation of part-of-speech (PoS) tagging technology for written, interactional discourse retrieved from social media genres (*computer-mediated communication*, *CMC*). Starting from an overview of related approaches, we give a summary of the results and discuss lessons learned from a community shared task on PoS tagging German CMC conducted in 2016. These results suggest that further effort should be put into the development of solutions for phenomena which, one the one hand, are too sparsely represented in data samples that could be used for training tagger models, but, on the other hand, are of special interest for the annotation of linguistic corpora. We present a case study in which we used a PoS tagger to find one particular phenomenon of that type, namely German verb-pronoun contractions, in chats and tweets. Whereas the adoption of over- and undersampling strategies to artificially enhance the frequency of the phenomenon in the training data does not lead to significant improvements, the choice of the tagger together with the expansion of the training data with relatively small amounts of additionally labelled instances turns out to be a promising way to let the tagger learn the local word context, and thus improve the recall of the phenomenon in focus while sustaining a high level of precision.

Keywords: CMC, social media, NLP, annotation, PoS tagging

1 INTRODUCTION

This paper reports on experiments in the adaptation of part-of-speech (PoS) tagging technology for written, interactional discourse retrieved from social media environments (tweets, chats, forums, blogs, wikis, social network sites, SMS, WhatsApp, Instagram, etc.). We refer to this type of written, interactional discourse as *computer-mediated communication* (*CMC*) and to the environments where CMC can be found (be it exclusively, as in the case of chatrooms, or as one among other types of discourse, as on Facebook and Wikipedia) as *social media*. The main challenge of adapting natural language processing (NLP) tools for an accurate automatic annotation of CMC data is dealing with linguistic peculiarities which result (i) from the dialogic, interactional conception of the written utterances, and (ii) from a spontaneous production strategy which is commonly adopted by CMC users, especially in informal settings. Starting from an overview of approaches that have been developed to deal with this issue (Section 2), and from an outline of the views of language technologists and linguists on PoS tagging of CMC data (Section 3), we give a summary of the results and discuss lessons learned from a community shared task on PoS tagging German CMC (*EmpiriST*) conducted in 2016 (Section 4). These results suggest that more effort should be put into the development of solutions for dealing with phenomena which, one the one hand, are too sparsely represented in data samples that could be used for training tagger models, but, on the other hand, are of special interest for the annotation of linguistic corpora. In Section 4, we present a case study in which we used a PoS tagger to find one particular phenomenon of this type, namely German verb-pronoun contractions (*haste*, *schreibste*, *gibts*, *geht's*, ...) in chats and tweets. The results open up some directions for further work and suggest close cooperation between language technologists and linguists as a promising approach for further advances in the automatic identification of rare phenomena in corpora.

2 STATE-OF-THE-ART

Robust part-of-speech (PoS) tagging of CMC still poses a challenge. Instead of tagging accuracy in the high nineties, as on edited text, which is close to the written standard (as can be found in newswire texts and similar text types), we see a big performance drop on CMC, where we only get accuracies of around 80% (Ritter et al. 2011) or even less, depending on the genre (e.g., 69% as a baseline for German chats, as reported by Horbach et al. 2014). The main reason for this performance drop, as noted in Eisenstein (2013), is the high number of out-of-vocabulary words in CMC. Authors, for instance, may neglect orthographic rules and join, add, omit, or swap letters. Bartz et al. (2013) give a typology of linguistic phenomena which affect this performance, and group them into six main types (with subtypes): speedwriting phenomena, written emulations of prosody, colloquial spellings, creative spellings, CMC-specific acronyms and CMC-specific 'interactive units,' which include emoticons, addressing terms and German inflectives. The dialogic character of written utterances in CMC, moreover, also affects syntax, as for example personal pronouns at the beginning of sentences are often omitted (*ellipsis*), as in "went to the gym," where the pronoun 'I' is implied (Ritter et al. 2011). There are two main paradigms to tackle these challenges, normalisation and domain adaptation, as discussed below.

Normalisation removes the orthographic and syntactical anomalies of a text and brings them into their correct form (Han and Baldwin 2011, Chrupala 2014). The text is fitted to the tagger, which is usually trained on edited text, prototypically newswire text, which enables the tagger to perform well. Easy as this might sound, normalisation is probably a more challenging task than domain adaptation. In order to perform normalisation, one has to know (i) that a certain word form is a non-standard form, and (ii) how to normalise it. This entails two tasks, detection and correction. For both steps, an external knowledge source is needed which, especially for the CMC domain, with its many non-standard word forms, can be expected to have a coverage problem. Since performance depends on the degree of coverage obtained, the resulting normalised sequence is not necessarily easier to tag. As such, we will use the second paradigm, domain adaptation, which is more suited to the current work, since it operates directly on the word forms as they appear in CMC data.

Domain Adaptation uses PoS annotated text from the CMC domain to retrain the tagger. The tagger thus learns the characteristics of the domain and is then able to tag CMC data with high accuracy. As existing manually annotated CMC data sets are rather small, a strategy to compensate for this data sparsity problem is to add knowledge from other discourse domains. There are two main strategies for this described in the literature. First, to add more labelled training data by adding foreign domain or machine-generated data (Daumé III 2007; Ritter et al. 2011). Machine-generated data can be created, for instance, by applying several newswire-trained PoS taggers to CMC discourse and adding the related data to the training set when the taggers agree. A second approach is to incorporate external knowledge from resources containing word distributional knowledge, and to guide the machine learning algorithm to extract more information from the existing data (Ritter et al. 2011, Owoputi et al. 2013). The first strategy is related to *which* kind of data is learned, while second to *what* is learned.

The main challenge in tagging CMC lies in dealing with the large number of unknown word forms. Van Halteren and Oostdijk (2014) estimate a range of 20%

to 36% non-word tokens and 4% to 11% out-of-vocabulary (OOV) tokens in (Dutch) tweets. The PoS annotated data sets from the CMC domain are usually too small to cover the high number of word forms which can occur in CMC data, and so cannot yield robust models. While for some languages (e.g. English and German) several data sets exist, these are not easy to combine as the annotation schemes and tagsets used differ, and cannot be easily harmonised.

In the face of these problems with regard to a lack of training data, three methods have been shown to yield considerable improvements with regard to tagging CMC data for English (Ritter et al. 2011, Owoputi et al. 2013) and German (Rehbein 2013, Neunerdt et al. 2013):

1. adding foreign domain data to add lexical and contextual knowledge,
2. adding PoS dictionaries created from other existing corpora,
3. adding word distributional knowledge obtained from unsupervised machine learning methods trained on large collections of plain text.

(1) With the use of *foreign domain data*, text from other existing corpora which have an at least partly compatible PoS tagset is added. Most of the time newswire corpora with edited text are used for this, and these are available for many languages; however, similar-domain text data – such as chat corpora, in the context of the current study – are used if available. Adding more data to the tagger and thus providing more lexical knowledge can be useful in the CMC domain, as it is very useful to know which words can occur together and which inflections are possible for a word (even if only in standard language).

(2) *PoS dictionaries* contain the most frequent PoS tags a word form can have. These dictionaries are created from various corpora, and mainly serve to provide a bias for OOV words. The usefulness of a dictionary is determined by the similarity of the source corpus to the CMC domain and its size. For instance, Neunerdt et al. (2014) created a verb lexicon from a website which also lists common contracted forms that may occur in informal written communication.

(3) *Word distributional knowledge* is provided by applying clustering methods to a large amount of unlabelled data from the CMC domain. Words are clustered according to their distributional similarity, i.e. by a similar word context in which they tend to appear. This property is particularly valuable for PoS tagging of CMC data, as many spelling variations of the same word (e.g. *tomorrow*, *tmr*, *2mr*, *tmrrow*, etc.) tend to be placed into the same cluster (Ritter et al. 2011). If at least one of the word forms in a cluster did occur in the training data, i.e. the correctly spelled form, the tagger receives a bias to assign an unknown word the same tag as that of the known word if both words appear in the same cluster.

The obtained word clusters are identified by ID numbers which can be understood as a kind of PoS tag. According to the similarity function used for clustering, all words which are placed into the same cluster occur in similar word contexts. Hence, one will find clusters with gerund verbs, happy emoticons, sad emoticons, plural nouns, and so on. The number of created clusters usually exceeds the number of tags in human-defined tagsets. Furthermore, the numbering of the clusters is arbitrary, and each time the clustering algorithm is executed the clusters will have different IDs. This arbitrary numbering limits the use of clustering methods for linguists, as cluster IDs are always changing. By using clusters in supervised machine learning scenarios, a mapping from the arbitrary numbering to the tags in a human-defined tagset can be learned, which enables the use of unsupervised methods in supervised setups.

Word clusters have been reported as highly effective if the clustering is applied over a large collection of plain text (Ritter et al. 2011, Rehbein 2013), with Brown clusters (Brown et al. 1992) being frequently used in the literature. Words in Brown clusters are identified by a binary string, and this can be used to express partial similarity between words by overlaps in the binary code. If this binary code is provided in varying length (Owoputi et al. 2013), then the tagging accuracy improves during training to a greater extent than just by providing the entire string as a cluster ID. Brown clustering is a hard-clustering algorithm, and a word will eventually be part of only one cluster. This contrasts with soft-clustering algorithms, such as *Latent Dirichlet Allocation* (*LDA*) (Blei et al. 2003; Chrupala 2011), which uses probabilistic word classes, and with which a word can belong to more than one cluster. Horsmann and Zesch (2015) show that Brown clustering is more suitable than LDA for PoS tagging of CMC data.

3 POS TAGGING CMC FROM THE PERSPECTIVES OF LANGUAGE TECHNOLOGISTS AND THE LINGUISTS

3.1 The language technologist's view

From a technical viewpoint, a PoS tagger performs well if it reaches a high accuracy and is robust against transfers to other domains of textual data. This high accuracy is a criterion readily fulfilled by many tagger implementations, while the criterion of robustness is often not. Taggers are usually evaluated by choosing one corpus and splitting it up into a training and testing set. The most prominent example of this approach for English is the same corpus evaluation of the Wall Street Journal (WSJ) (Marcus et al. 1993) based on a de-facto standard data split. Each new tagger implementation reports the tagging results on this data split as point of reference to other implementations. Such evaluations reach high accuracies, but they also evaluate under ideal conditions, since the training and testing data are very similar to each other (Giesbrecht and Evert 2009). This high similarity is unrealistic for real setups, however, and as soon foreign domain data is used for such evaluations the tagging accuracy decreases, with the severity of this decline depending on the degree of dissimilarity. The CMC domain is a such a severe case, with the Stanford tagger (Toutanova et al. 2003), for instance, achieving over 97% accuracy with the WSJ data (Manning 2011), but only 80% with the CMC data set examined by Ritter et al. (2011).

It thus seems as if there is no all-round tagger within reach, as no newswire-trained tagger has a sufficiently high robustness to work on the CMC domain with a similar high accuracy as that seen on edited standard-text. This lack of robustness has motivated considerable research into domain adaptation to re-train tagger models on a mixture of data from several domains, and provide supplementary knowledge from other resources.

3.2 The linguist's view

For qualitative and quantitative empirical analyses of authentic language data, linguists are interested in using corpora which provide highly accurate PoS annotations, and can thus be queried not only for word tokens, but also for morphosyntactic patterns. For the domain of edited text (fictional prose, scientific and newspaper text and similar genres), the reference corpora provided by the Berlin-Brandenburg Academy of Sciences (DWDS corpus, Geyken 2007) and by the Institute for the German Language (DeReKo, Kupietz et al. 2010) are examples which meet this requirement. For the domain of CMC, corpora with highly accurate linguistic annotations still need to be developed, since existing taggers still cannot sufficiently deal with the linguistic peculiarities of CMC discourse.

From a linguistic perspective, and especially for research on the commonalities and differences between the written, interactional language of CMC, the written language of edited text and the language of spoken interactions, a PoS layer in CMC corpora should, on the one hand, adequately represent units which are specific to CMC discourse – such as emoticons, hashtags, non-inflected verb stems (*grins*, *lach*, *grübel*), addressing terms, email addresses and URLs. On the other hand, taggers should also be able to deal with phenomena which are not unique to CMC data but are typical for all types of discourse in informal, interactional settings with spontaneous language production. Besides CMC genres, phenomena of that type occur in spoken language and even in certain domains of edited text (e.g. in direct speech or quotations as parts of literary prose or newspaper articles). Examples of phenomena of this type are interjections, discourse markers, modal particles and intensifiers, colloquial contractions, and onomatopoeia – phenomena which are only rudimentarily covered by PoS tagsets which have been created for processing edited or newswire texts. The Stuttgart-Tübingen Tagset (STTS, Schiller et al. 1999) for instance, which is a de-facto standard for the tagging of German text corpora, includes a tag for interjections (ITJ), whereas modal particles, downtoners, intensifiers, focus and gradation particles are not represented as unique categories (instead, they are included in the ADV category for adverbs). For contractions, the tagset only covers preposition-article contractions (APPRART) which are part of the written standard, and which are characterised by a high degree of grammaticalisation (German *im, am, zum, vom, ins*); the vast variety of contractions beyond the APPRART type which are typical of colloquial language (e.g., verb-pronoun, conjunction-pronoun, adverb-article) cannot be adequately labelled using STTS.

A precise PoS annotation which covers the aforementioned phenomena can, moreover, form the basis for the (manual or NLP-assisted) creation of more sophisticated corpus annotations, e.g. on syntactic, semantic, pragmatic or interactional patterns.

4 *EMPIRIST*: A COMMUNITY SHARED TASK FOR POS TAGGING GERMAN CMC DATA

In this section, we give a summary of the design and results of a community shared task which was organised to foster the adaptation of NLP tools for the automatic annotation of German CMC data. *EmpiriST* ("Empirikom Shared Task") resulting from an initiative of the interdisciplinary scientific network "Empirical Research on Computer-mediated Communication" (Empirikom, http://www.empirikom.net) which was funded by the DFG 2010–2014, and in which linguists, language technologists, computer scientists and psychologists worked on solutions for open issues related to the acquisition, design and analysis of CMC data sets. A detailed documentation of the task including descriptions of the participating systems is given in WAC-X/EmpiriST (2016).

4.1 Focus and layout of the task

The focus of EmpiriST was on PoS tagging of German CMC data in two types of resources: (1) as part of genuine CMC corpora, (2) as part of large corpora crawled from the web (web corpora). The task provided annotated data sets of CMC and web text to participants as training data to adapt PoS taggers to the CMC domain. EmpiriST consisted of the two subtasks, (1) tokenisation and (2) PoS tagging. These subtasks were performed on two data sets: (i) a CMC data set with samples from several CMC genres (tweets, chats, Wikipedia talk pages, WhatsApp interactions, blog comments), and (ii) a web corpora data set of CC-licensed web pages (including a small portion of CMC discourse). All in all, 23k tokens of training and testing data were annotated, each subset by at least two trained annotators.

4.2 Tagset

EmpiriST adopted the 'STTS 2.0' tagset (Beißwenger et al. 2015), which expands the canonical version of the Stuttgart-Tübingen-Tagset (Schiller et al. 1999, henceforth 'STTS 1.0') with 18 new tags that are relevant for the tagging of linguistic peculiarities in written CMC interactions that cannot be adequately handled with the STTS 1.0 categories (Table 1). According to the linguist's view described in Section 3.2, STTS 2.0 introduces two 'families' of new tags:

(i) tags for phenomena that are specific to CMC discourse: ASCII emoticons and emojis, 'interaction words' describing facial expressions, gestures, bodily actions, or virtual events (cf. Beißwenger et al. 2012: 3.5.1.3), hashtags, addressing terms, URLs and e-mail addresses.

(ii) tags for phenomena that are typical of spontaneous (spoken or 'conceptually oral') language in colloquial registers: tags for types of colloquial contractions which frequently occur in German chats, tags for discourse markers and onomatopoeia, and, finally, three tags which allow for the description of different types of particles which in STTS 1.0 are treated as adverbs without further subclassification:

- a tag for intensifiers, focus and gradation particles (which – besides units that belong to the written standard (*sehr*, *höchst*, *nur*) – also covers forms which are associated with colloquial registers (*<u>voll</u> geil*, *<u>krass</u> unterschiedlich*)),
- a tag for modal particles and downtoners (*Das ist <u>ja</u> / <u>vielleicht</u> doof*),
- a tag for particles which are part of multi-word lexemes (*keine <u>mehr</u>*, *<u>noch</u> mal*).

Table 1: Tagset extensions for CMC phenomena according to STTS 2.0.

PoS tag	Category	Examples
I.Tags for phenomena specific for CMC / social media discourse:		
EMO ASC	ASCII emoticon	:-) :-(^^ O.O
EMO IMG	Graphic emoticon (emoji)	🤗 😜 🙊
AKW	Interaction word	*lach*, freu, grübel, *lol*
HST	Hash tag	Kreta war super! #urlaub
ADR	Addressing term	@lothar: Wie isset so?
URL	Uniform resource locator	http://www.uni-due.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena typical for spontaneous (spoken or conceptually oral) language in colloquial registers:		
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART already exists in STTS 1.0)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfste, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART		son, sone
PTK IFG	Intensifier, focus and gradation particles	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal particles and downtoners	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder
DM	Discourse markers	weil, obwohl, nur, also, ... with V2 clauses
ONO	Onomatopoeia	boing, miau, zisch

STTS 2.0 is downward compatible to STTS 1.0, and therefore allows for interoperability with existing corpora and tools. In addition, the tagset extensions in STTS 2.0 are compatible with the STTS extensions defined at IDS Mannheim for the PoS annotation of FOLK, the Mannheim "Research and Teaching Corpus of Spoken German" (Westpfahl and Schmidt, 2016). Further details and examples for the tag categories introduced in STTS 2.0 are given in Beißwenger et al. (2015).

4.3 Results for the subtask of PoS tagging the CMC data set

Six teams submitted results for the PoS subtask from eight different systems. The subtask was evaluated in terms of the accuracy of the PoS tag assignments in the participants' submissions. For each system, the submitting team could submit up to three different runs, and only the best was considered in the task results. To put the performance of submissions into perspective, three widely used off-the-shelf tools were additionally evaluated as baselines: TreeTagger v3.2 (Schmid 1995), Stanford tagger v3.6.0 (Toutanova et al. 2003), and the COW pipeline (Schäfer and Bildhauer 2012, Schäfer 2015). Agreement was calculated (1) for the official gold standard on the basis of STTS 2.0, and (2) for the canonical STTS 1.0 on the basis of a coarse-grained mapping of the 18 new tags in STTS 2.0 to the most acceptable corresponding tag(s) in STTS 1.0. The latter was done to allow for a better comparison of the submitted systems with off-the-shelf taggers which are not aware of the STTS 2.0 tagset extensions. Table 2 gives a summary of the results of the submissions and of the three baseline systems for the PoS subtask on the CMC data set. A detailed description of the evaluation metrics and the results is given in Beißwenger et al. (2016).

Table 2: Summary of results of the EmpiriST subtask on PoS tagging for CMC data (Beißwenger et al. 2016).

System	acc (**STTS 2.0**)	acc (**STTS 1.0**)
UdS-distributional	87.33	90.28
UdS-retrain	86.40	89.07
UdS-surface	86.45	89.28
LTL-UDE	86.07	88.84
AIPHES	84.22	87.10
bot.zen (*non-competitive*)	85.42	87.47
$WAGMOB (*non-competitive*)	84.77	87.03
COW (*baseline*)	77.89	81.51
TreeTagger (*baseline*)	73.21	76.81
Stanford (*baseline*)	70.60	75.83

The improvements shown by the submitted systems compared to the baseline systems is striking: the best submitted tagger achieved an accuracy of 87.33% evaluated against STTS 2.0 (vs. 77.89% baseline), and an accuracy of 90.28% against STTS 1.0 (vs. 81.51% baseline). Nevertheless, since the EmpiriST training and testing data sets were compiled of snippets of authentic CMC interactions, the number of occurrences of the 18 newly introduced PoS tags in STTS 2.0 was extremely varied, as shown in Table 3.

Table 3: All 18 newly introduced PoS tags from STTS 2.0 with their frequency of occurrence in the training data compared to the frequency of the 18 least frequent STTS 1.0 tags (Horsmann and Zesch 2016).

Tags specific of STTS 2.0	Freq	Least frequent tags in STTS 1.0	Freq
EMOASC	115	PTKANT	42
PTKMA	103	PWAV	39
PTKIFG	99	KOKOM	28
AKW	49	XY	28
HST	46	PDAT	28
ADR	35	VAINF	26
PTKMWL	28	PWS	23
EMOIMG	22	VVIMP	18
URL	18	TRUNC	12
VVPPER	7	KOUI	10
VAPPER	4	PWAT	8
DM	3	VVIZU	7
VMPPER	1	PIDAT	7
ADVART	1	PTKA	5
KOUSPPER	1	APZR	5
ONO	1	VMINF	3
PPERPPER	1	VAPP	3
EML	0	VMPP	1

From the view of corpora representing natural language, the uneven distribution of occurrences with regard to the PoS categories is a notable feature. From the view of language technology, it is an issue that has to be addressed.

4.4 Discussion of the results from the language technologist's perspective: The challenge of rare phenomena

Evaluations of PoS taggers usually focus on the accuracy computed over all PoS tag classes as the main metric of assessment. The frequency of the individual PoS tags varies greatly, which is why a high level of correctness with regard to frequent tags will automatically lead to a high accuracy. At least for English and German, those classes are typically nouns, verbs, adjective and adverbs. Conversely, errors in tagging infrequent tag classes barely have an influence on the accuracy, and thus an accuracy in the mid-nineties tells us little about the system's performance on infrequent tags. More suitable measures do exist, computed for each individual tag, such as the F-score. However, the convenience of having a single value which expresses the overall performance makes accuracy the preferred metric of evaluation.

PoS tagsets for the CMC domain tend to add additional PoS tag classes (Rehbein, 2013, Beißwenger et al. 2015) to address the phenomena of informal language use. Some of these additional tag classes are extremely infrequent, which makes it difficult for the tagger to learn to recognise them during model training. In particular when CMC corpora which ought to represent a certain (sub-)domain are compiled, the problem of infrequency becomes more extreme when tags occur only once or twice. Horsmann and Zesch (2016b) show that such ultra-rare phenomena are not learned by a tagger, even it is able to reach an accuracy of around 90%.

The lesson learned from the EmpiriST shared task is that annotation of rare phenomena is only reasonable when a sufficient number of samples can be provided for each tag. This certainly conflicts with the goal of having a corpus that represents the natural distribution in a domain. Under practical considerations, when rare phenomena need to be studied, it is more reasonable to give up on the natural distribution and provide additional annotated sequences with the phenomena of interest in order to provide enough training instances to be learned by the tagger.

5 EXPERIMENTS IN POS TAGGING LOW-FREQUENT LINGUISTIC PHENOMENA: THE CASE OF GERMAN VERB-PRONOUN CONTRACTIONS

In this section, we present an experiment in which we investigate how to improve the tagging accuracy on German **verb-pronoun contractions**. Verb-pronoun contractions belong to the class of phenomena which are not unique to CMC discourse, but typical for spontaneous – spoken or ’conceptually oral’ – language in colloquial registers. Phenomena of this type are of special interest to linguists who want to use corpora to compare written discourse from the CMC domain with the language of edited text and that found in informal, spoken interactions. Table 4 shows examples of such contractions taken from the Dortmund Chat Corpus (Beißwenger 2013, Lüngen et al. 2016). Compared to other PoS classes, verb-pronoun contractions must be considered a rarely occurring phenomenon; at the same time, the number of possible forms for this pattern that may occur in a corpus cannot be predicted. In the EmpiriST training data, we found 12 occurrences (seven of the type full verb + pronoun, four of the type auxiliary + pronoun, one of the type modal verb + pronoun, cf. Table 3). Since the use of verb-pronoun contractions is considered typical for informal settings, the frequency of its occurrence may vary in different CMC genres and contexts (e.g., social chats vs. chats in the context of learning and teaching). Verb-pronoun contractions are therefore an excellent case to explore how a tagger can be adapted to the identification of phenomena which typically (1) occur rarely, (2) in a big variety of possible forms, and without (3) the number of occurrences and the variety of forms being able to be anticipated.

Table 4: Examples of contractions of a full verb with a personal pronoun.

wiederholen (to repeat) + es (it)	1st person
	ich **wiederhols** nochmal, ihr redet hier öffentlich! *I repeat it [repeat-it] again, you're talking in public!*
kommen (to come) + du (you)	2nd person
	wieso? wo **kommste** denn her? ich besuch dich auch! *why? where do you come [come-you] from? i will visit you too!*
finden (to find) + du (you)	2nd person
	nö,dat ebste **findeste** eigentlich wenn du gar nich suchst sondern einfach guckst was da ist *nope, you find [find-you] the best when you're not searching for it but just look what's there*
machen (to make) + es (it)	3rd person
	shortnews.de **machts** möglich wenn die supermarktwebcams reinverlinkt werden:-) *shortnews.de makes it [makes-it] possible when they link to the super market webcams:-)*

As a prerequisite for studying the use of this phenomenon in the CMC domain, we are adapting a tagger for dealing with VVPPER contractions so that it may be used as a tool for retrieving new instances of VVPPER in raw data. This tagger needs high precision to avoid screening through countless false positive instances, and at the same time we need to be able to find new lexical instances for our studies, which requires a high level of recall. Building such a tagger needs a sufficiently large number of training instances, which poses the biggest challenge to this project, as such data is not readily available. We will thus address two sub-problems: first, how to deal with the lack of training data, and second, how to reach a reasonable trade-off between precision and recall. The focus of our experiments will lie on verb-pronoun contractions of the type *full verb + personal pronoun,* for which STTS 2.0 introduces the tag **VVPPER** with 'VV' representing the full verb (German *Vollverb*) and 'PPER' the personal pronoun (German *Personalpronomen*) component.

5.1 Data set

For building our training data set, we build on the (small) set of 23k manually PoS annotated tokens provided in the context of the EmpiriST project (cf. Section 4) which was annotated using STTS 2.0 (Beißwenger et al. 2015). There are 13 VVPPER instances in the EmpiriST data set, which we split into the training set (seven occurrences, cf. Table 3) and testing set (six occurrences).

Since the VVPPER tag is not included in the canonical STTS, the low representation of the phenomenon in the data cannot be increased using existing corpora which are tagged with STTS 1.0. Therefore, to arrive at meaningful results, we have to increase the number of verb contractions artificially. To do so, we manually select 230 user posts containing this phenomenon from the Dortmund Chat Corpus and machine-tagged these using the Stanford tagger. We manually assign the correct PoS tag from the STTS 2.0 to all VVPPER occurrences, but leave the remaining tags untouched. We have no interest in reaching a new *best-accuracy result*, and thus the performance on other tags is not of primary importance. Of course, ensuring the correctness of the surrounding tags is desirable, but we want to avoid labour intensive, manual annotation as much as possible. We therefore focus on providing verified lexical (context) knowledge of VVPPER and risk wrong surrounding tags as a result of the machine tagging. This enables us to add many additional sequences and inform the tagger more extensively about the phenomenon of interest. Of the 230 instances, we add one half (115) to the test set and one sixth (38) to the training set. The remaining two sixths (77) are the (held back) development set, and will be used in the experiment to increase the number of instances. Hence, our enhanced data set now contains 45 (7+38) VVPER instances in the training set (seven from the EmpiriST data set and 38 from the additional chat data set) and 121 VVPPER instances in the test set (six EmpiriST, 115 chat). These should be enough training instances for learning the phenomenon, and enough instances for evaluating the tagger, especially with respect to generalisation.

The set of 230 chat posts with PoS annotations can be retrieved from the CLARIN repository at IDS Mannheim via http://hdl.handle.net/10932/00-0374-4A34-CED0-0801-B and may be re-used by developers under a CC-BY-SA license.

5.2 PoS Taggers

To find the system which is best suited to the task, we experiment with various PoS taggers and compare different tagger implementations to each other:

Stanford: We include the Stanford (Toutanova et al. 2003) tagger as a widely-used system and train maximum entropy models. We use the default configuration provided for training the German STTS (1.0) model.

HunPos: A Hidden-Markov model based tagger by Halácsy et al. (2007) which is a freely available re-implementation of the TnT tagger by Brants (2000). We choose this tagger to have a further well-known tagger in our setup which is frequently used in the literature, and thus to provide a comparison with the results achieved with the Stanford tagger.

LSTM: A deep-learning PoS tagger by Plank et al. (2016) which is based on Long-Short-Term-Memory (Hochreiter and Schmidhuber 1997) neural networks. This tagger has an interesting property, as it considers the word frequency during model training, which leads to an improved performance on rare words. For our purposes, we argue that rare words and the tagging of rare tags are highly related, as rare tags often also have only rarely occurring word forms. This particular implementation might thus offer some advantages for our use case. We run the tagger with the same parametrisation as Plank et al. (2016), and use a German word embedding which we create from 195 million tokens of German Twitter messages we crawled between 2011 and 2017.

Two-Step: Horsmann and Zesch (2016a) proposed a tagger architecture for CMC data that first uses a highly generalised *coarse-grained* tagger, and as a second step applies a specialised non-sequential tagger for *fine-grained* tagging. The second tagger is tailored towards recognising the tag of interest, while the first tagging step constrains the second tagger, e.g. the non-sequential tagger fitted to verbs contractions would be only applied if the sequence model has tagged a word as a verb. We train the coarse-grained sequence tagging model by using Conditional Random Fields (Lafferty et al. 2001) on the abovementioned training set of EmpiriST data and additionally annotated VVPPER instances. The STTS 2.0 tags are mapped to the coarse-grained tagset by the Universal Dependencies project. We add mappings for the contraction phenomena which are not part of the canonical STTS, and treat the VVPPER instances as a verb form. We include a PoS dictionary and Brown clusters (Brown et al. 1992) created from German Twitter messages to compensate for the lack of CMC training data. This coarse tagger reaches an F1 of 0.93 on the coarse-tag *Verb* in the test data set, which is essential for tagging VVPPER.[1]

We train a Support Vector Machine (SVM) for the second step using Weka (Hall et al. 2009), a machine learning toolkit. The SVM is trained on the same data as the sequence model, and is fitted to the local word context in which the VVPPER instances occur. As context features, we use the current word and the first and second words to the right and left. We also use character bigrams over all verbs.

[1] As such, some VVPPER instances might be missed if the coarse-model does not predict 'verb'.

5.3 Experiment: Frequency weight vs. lexical knowledge

In this experiment, we want to learn which information is more relevant for tagging VVPPER instances. We experiment with altering the frequency in the training data by over- and undersampling, and compare the performance to when adding newly annotated instances.

Setup: While annotation of more data will certainly improve the performance, we also want to investigate if we can improve tagging of this particular PoS tag by altering the overall tag distribution. This can either be done by **oversampling** the few instances in the data set (cf. weighting of data, Daumé III, 2007) or by **undersampling**, i.e. removing data from the large other PoS tag classes. Both approaches lead to an increased frequency weight of the focal phenomenon by increasing its frequency relative to the rest of the corpus. If undersampling is applied, sentences which *do not* contain the tag of interest are removed. This shrinks the overall corpus size, so that the tag becomes more frequent than in the original distribution. If oversampling is applied, the sentences with the phenomenon are added several times to increase its frequency weight, but leaving the rest of the corpus untouched. We use the following sampling levels:

- *Downsampling:* We remove 25, 50 and 75 percent of the training data instances which do not contain any VVPPER instances.
- *Oversampling/new instances:* To reach comparable results between oversampling and adding new training instances, we constrain the oversampling to fit the number of held back hand annotated sequences. We thus oversample the additionally added training data two and three times and compare this to adding the same amount of newly annotated data from the held back data.

Results: In Figure 1, we show the results on *out-of-vocabulary* (*OOV*) instances which did not occur in the training set and, hence, show the performance of the taggers to find new lexical forms. We focus on OOV instances because all taggers perform well in recognising *in-vocabulary* words, with an F1 between 0.96 to 0.99. Neither downsampling nor oversampling helps to achieve a substantial improvement on the tag. Furthermore, downsampling shows that the already small amount of training data becomes a large problem for the LSTM if this is further reduced. The Stanford tagger lags behind the other taggers with both sampling methods. Unsurprisingly, the only effective method is providing new data. With this approach, the LSTM needs considerably more data to improve, while the other taggers improve linearly with each new data set.

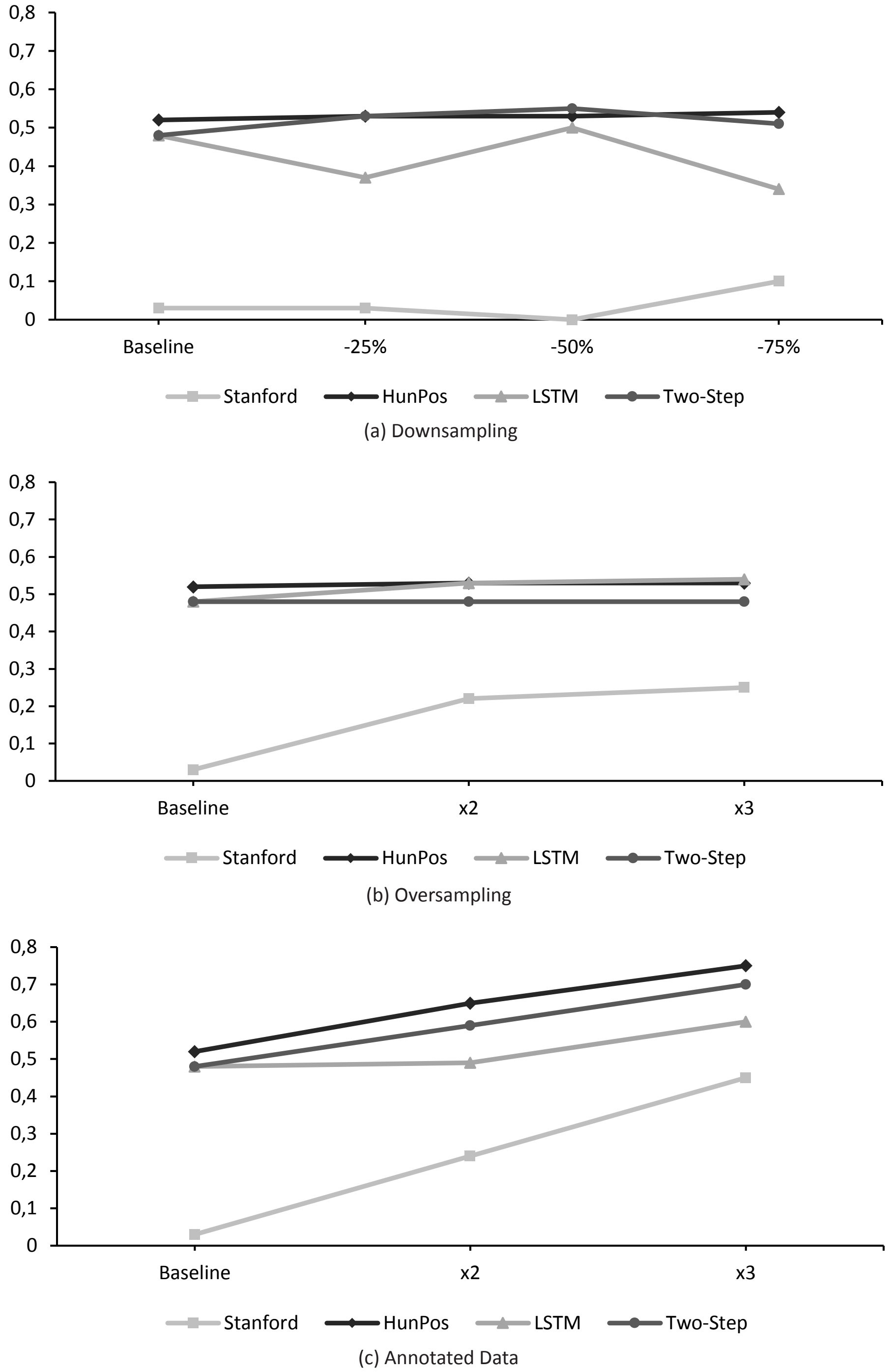


Figure 1: Results on *unknown* VVPPER word forms with various methods.

Discussion: Table 5 shows details of the two best taggers, HunPoS and Two-Step. We focus again on the out-of-vocabulary instances, this time presenting also precision (P) and recall (R). The overall F1 score shows that the overall performance of both taggers is rather similar. When looking at precision and recall, highlighted in grey, we see that Two-Step is considerably more precise than HunPos, which has a better recall.

Since oversampling showed barely any effect, we suspect that the added lexical knowledge can account for the improvements we see when adding more data. This would mean that the tagger focuses too much on the lexical forms and does not weight the word context sufficiently.

Table 5: F_1 on all and on out-of-vocabulary instances.

		All	Out-Of-Vocabulary		
	Setup	**F1**	**P**	**R**	**F1**
HunPos	Baseline	.78	.80	.38	.52
	Downs. 75%	.78	.63	.48	.54
	Downs. 50%	.79	.74	.41	.53
	Downs. 25%	.79	.81	.40	.53
	Overs. x2	.79	.78	.40	.53
	Overs. x3	.79	.74	.41	.53
	Annotated x2	.83	.80	.56	.65
	Annotated x3	**.88**	.81	.70	.75
Two-Step	Baseline	.77	.95	.32	.51
	Downs. 75%	.78	.96	.38	.55
	Downs. 50%	.80	.96	.38	.53
	Downs. 25%	.79	.92	.32	.48
	Overs. x2	.77	.95	.32	.48
	Overs. x3	.77	.95	.32	.48
	Annotated x2	.81	.93	.43	.59
	Annotated x3	**.85**	.92	.56	.69

5.4 Experiment: Forced generalisation

In this experiment, we examine if we can improve the performance of the Two-Step tagger by forcing it to rely more on the local word context, and thus improve the recall. Since this tagger is self-implemented, we can easily adjust the implementation. We alter the feature space of the SVM and exclude all features which contain the lexical form of the positive instances. The SVM is thus not aware of any lexical forms that can occur with the PoS of interest, and must now rely more strongly on the word context.

Results: In Table 6, we show the changes in performance of the contextualised Two-Step tagger. In parentheses, we show the differences compared to the non-contextualised tagger in Table 5. For both setups, we see an improvement on the overall F1, but the recall especially increases for out-of-vocabulary instances. The overall F1 reached by HunPos (.88) is still better, but the trade-off between precision and recall of Two-Step more efficiently supports the use case of using the tagger as a filtering tool.

Table 6: Results of the contextualised Two-Step.

	All	Out-Of-Vocabulary		
Configuration	F1	P	R	F1
Baseline	.81 (+.04)	.93 (+.02)	.41 (+.09)	.57 (+.09)
Annotated x3	.86 (+.01)	.89 (-.03)	.62 (+.06)	.73 (+.04)

5.5 Experiment: Field trial in CMC

So far, we have only simulated our use case of using a tagger as a filtering tool. Now we turn to a real setting: we tag plain CMC data to find VVPPER instances. Working on unlabelled text means that the ground truth for computing the recall is unknown. We will thus focus on evaluating the precision of the tagging and evaluate how many new instances are found. We choose the Twitter domain for its ease of obtaining data, but also for its linguistic diversity. Some tweets may grammatically and orthographically conform to the written standard while others – more similar to social chat than to edited standard-text – may be noisy and deviant from the orthographic standard, and contain conceptually oral and colloquial language. Tweets of the latter type are the kind of data in which we expect occurrences of VVPPER and other types of colloquial contractions. Twitter thus provides us with a text domain which contains a large amount of naturally occurring noise (which, of course, from the linguist's view, may be the data which is most interesting for analysing the peculiarities of CMC). Evaluating this domain will provide us with a conservative, lower-bound performance for finding this phenomenon. We use the contextualised Two-Step tagger for its higher precision while still providing reasonably high recall.

Twitter Data: We use tweets that we crawled between 2011 and 2017 from the public Twitter API[2] endpoint, which allows retrieval of a random subsample of all world-wide posted Twitter messages when this endpoint is accessed. We language-filter those tweets and extract a random sample of 50k German tweets

[2] https://dev.twitter.com/streaming/public?lang=en, last accessed 6th of June, 2017.

(about 1.7 million tokens) between the years 2011 to 2017. All occurrences of addressing terms, hashtags and URLs are replaced by a text constant. The tweets are tokenised by Gimpel et al.'s (2011) ArkTools tokeniser.

Tagger setup: We train the coarse model and the SVM on the full EmpiriST data set including the additionally annotated data. To provide more lexical knowledge and increase the robustness when facing standard language text, we also add 100k tokens of the German newswire Tiger (Brants et al. 2004) corpus to both tagging steps.

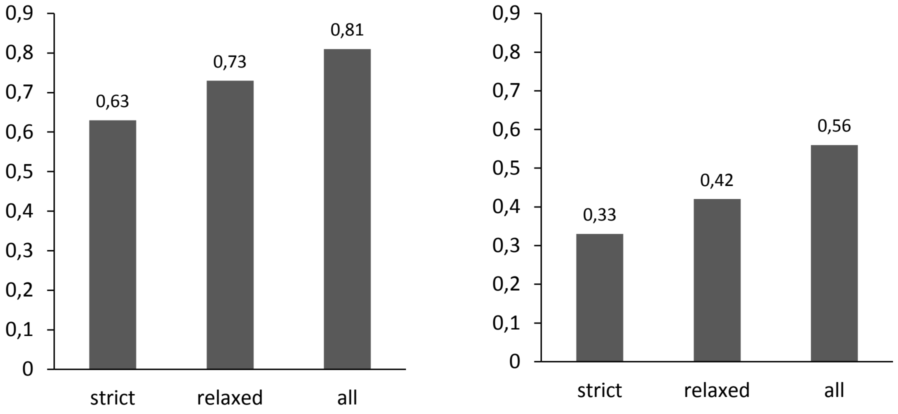
Evaluation setup: We evaluate the tagged instances with two annotators. The annotators make four distinctions: *strict*, *relaxed*, *all* and *none*. *Strict* are full verb contractions with personal pronoun (VVPPER), the exact phenomenon we intended to tag. *Relaxed* counts all verb contractions with a personal pronoun as correct, which also includes contractions with modal and auxiliary verbs as the first component (VMPPER and VAPPER according to STTS 2.0). *All* counts all phenomena as correct which, from a linguistic perspective, can be considered contractions. This additionally includes, for instance, contractions of conjunctions with personal pronouns, of adverbs with articles, or of two personal pronouns. The remaining cases are not contractions, and thus treated as false positives (= *none*).

We evaluate two setups. The first selects the first 250 of all found instances, which is the basis for the overall evaluation. The second evaluation focuses on out-of-vocabulary instances in which we remove all tagged instances that are known from the training set until we have gathered 250 instances. This set of instances is used to evaluate how frequently new instances are found.

Results: On 50k tweets we find 1,091 instances in total in which one word was tagged as VVPPER. The two annotators reach perfect agreement on the subset of the first 250 instances that are evaluated manually. Figure 2a shows the precision of the overall evaluation. The *strict* result shows that the majority of found instances are the targeted VVPPER contractions. Including modal and auxiliary verbs in the *relaxed* mode, three quarters of all matches are true positives. When considering *any* type of contractions true positives (in *all*), almost all instances are true positives. We also analysed the type[3]/token ratio, which is 0.33 for the *strict* evaluation, showing that few instances re-occur with high frequency.

In Figure 2b, we take a closer look at the performance of detecting new contractions, e.g. out-of-vocabulary instances. We focus our discussion on the *strict* results where only VVPPER instances count as true positives. The precision is

[3] Many word-forms differ by an apostrophe and are, thus, distinct types, e.g. *geht's* vs. *gehts* vs. *geht's* which are counted as three types.



(a) In- and out-of-vocabulary contractions

(b) Out-of-vocabulary contractions

Figure 2: Results of manual evaluation.

drastically reduced to almost half the value when including all instances. The type/token ratio of 0.69 is almost twice as high as the overall evaluation. This confirms that the tagger is able to recognise many new instances of the phenomenon. Furthermore, when ignoring the known instances, almost every correct instance is a new lexical form.

Table 7: Examples of tagged instances (bold) in context and PoS category according to STTS 2.0.

Strict		
	Savegames - jetzt **langts** aber !	VVPPER
	Da **lernste** pragmatisch zu sein .	VVPPER
	Ich **sachs** dir noch .	VVPPER
Relaxed		
	Ich **bins** auf jeden Fall nicht .	VAPPER
	Wer **hats** gedacht .	VAPPER
	Ich **wills** nicht ich will aber auch nicht [...]	VMPPER
All		
	So schlimm hab **ich's** mir mit noch keiner Ex verscherzt .	PPERPPER
	Warum einfach , **wenn's** auch kompliziert geht ? URL	KOUSPPER
	Ich beschränke mich **auf's** nicht im Weg stehen .	APPRART
Frequent Confusion Cases		
	Und keiner **weiss** warum .	VV
	Ich **weiss** gar nicht , was du beruflich machst .	VV
	Ich **weis** wie immer nicht ... URL	VV

Discussion: Table 7 shows examples of each of the three evaluation modes (*strict*, *relaxed*, *all*) and additionally presents three instances of a frequent confusion case which is erroneously tagged as contraction. In the *strict* case there are instances in quite different local word contexts, which supports our motivation for studying this phenomenon. A general observation about the SVM is that it seems to be biased on word endings on <*s*> or <*'s*>. Such words have a high chance of being tagged as contractions. This bias also seems to account for a rather common confusion case with the verb *weiß* (*to know*), where the German <*ß*> is erroneously replaced by <*ss*> but at the same time accounts for the related phenomena in the *relaxed* and *all* evaluation. We are planning to address the further reduction of false positives in future work.

6 CONCLUSION

In view of the heterogeneous frequency of CMC phenomena in CMC data, the results and lessons learned from the EmpiriST shared task suggest that it is not realistic to train a tagger which performs well on any phenomena on the token/PoS level.

In particular, finding rare or ultra-rare phenomena poses serious challenges, and the small size of hand-annotated CMC training data sets causes the under-representation of such phenomena. The EmpiriST project conducted by Beißwenger et al. (2016) showed that the degree of under-representation can be so severe that machine learning methods fail almost entirely to learn how to recognise these phenomena. Increasing the frequency of rare phenomena artificially by over- and undersampling has no impact on this, as the phenomena occur just too infrequently. We thus presented a case study in which we used a PoS tagger as a filtering tool to find instances of German verb-pronoun contractions. We started from the EmpiriST training data and added an additional set of 230 hand-annotated user posts which had been selected manually from the Dortmund Chat Corpus as further instances of the phenomenon of interest. The results shows that the choice of the tagger together with the expansion of the training data with relatively small amounts of additional instances turns out to be a promising way to let the tagger learn the local word context, and thus enables tagging such phenomena with a sufficiently high recall and precision. To reduce the number of false positives, we are planning to add the results of the manual evaluation of the first 250 positives found in tweets to our training data set, and then retrain the SVM on the expanded data in a bootstrapping approach. In future work we will also investigate how tagging improves if not just the instances of interest are hand-annotated, but also their local word context, in order to find the ideal trade-off between avoiding annotation of full sentences and yet achieving improved results for a certain phenomenon.

To be able to estimate if the results of our case study may provide a general and more efficient approach to "nasty" phenomena in CMC corpora, the study should be repeated for other CMC phenomena which are either rare and/or difficult to handle with approaches from the literature. More close cooperation between language technologists and linguists is thus recommended, as this would enable the creation and annotation of the high-quality samples from CMC corpora which are needed for training.

References

Bartz, Thomas, Michael Beißwenger and Angelika Storrer, 2013: Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics* (JLCL) 28/1. 157–198.

Beißwenger, Michael, 2013: Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik* 41/1. 161–164.

Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer and Angelika Storrer, 2012: A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative* (jTEI) 3. http://jtei.revues.org/476. (Last accessed 5 May 2017.)

Beißwenger, Michael, Thomas Bartz, Angelika Storrer and Swantje Westpfahl, 2015: *Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation / Tagset and guidelines for the PoS tagging of langauge data from genres of computer-mediated communication / social media. EmpiriST guideline document (German and English version).* https://sites.google.com/site/empirist2015/home/annotation-guidelines. (Last accessed 5 May 2017.)

Beißwenger, Michael, Sabine Bartsch, Stefan Evert and Kay-Michael Würzner, 2016: EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics. 44–56. http://aclweb.org/anthology/W/W16/W16-2606.pdf. (Last accessed 5 May 2017.)

Blei, David M., Andrew Y. Ng and Michael I. Jordan, 2003: Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3. 993–1022.

Brants, Thorsten, 2000: TnT: A Statistical Part-of-speech Tagger. *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Seattle: Association for Computational Linguistics. 224–231.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith and Hans Uszkoreit, 2004: TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation* 2/4. 597–620.

Brown, Peter F., DeSouza, Peter V., Mercer, Robert L., Pietra, Vincent J. Della, Lai, Jenifer C., 1992: Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18. 467-479

Chrupala, Gzegorz, 2011: Efficient induction of probabilistic word classes with LDA. *Proceedings of the Fifth International Joint Conference on Natural Language Processing.* Chiang Mai: Asian Federation of Natural Language Processing. 363–372.

Chrupala, Gzegorz, 2014: Normalizing tweets with edit scripts and recurrent neural embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore: Association for Computational Linguistics. 680–686.

Cook, Paul, Stefan Evert, Roland Schäfer and Egon Stemle (eds.) *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26). http://aclweb.org/anthology/W/W16/W16-26.pdf. (Last accessed 5 May 2017.)

Daumé III, Hal, 2007: Frustratingly Easy Domain Adaptation. *Conference of the Association for Computational Linguistics* (ACL). Czech Republic: Association for Computational Linguistics. 256–263.

Eisenstein, Jacob, 2013: What to do about bad language on the internet. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics. 359–369.

Geyken, Alexander, 2007: The DWDS corpus: A reference corpus for the German language of the 20th century. Fellbaum, Christiane (ed.): *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. London: Bloomsbury Publishing. 23–41.

Giesbrecht, Eugenie and Stefan Evert, 2009: Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. *Proceedings of the Web as Corpus Workshop* (WAC). San Sebastian.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan and Noah A. Smith, 2011: Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers 2*. Stroudsburg: Association for Computational Linguistics. 42–47.

Halácsy, Péter, András Kornai and Csaba Oravecz, 2007: HunPos: An open source trigram tagger. *Proceedings of the 45th Annual Meeting of the ACL*. Association for Computational Linguistics. 209–212.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten, 2009: The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter* 11/1, http://dl.acm.org/citation.cfm?id=1656278. (Last accessed 5 May 2017.)

Han, Bo and Timothy Baldwin, 2011: Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 1. Stroudsburg: Association for Computational Linguistics. 368–378.

Hochreiter, Sepp and Jürgen Schmidhuber, 1997: *Long Short-Term Memory. Neural Computation.* MIT Press. 1735–1780.

Horbach, Andrea, Diana Steffen, Stefan Thater and Manfred Pinkal, 2014: Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. *Proceedings of KONVENS 2014*. Hildesheim.171–177.

Horsmann, Tobias and Torsten Zesch, 2015: Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging. *Proceeding of the Second Italian Conference on Computational Linguistics.* Trento: Accademia University Press. 166–170.

Horsmann, Tobias and Torsten Zesch, 2016a: Assigning Fine-grained PoS Tags based on High-precision Coarse-grained Tagging. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka: Dublin City University and Association for Computational Linguistics.

Horsmann, Tobias and Torsten Zesch, 2016b: LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text. *Proceedings of the 10th Web as Corpus Workshop* (WAC-X). Berlin: Association for Computational Linguistics. 120–126.

Kupietz, Marc, Cyril Belica, Holger Keibel and Andreas Witt, 2010: The German Reference Corpus DeReKo: A primordial sample for linguistic research. Calzolari, Nicoletta et al. (eds.): *Proceedings of the 7th conference on International Language Resources and Evaluation* (LREC 2010). Valletta: European Language Resources Association (ELRA). 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf. (Last accessed 5 May 2017.)

Lafferty, John D., Andrew McCallum and Fernando C. N. Pereira, 2001: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc. 282–289.

Lüngen, Harald, Michael Beißwenger, Axel Herold and Angelika Storrer, 2016: Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. Dipper, Stefanie, Friedrich Neubarth and Heike Zinsmeister (Eds.): *Proceedings of the 13th Con-*

ference on Natural Language Processing (KONVENS 2016). 156–164. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf. (Last accessed 5 May 2017.)

Manning, Christopher D., 2011: Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing* 1. Tokyo: Springer-Verlag Berlin, Heidelberg. 171–189.

Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini, 1993: Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19/2. Cambridge: MIT Press. 313–330.

Neunerdt, Melanie, Michael Reyer and Rudolf Mathar, 2013: A POS Tagger for Social Media Texts trained on Web Comments. *Polibits* 48. 61–68.

Neunerdt, Melanie, Michael Reyer and Rudolf Mathar, 2014: Efficient Training Data Enrichment and Unknown Token Handling for POS Tagging of Nonstandardized Texts. *12th Conference on Natural Language Processing* (KONVENS). Hildesheim. 186–192.

Owoputi, Olutobi, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith, 2013: Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Plank, Barbara, Anders Søgaard and Yoav Goldberg, 2016: Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL-16). Berlin: Association for Computational Linguistics. 412–418.

Rehbein, Ines, 2013: Fine-Grained POS Tagging of German Tweets. *Language Processing and Knowledge in the Web*. Springer-Verlag Berlin, Heidelberg. 162–175.

Ritter, Alan, Sam Clark, Mausam Etzioni and Oren Etzioni, 2011: Named Entity Recognition in Tweets: An Experimental Study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Edinburgh: Association for Computational Linguistics. 1524–1534.

Schäfer, Roland, 2015: Processing and querying large web corpora with the COW14 architecture. Bánski, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen and Andreas Witt (eds.): *Proceedings of Challenges in the Management of Large Corpora* 3. Lancaster: UCREL.

Schäfer, Roland and Felix Bildhauer, 2012: Building large corpora from the web using a new efficient tool chain. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC '12). Istanbul: ELRA. 486–493.

Schiller, Anne, Simone Teufel, Christine Stöckert and Christine Thielen, 1999: *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* Stuttgart: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. http://www.sfs.unituebingen.de/resources/stts-1999.pdf. (Last accessed 5 May 2017.)

Schmid, Helmut, 1995. Improvements in part-of speech tagging with an application to German. *Proceedings of the ACL SIGDAT Workshop.*

Toutanova, Kristina, Dan Klein, Christopher D. Manning and Yoram Singer, 2003: Feature rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.*

Universal Dependencies 1.2. Universal Dependencies Consortium. http://universaldependencies.github.io/docs/. (Last accessed 5 May 2017.)

van Halteren, Hans and Nelleke Oostdijk, 2014: Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens text. *Journal of Langauge Technology and Computational Linguistics* (JLCL) 29/2. 97–123.

Westpfahl, Swantje and Thomas Schmidt, 2016: FOLK-Gold – A GOLD standard for Part-of-Speech- Tagging of Spoken German. *Proceedings of the Tenth conference on International Language Resources and Evaluation* (LREC16). Paris. 1493–1499.

About the authors

Solange Aranha is Assistant Professor at the Modern Languages Department at São Paolo State University at São José do Rio Preto. She teaches English and academic writing for undergraduate students and methodology, genres, EAP and telecollaboration at the graduate level. She advises graduate students on telecollaboration studies, genre analysis and teaching and learning technologies. As a researcher, she investigates data on teletandems and is responsible for developing two multimodal corpora: DOTI (Data of Oral Teletandem Interactions) and MulTeC (Multimodal Teletandem Corpus). Her research is sponsored by FAPESP (Fundação de Amparo a pesquisa do Estado de São Paulo).

Michael Beißwenger is Professor for German Linguistics and Language Teaching at the Department of German Studies of the Faculty of Humanities, University of Duisburg-Essen. Besides the field of computer-mediated communication, which he has been researching since 1999, his research interests include corpus linguistics, digital humanities, text technology, collaborative writing, and the development of e-learning scenarios for language teaching and higher education. He is one of the initiators and members of the steering committee of the annual Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora.org), convener of the TEI special interest group on computer-mediated communication and membe of the CLARIN-D working groups German Philology and Applied and Computational Linguistics.

Steven Coats is a lecturer in English Philology in the Faculty of Humanities at the University of Oulu, Finland. He teaches courses on digital humanities, sociolinguistics, and corpus linguistics. His research interests include the discourse of computer-mediated communication, the linguistics of English varieties, bi- and multilingualism, and scripting in Python and R. He is a member of the Finnish Society for the Study of English and the European Association of Digital Humanities.

Darja Fišer is Assistant Professor and Chair of the unit for lexicology, terminology and language technologies at the Department of Translation Studies of the Faculty of Arts, University of Ljubljana and Research Associate at the Department of Knowledge Technologies at the Jožef Stefan Institute. She teaches courses on corpus linguistics and translation technologies. As a researcher, she is currently active in the fields of computer-mediated communication and lexical semantics using corpus-linguistics methods and natural language processing. She is President of the Slovenian Language Technologies Society, Chair of the FoLLI Steering Committee of the biggest European summer school on language, logic and computation ESSLLI and Director of User Involvement of the European research infrastructure for language resources and technology CLARIN.

Lydia-Mai Ho-Dac is Assistant Professor at the Department of Linguistics, University of Toulouse - Jean Jaurès. She teaches courses on corpus linguistics and natural language processing (NLP). As a researcher, her main interests are the study of genres and discourse organisation in a corpus-linguistics approach using data-driven analysis, NLP techniques and quantitative analysis. She is currently involved in projects concerned with computer-mediated communication as an active member of the French CORLI consortium and the discussion group coordinated by Michael Beißwenger and Ciara Wigham about standards for CMC corpora and for the creation of a CMC corpus infrastructure across languages and genres.

Tobias Horsmann is a doctoral researcher at the Language Technology Lab at the University of Duisburg-Essen in Germany. He holds a master's degree in Computer Science from the Technische Universität Darmstadt with a minor in English studies. His research focuses on robust part-of-speech tagging of both standard and non-standard text. He is particularly interested in building taggers suited for cross-domain tagging. In the non-standard text domain, his main interest is on social media text where he tries to find new methods to deal with the many challenging phenomena unknown from standard text.

Veronika Laippala is a Postdoctoral Researcher in the School of Languages and Translation Studies at the University of Turku, Finland. Her research focuses on corpus linguistics and computational linguistics. In particular, she has worked on the development of web-crawled corpora and corpora of computer-mediated communication in various languages and on enhancing computational methods for text linguistics and discourse analysis. She has also studied the variation of language use across different digital genres by applying methods from corpus linguistics and natural language processing.

Paola Leone is Professor at the University of Salento, Italy. She teaches Teaching Italian as L2 (graduate program) and Methodologies in Foreign Language Teaching (graduate program). Her main research interest is computer-mediated communication and language learning. She particularly focuses on the structure and conversational management of teletandem interaction and the use of discourse markers in Italian as L2. She has participated in a number of EU-funded projects. Currently, she is involved in the project Lecturio+, which focuses on the implementation of learning scenarios for developing students' intercomprehension ability.

Nikola Ljubešić is Assistant Professor at the Department of Information and Communication Sciences, University of Zagreb, and Postdoctoral Researcher at the Department of Knowledge Technologies at the Jožef Stefan Institute in Ljubljana. His main research interests are representation learning for lexical semantics and social media analytics, semantic shift detection, cross-lingual lexical feature prediction, linguistic processing of non-canonical texts, non-canonical text normalisation, user profiling and detection of inappropriate content on social media. He teaches introductory courses on natural language processing and machine learning. He is a member of the Association for Computational Linguistics, the Slovene Society for Language Technologies and the Croatian Society for Language Technologies.

Maja Miličević is Associate Professor at the Department of General Linguistics in the Faculty of Philology, University of Belgrade. She teaches courses on second language acquisition, psycholinguistics, corpus linguistics and quantitative methods in language studies. Her research interests include the role of transfer in second language acquisition, linguistic properties of translations, and computer-mediated communication. The languages she works on most are Serbian, Italian and English. She is particularly interested in research methodology and has held or co-held a number of seminars and online courses dedicated to statistical analysis and general methodological issues in the study of language.

Céline Poudat is Assistant Professor in corpus linguistics and computer-assisted discourse analysis at the University of Nice in France. She teaches courses on discourse analysis and corpus exploration. As a researcher, she is currently active in the corpus community, in the fields of computer-mediated communication and corpus exploration using corpus-linguistics, textual data analysis and NLP methods. She is an active member of the TGIR Huma-Num CORLI consortium, a national consortium for the study of Language, Corpora and Interactions. She is a member of the steering committee and she participates in the coordination of the working groups on Multimodality and CMC, and Interoperability and corpus exploration. She is also a member of the National Council of Universities – 7e section Linguistics.

Mohamed Tristan Purvis is Assistant Professor for English Language and Linguistics in the School of Arts and Sciences at the American University of Nigeria, where he teaches courses in writing and linguistics. His research interest lies in discourse analysis, corpus linguistics, and language documentation, and his areal focus in African languages and linguistics has led him to pursue research in Ghana, Ethiopia, Nigeria, and Kenya.

Tatjana Scheffler teaches Computational Linguistics in the Department of Linguistics at the University of Potsdam, Germany. She received her PhD in Linguistics from the University of Pennsylvania, and has worked as a researcher in intelligent multimodal interfaces at the German Research Center for Artificial Intelligence (DFKI). Her research interests are discourse and dialog, the analysis of computer-mediated communication and computational social science. She uses formal theoretical linguistic as well as corpus linguistic and computational methods. She is Co-PI of a research project on Discourse Strategies across Social Media and is investigating linguistic variability within individuals and across channels, within the Collaborative Research Cluster on Limits of Variability in Language at the University of Potsdam.

Ludovic Tanguy is Assistant Professor at the Department of Language Sciences, University of Toulouse. He teaches computational linguistics, natural language processing and data-based linguistics. He is a member of the CLLE research laboratory; his current research topics are corpus-based computational semantics and its various applications in NLP.

Lieke Verheijen is a PhD candidate at the Department of Dutch Language and Culture in the Faculty of Arts, Radboud University in Nijmegen, the Netherlands. She also works as a lecturer at the Department of Communication and Information Sciences at Tilburg University, specifically in the track Business Communication and Digital Media. She has a background in English Language and Culture (BA and MA degrees) and Language and Communication (Research Master degree). She teaches courses on content analysis, corporate (online) communication, research skills, and academic English. Her research focuses on language use in Dutch social media and the effects of such informal computer-mediated communication on the literacy skills of young people.

Torsten Zesch is Assistant Professor and Chair of the Language Technology Lab at the Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen. He holds a doctoral degree in Computer Science from Technische Universität Darmstadt and has worked as a substitute professor at the German Institute for International Pedagogical Research. He is currently the Co-President of the German Society for Computational Linguistics and Language Technology. His research interests include the processing of non-standard, error-prone language as found in social media and learner language.

Name index

Univerza *v Ljubljani*
FILOZOFSKA
FAKULTETA

ISBN 978-961-237-950-6
9 789612 379506 >
20,00 €

Univerza v Ljubljani
FILOZOFSKA
FAKULTETA

Darja Fišer is Assistant Professor and Chair of the Unit for lexicology, terminology and language technologies at the Department of Translation Studies of the Faculty of Arts, University of Ljubljana and Research Associate at the Department of Knowledge Technologies at the Jožef Stefan Institute. She teaches courses on corpus linguistics and translation technologies. As a researcher, she is currently active in the fields of computer-mediated communication and lexical semantics using corpus-linguistics methods and natural language processing. She is President of the Slovenian Language Technologies Society, Chair of the FoLLI Steering Committee of the biggest European summer school on language, logic and computation ESSLLI and Director of User Involvement of the European research infrastructure for language resources and technology CLARIN ERIC.

Michael Beißwenger is Professor of German Linguistics and Language Teaching at the Department of German Studies of the Faculty of Humanities, University of Duisburg-Essen. Besides the field of computer-mediated communication, which he has been researching since 1999, his research interests include corpus linguistics, digital humanities, text technology, collaborative writing, and the development of e-learning scenarios for language teaching and higher education. He is one of the initiators and a member of the steering committee of the annual Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora.org), convener of the TEI special interest group for computer-mediated communication and member of the CLARIN-D working groups for German Philology and for Applied and Computational Linguistics.

ISBN 978-961-237-950-6
9 789612 379506 > 20,00 €

Edited by
Darja Fišer and
Michael Beißwenger

INVESTIGATING COMPUTER-MEDIATED COMMUNICATION:
CORPUS-BASED APPROACHES TO LANGUAGE IN THE DIGITAL WORLD

Univerza v Ljubljani
FILOZOFSKA
FAKULTETA

TRANSLATION STUDIES
AND APPLIED
LINGUISTICS

Univerza v Ljubljani
FILOZOFSKA
FAKULTETA

Edited by **Darja Fišer** and **Michael Beißwenger**

INVESTIGATING COMPUTER-MEDIATED COMMUNICATION:

CORPUS-BASED APPROACHES TO LANGUAGE IN THE DIGITAL WORLD

The increasing popularity of Web 2.0 has resulted in an unprecedented surge of user-generated and social media content which is becoming a major source of knowledge and opinion, and is considered a catalyst of bottom-up communication practices that contribute towards the democratization of language. As a consequence, we are seeing a growing need for a thorough multidisciplinary understanding of this type of communication that is significantly shaped by the specific social and technical circumstances in which it is produced: rich in colloquialisms and foreign language elements, non-canonical spelling variants and syntax, idiosyncratic abbreviations and neologisms.

This volume brings together researchers active in the initiative called Computer-Mediated Communication and Social Media Corpora for the Humanities (http://www.cmc-corpora.org/) that is dedicated to the discussion of best practices on all aspects of open issues regarding the development, annotation, processing and analysis of corpora of computer-mediated communication (CMC). It includes eight chapters that have been written by 16 authors from 13 different countries and deal with the creation of CMC corpora, and with the analysis of CMC phenomena in 10 different languages. They tackle a diverse range of research questions and use a rich set of approaches, which is why they are organized into four broad thematic and methodological parts: Part 1 - Lexical analysis of CMC, Part 2 - Sociolinguistic analysis of CMC, Part 3 - Conversation and conflict in CMC, and Part 4 - Building and processing CMC resources.

Darja Fišer and Michael Beißwenger
Editors