Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Proceedings of the
4th Conference on
CMC and Social Media Corpora
for the Humanities

27–28 September 2016
Faculty of Arts, University of Ljubljana
Ljubljana, Slovenia

Editors:

Darja Fišer
Michael Beißwenger

PROCEEDINGS OF THE 4TH CONFERENCE ON CMC AND SOCIAL MEDIA CORPORA FOR THE HUMANITIES

Editors: Darja Fišer, Michael Beißwenger
Technical editors: Jaka Čibej, Katja Zupan

Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
(Ljubljana University Press, Faculty of Arts)
Issued by: Department of Translation Studies
For the publisher: Branka Kalenić Ramšak, the dean of the Faculty of Arts

Ljubljana, 2016
First edition

Conference web site: http://nl.ijs.si/janes/cmc-corpora2016/

Publication is available at:
http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-conference-proceedings-2016.pdf

The publication was supported by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017).

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

COBISS.SI-ID=286436096

ISBN 978-961-237-859-2 (pdf)

Preface

This volume presents the proceedings of the 4th edition of the *Conference on CMC and Social Media Corpora for the Humanities* (*cmc-corpora2016*) which was held on September 27–28 at the University of Ljubljana, Slovenia. The conference series (http://cmc-corpora.org/) is dedicated to the collection, organization, annotation, processing, analysis and sharing of data and corpora from computer-mediated communication (CMC) and social media genres for research purposes. The genres of interest to the cmc-corpora conference community include e-mail, chats, forums, newsgroups, blogs, news comments, wiki discussions, SMS and mobile messaging applications (WhatsApp, etc.), interactions on social network sites (Facebook, Twitter etc.), on YouTube and in multimodal online environments. The conference brings together research questions from linguistics, philology, communication sciences, media and social sciences with methods, tools and infrastructures from the fields of corpus and computational linguistics, natural language processing, text technology and digital humanities. The focus of the conferences is on

- language-centered research using computational methods and tools for the empirical analysis of CMC and social media phenomena,
- approaches towards automatic processing and annotation of CMC and social media data,
- corpus-linguistic research on collecting, processing, representing and providing CMC and social media corpora on the basis of standards in the field of digital humanities.

Previous conferences have been held in Dortmund/Germany (2013 and 2014) and in Rennes/France (2015).

Besides keynote talks by two invited speakers, *Dawn Knight* from Cardiff University (UK) and *Petra Kralj Novak* from the Jožef Stefan Institute (Slovenia), the 4th cmc-corpora conference featured 17 papers, 4 posters and 1 student paper written by 40 authors and co-authors from 24 research institutions in 11 countries, addressing key issues and current trends in the research field on data from 8 different languages.

We thank all colleagues who have contributed to the conference and to this volume with their papers, talks and posters, and as members of the scientific committee. We hope that the results of the conference will mark another step towards a lively exchange of approaches, expertise, resources, tools and best practices between researchers and existing networks in the field and pave the ground for future standards in building and using CMC and social media corpora for research in the humanities.

Darja Fišer, University of Ljubljana, Slovenia
Michael Beißwenger, University of Duisburg-Essen, Germany

Co-chairs of the Scientific Committee

Ljubljana and Essen, September 2016

Coordinating Committee

Michael Beißwenger (UDE, Germany)
Ciara R. Wigham (ICAR, France)
Thierry Chanier (LRL, France)

Scientific Committee

Co-chairs

Darja Fišer (UL, Slovenia)
Michael Beißwenger (UDE, Germany)

Members

Thierry Chanier (LRL, France)
Isabela Chiari (SAPIENZA, Italy)
Tomaž Erjavec (JSI, Slovenia)
Axel Herold (BBAW, Germany)
Gudrun Ledegen (UR2, France
Nikola Ljubešić (UZ, Croatia)
Julien Longhi (UCP, France)
Harald Lüngen (IDS, Germany)
Maja Miličević (UB, Serbia)
Céline Poudat (UN, France)
Egon W. Stemle (EURAC, Italy)
Ciara R. Wigham (ICAR, France)

Organizing Committee

Chair

Darja Fišer (UL)

Members

Simon Krek (JSI)
Jaka Čibej (UL)
Katja Zupan (JSI)

Organizers

SDJT

The Slovenian Language Technologies Society

CLARIN.SI

Common Language Resources and
Technology Infrastructure, Slovenia

Univerza *v Ljubljani*

Table of Contents

Table of Contents

Constructing E-Language Corpora: a focus on CorCenCC (The National Corpus of Contemporary Welsh)

Dawn Knight
Centre for Language and Communication Research, Cardiff University,
2 Column Drive, CF10 Cardiff, UK
E-mail: knightd5@cardiff.ac.uk

Abstract

Digital communication in the age of 'web 2.0' (that is the second generation of in the internet: an internet focused driven by user-generated content and the growth of social media) is becoming ever-increasingly embedded into our daily lives. It is impacting on the ways in which we work, socialise, communicate and live. Defining, characterising and understanding the ways in which discourse is used to scaffold our existence in this digital world is, therefore, emerged as an area of research that is a priority for applied linguists (amongst others). Corpus linguists are ideally situated to contribute to this work as they have the appropriate expertise to construct, analyse and characterise patterns of language use in large-scale bodies of such digital discourse (labelled 'e-language' here). Indeed, an increasing amount of e-language corpora are being developed to allow us to investigate e-language use.

This presentation discusses some of the methodological, technical, practical and ethical considerations and challenges faced in the construction of e-language corpora. It will outline, for example, some of the approaches used when planning the construction of e-language corpora including: obtaining consent; approaches to sampling, collecting and anonymising data; sourcing and attributing metadata, as well as some reflections on constructing a corpus infrastructure.

Discussions will be contextualised with reference to the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC)-funded CorCenCC corpus (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh) project. CorCenCC will be the first large-scale corpus of Welsh representative of language use across communication types, including 2 million words of e-language and 4 million words each of spoken and written language. CorCenCC will be open-source and freely available for use by professional communities and anyone with an interest in language. Bespoke applications and instructions will be provided for different user groups. The corpus will enable, for example, community users to investigate dialect variation or idiosyncrasies of their own language use; professional users to profile texts for readability or develop digital language tools; to learn from real life models of Welsh; and researchers to investigate patterns of language use and change.

Sentiment of Emojis

Petra Kralj Novak
Department of Knowledge Technologies, Jožef Stefan Institute,
Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: petra.kralj.novak@ijs.si

Abstract

Emojis are one of the phenomenons of the technological age. What started out as the odd smiley face at the end of a text message :) has evolved into being an indispensable part of informal computer mediated communication. For example, Instagram reports that in March 2015 nearly half of the texts on their platform contained emojis. But what is the emotional context of emojis? We engaged 83 human annotators to label over 1.6 million tweets in 13 European languages by the sentiment polarity (negative, neutral, or positive). About 4% of the annotated tweets contain emojis. By computing the sentiment of emojis from the sentiment of the tweets in which they occur, we constructed the first emoji sentiment lexicon, called the Emoji Sentiment Ranking, and draw a sentiment map of the 751 most frequently used emojis. The sentiment analysis of the emojis allows us to draw several interesting conclusions. It turns out that most of the emojis are positive, especially the most popular ones. The sentiment distribution of the tweets with and without emojis is significantly different. The inter-annotator agreement on the tweets with emojis is higher. Emojis tend to occur at the end of the tweets, and their sentiment polarity increases with the distance. We observe no significant differences in the emoji rankings between the 13 languages and the Emoji Sentiment Ranking. Consequently, we propose our Emoji Sentiment Ranking as a European language-independent resource for automated sentiment analysis.

Syntactic Annotation of Slovene CMC: First Steps

Špela Arhar Holdt[*♦], Darja Fišer[*‡], Tomaž Erjavec[‡], Simon Krek[‡]

[*] Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
[♦]Institute for Applied Slovene Studies Trojina
Trg republike 3, 1000 Ljubljana
[‡] Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
E-mail: spela.arharholdt@ff.uni-lj.si, darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si, simon.krek@ijs.si

Abstract

This paper presents the first steps towards the syntactic annotation of Slovene CMC, namely the annotation of 200 Slovene tweets with the JOS dependency model. After a presentation of the dataset we present the selected annotation model, the annotation procedure, and results. The focus of the paper is on the decisions regarding the annotation of CMC-specific elements that required special treatment: Twitter-specific features, foreign language elements, ellipsis and fragments, non-standard use of punctuation, and other non-standard language features. The dataset, together with the CMC-adapted annotation guidelines, can be used for further annotation of language data (from Twitter or other CMC genres), and in the second step to train a parser for the selected CMC domain(s). The large-scale corpus-based research of non-standard Slovene syntax, which will be facilitated by the described activities, will help disprove the myths surrounding CMC that are still present in the field of Slovene studies.

Keywords: computer mediated communication, syntactic annotation, JOS dependency model, Slovene language, tweets

1. Introduction

With the advent of digital media and the Internet, communication practices began to change significantly, challenging the traditionally established dichotomies of public vs. private, formal vs. informal, written vs. spoken, and standard vs. non-standard language use. Initially, the linguistic research community observed the new situation with a somewhat reserved attitude, whereas in the last years, more and more studies aim to disprove the myths surrounding computer mediated communication and its possible negative impact on the evolution of language (Crystal, 2011). Since computer-mediated communication is a global phenomenon, work on languages other than English soon followed (Myslin and Gries, 2010; Storrer, 2013; Chanier, 2015).

While studies have been performed on Slovene as well, they mostly focused on orthographic (Jakop, 2008; Arhar Holdt and Dobrovoljc, 2015), lexical (Michelizza, 2015; Zwitter Vitez and Fišer, 2015) and processing issues (Ljubešić et al., 2016a; Ljubešić et al., 2016b) whereas no larger-scale corpus-based work exists on the syntax of Slovene CMC. The goal of this paper is to present the first steps in bridging this gap, the annotation of 200 Slovene tweets with the JOS dependency model (Erjavec et al., 2010), which will serve as the groundwork for syntactic annotation and analysis of Slovene CMC.

2. Dataset

A dataset of 200 tweets (475 sentences) was extracted from the Janes corpus of Slovene CMC (Fišer et al., 2015), sampled to include an equal amount of linguistically and technically standard and non-standard tweets (Ljubešić et al., 2015). The dataset only includes tweets longer than 120 characters published by private individuals. This material was lemmatized and POS-tagged with the tools described in (Erjavec et al., 2005; Ljubešić et al., 2014). In the next step, the sentence segmentation and tokenization was manually corrected, the tweets were normalised on the lexical and morphological level (Čibej et al., 2016a), and finally, the attributed lemmas and POS-tags were manually corrected (Čibej et al., 2016b).

3. The JOS Dependency Model

For the annotation, the JOS dependency model was used. The system, which was designed in the project "Linguistic Annotation of Slovene" (Erjavec et al., 2010), is based on syntactic dependencies. The categories of the system are presented in Table 1.

Groups of labels	Labels	Description
First level labels link elements in different types of phrases (green and yellow colour in the visualisation).	*dol*	Links heads and modifiers in phrases.
	del	Links parts of verbal phrases.
	prir	Links heads in coordinate structures within clauses.
	vez	Links words or commas in conjuctive function.
	skup	Links (function) words in frozen multi-word structures.
Second level labels link sentence elements (red colour in the visualisation).	*ena*	Clause subject.
	dve	Clause object.
	tri	Adverbial of manner.
	štiri	Other adverbials.
Third level label links all other structures (blue colour in the visualisation).	*modra*	Links to the root, punctuation, fragments, etc.

Table 1: The labels in the JOS dependency model.
(http://eng.slovenscina.eu/tehnologije/razclenjevalnik)

As the JOS dependency model is based on the principle that the relations inducible from the tags on lower levels (lemmas and POS) are not annotated again on the syntactic level, it is significantly simpler and more robust than similar models, e.g. the Prague Dependency Treebank (Böhmová et al., 2003). The main features of the model are described in Erjavec et al. (2010), and in more detail in the annotation guidelines (Holozan et al., 2008). The model was applied in the "Communication in Slovene (SSJ)" project to annotate the ssj500k training corpus (Krek et al., 2015), on the basis of which a parser for Slovene was trained (Dobrovoljc et al., 2012). Additionally, a specialised program was developed for the visualisation, manual annotation and search of the data (the screenshots on Figures 1 to 4 are from this program, the author of the program is Janez Brank).

4. Annotation and Results

The dataset, described in Section 2, was automatically parsed and imported into the SSJ annotation program. Syntactic annotations were then manually corrected, following the guidelines for the annotation of the jos500k corpus. During annotation, the majority of the problems could be adequately addressed by the existing guidelines, while for some specific questions, the guidelines had to be complemented by additional rules. In the remainder of this paper, we present the decisions regarding the annotation of: Twitter elements; foreign language; syntactical fragments and ellipsis; non-standard use of punctuation; and other non-standard language features. The implement solutions are exemplified in Figures 1 to 4. The examples are in Slovene, with English translation provided in the corresponding figure title.[1]

4.1 Twitter Elements

We considered hashtags, usernames, URLs, and emoticons of two kinds. The elements that were syntactically part of a sentence were annotated in accordance with their function, while function-free elements (typically appearing at the beginning or the end of the tweet) were connected to the node. This decision is in accordance with similar projects (Kong et al., 2012), and the annotation of the dataset indicates that the separation of the two groups is sufficiently straightforward. Figure 1 presents an example, where the first hashtag (*#zooljubljana*) connects to the node, while the second one is annotated as a part of a noun phrase (*a plane to #sochi*).

4.2 Foreign Language

Foreign language elements (primarily from English and related South Slavic languages) appear in Slovene tweets as single words, word phrases, or longer segments/clauses. Different levels of adaptation to Slovene can be observed regarding the spelling and morphology of these elements. The questions about how to lemmatise and POS-tag them (including the question how to separate the ones to be tagged as *foreign* from the ones to be treated as *Slovene*) were addressed at the earlier stages of the project (Čibej et al., 2016b). On the syntactic level, we followed a principle that single words and two-part phrases with a clear dependency relation are attached into the syntactic tree, whereas in longer phrases and segments, all of the foreign elements get attached to the node instead. Figure 2 presents an example of the first type, where the English phrase *personal message* is connected to the tree.

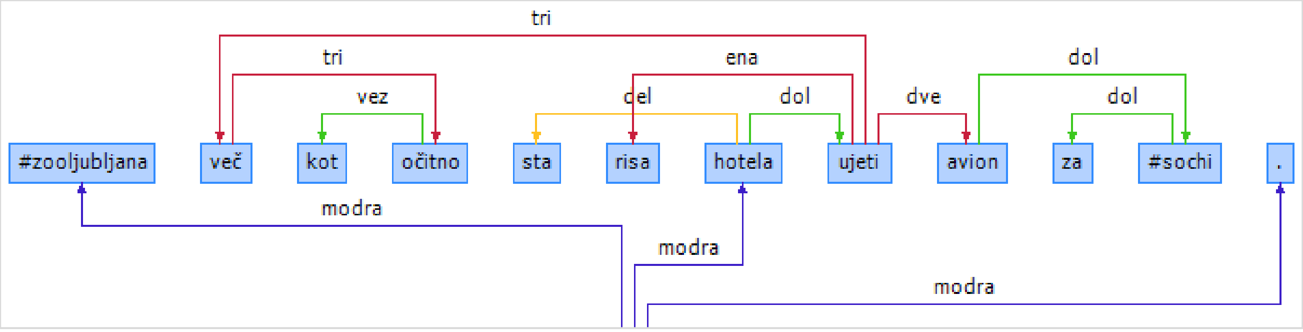


Figure 1: ***#zooljubljana*** *more than obviously the lynx wanted to catch* ***a plane to #sochi****.*

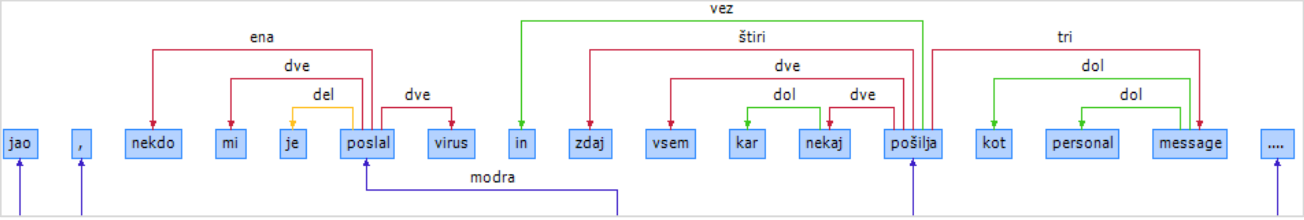


Figure 2: *Jeez, somebody sent me a virus and now it's sending random stuff to everyone as a {****personal message****}.*

[1] The translation is somewhat word-by-word to facilitate comprehension of annotated syntactic relations to non-Slovene speakers, however certain adaptations were obviously required due to language differences.

4.3 Ellipses and Fragments

Tweets are especially challenging to annotate syntactically due to their fragmented nature and a large number of ellipses. One possible solution to this problem is to use a system that allows for the orphan node to be promoted to the place of the missing parent (Dobrovoljc and Nivre, 2016). Another alternative is to use a system that attaches such elements directly to the root (Kong et al., 2012). The JOS dependency model was designed as the latter: while in regular clauses only the head of the predicate is attached to the root (or the relevant ordinate clause), in fragments or clauses without the predicate, each separate phrase head attaches to the node as well (Figure 3).

4.4 Non-standard Use of Punctuation

The annotation of the dataset revealed that lexical and morphological normalisation of tweets and subsequent manual correction of lemmas and POS tags successfully eliminated many of the potential problems for syntactic annotation. However, the non-standard use of punctuation in tweets remains an important factor of negative influence. The existing parser is trained on standard Slovene language, where punctuation – especially the use of the comma – plays an important role in determining the borders between clauses and other types of sentence segments. Omitted, redundant, and misplaced commas thus as a rule lead to parsing mistakes, and with the comma being notoriously difficult to master for Slovene speakers, such instances are frequent. For the annotation of the dataset, the parsing errors were manually corrected, however the findings indicate that it might be beneficial to include a step of punctuation normalisation before the attempts on the syntactical level (some work for Slovene has been presented by Kranjc and Robnik Šikonja, 2015).

4.5 Other Non-standard Language Features

Last but not least, the annotated dataset exhibits a number of other syntactic features that have been previously attributed to non-standard written Slovene (Michelizza, 2015), e.g. atypical word order, non-standard use of conjunctions, cases, grammatical number, high number of demonstrative pronouns and certain particles. A preliminary analysis of the annotated data reveals that 49 % of the (linguistically and technically non-standard) tweets exhibit at least one of the listed features. While it is clear that these phenomena need to be linguistically addressed in the future, they did not pose a problem for the annotation. Figure 4 presents an example, where the predicate *to be similar* is accompanied by two objects in dative (*he is similar to the members of the parliament* and *similar to me = to me he seems similar*). While valence in this example is atypical, the annotations are relatively straightforward.

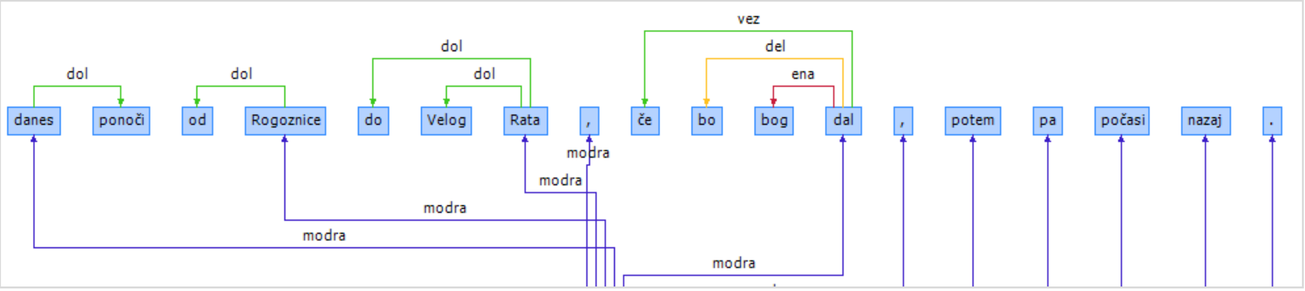


Figure 3: *During the night from Rogoznica to Veli Rat, if god allows it, and then slowly back.*

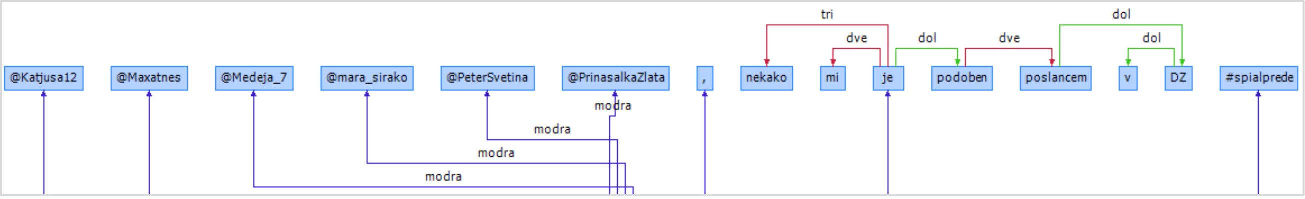


Figure 4: *[...] @PrinasalkaZlata, somehow* ***to me*** *he is similar* ***to the members of the parliament*** *#spialprede.*

5. Conclusion and Further Work

The paper presented the first steps towards a syntactic annotation of Slovene CMC. In this first stage, 200 tweets were annotated with the JOS dependency model and annotator guidelines were supplemented with examples for the annotation of Twitter-specific and non-standard language features. The dataset, together with the guidelines, can be used for further annotation of language data (from Twitter or other CMC genres), and in the second step to train a parser for the selected CMC domain(s).

The JOS dependency model in combination with the SSJ annotation program proved to be adequate for the described task, the main advantages of the system being its robustness and the ability to allow multiple attachments to the root element. A major drawback is that the system is language-specific and as such offers little possibility for cross-lingual comparison. Recent attempts to translate the annotations of the ssj500k corpus to the Universal Dependencies system (Dobrovoljc et al., 2016) suggest a possible solution to this problem in the future.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Non-Standard Internet Slovene" (J6-6842, 2014–2017).

7. References

Arhar Holdt, Š. and Dobrovoljc, K. (2015). Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. In D. Fišer (Ed.), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 4–9.

Böhmová, A., Hajič, J., Hajičová, E. and Hladká, B. (2003). The Prague dependency treebank. In *Treebank: Building and Using Parsed Corpora*. Netherlands: Springer, pp. 103–127.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J. and Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, 29(2), pp.1-30.

Čibej, J., Fišer, D. and Erjavec, T. (2016a). Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. In *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož: ELRA, pp. 5–10.

Čibej, J., Arhar Holdt, Š., Erjavec, T. and Fišer, D. (2016b). Razvoj učne množice za izboljšano označevanje spletnih besedil. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).

Crystal, D. (2011). *Internet Linguistics: A Student Guide.* London, New York: Routledge.

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16).* Portorož, pp. 1566–73.

Dobrovoljc, K., Erjavec, T. and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).

Dobrovoljc, K., Krek, S. and Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. In T. Erjavec, J. Žganec Gros (Eds.), *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan, pp. 42–47.

Erjavec, T., Fišer, D., Krek, S. and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In: *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta, pp. 1806–1809.

Erjavec, T., Ignat, C., Pouliquen, B. and Steinberger, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference*. Poznan, pp. 32–36.

Fišer, D., Ljubešić, N. and Erjavec, T. (2015). The JANES corpus of Slovene user generated content: construction and annotation. In *International Research Days: Social Media and CMC Corpora for the eHumanities: Book of Abstracts*. Rennes,.p. 11.

Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. and Velušček, A. (2008). *Specifikacije za učni korpus.* Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.

Jakop, N. (2008). Pravopis in spletni forumi – kva dogaja? In *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*, pp. 315–327.

Kranjc, A. and Robnik Šikonja, M. (2015). Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. In D. Fišer (Ed.), *Zbornik konference Slovenščina na spletu in v novih medijih,* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 38–43.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C. and Smith, N. A. (2014). A dependency parser for tweets. In *Proc. of EMNLP*. Doha, Qatar, pp. 1001–1012.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N. and Holz, N. (2015). *Training corpus ssj500k 1.4, Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference.* Hissar, pp. 371–378.

Ljubešić, N., Erjavec, T. and Fišer, D. (2014). Standardizing tweets with character-level machine translation. In *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal: Proceedings: part II.* Springer, Heidelberg, pp. 164–175.

Michelizza, M. (2015). *Spletna besedila in jezik na spletu.* Založba ZRC, ZRC SAZU, Ljubljana.

Myslin, M. and Gries, S. T. (2010). k dixez? A corpus study of Spanish Internet orthography. *Literacy and Linguistic Computing*, 25(1), pp. 85–104.

Storrer, A. (2013). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013.* De Gruyter Mouton, pp. 171–196.

Zwitter Vitez, A. and Fišer, D. (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*, *Herstmonceux Castle, UK.* Ljubljana: Trojina, Institute for Applied Slovene Studies; Birmingham: Lexical Computing, pp. 250–267.

(Best) Practices for Annotating and Representing CMC and Social Media Corpora in CLARIN-D

Michael Beißwenger*, Eric Ehrhard[†], Axel Herold[ɣ], Harald Lüngen[‡], Angelika Storrer[†]

* Department of German Studies, University of Duisburg-Essen, Berliner Platz 6–8, D-45127 Essen
[†] Department of German Linguistics, University of Mannheim, Schloss, Ehrenhof West, D-68131 Mannheim
[ɣ] Berlin-Brandenburg Academy of Sciences and Humanities, Jägerstraße 22/23, D-10117 Berlin
[‡] Institute for the German Language, R5, 6–13, D-68161 Mannheim

E-mail: michael.beisswenger@uni-due.de, eric.ehrhardt@gmx.de, herold@bbaw.de,
luengen@ids-mannheim.de, astorrer@mail.uni-mannheim.de

Abstract

The paper reports the results of the curation project *ChatCorpus2CLARIN*. The goal of the project was to develop a workflow and resources for the integration of an existing chat corpus into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences (http://clarin-d.de). The paper presents an overview of the resources and practices developed in the project, describes the added value of the resource after its integration and discusses, as an outlook, to what extent these practices can be considered *best practices* which may be useful for the annotation and representation of other CMC and social media corpora.

Keywords: CMC corpora, TEI encoding, tagging, corpus infrastructures, legal issues, CLARIN

1. Introduction

This paper reports the results of the curation project *ChatCorpus2CLARIN*. The goal of the project was to develop a workflow and resources for the integration of an existing chat corpus (the *Dortmund Chat Corpus*, Beißwenger 2013) into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences[1] as part of the European *Common Language Resources and Technology Infrastructure*[2]. The paper presents an overview of the resources and practices developed in the project, describes the added value of the resource after its integration and discusses, as an outlook, to what extent these practices can already be considered as *best practices* which may be useful for the annotation and representation of other CMC and social media corpora.

2. Goals of the Project

The goal of the project was twofold: On the one hand, (1) the project aimed to integrate an existing chat corpus into the CLARIN-D corpus infrastructures at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and at the Institute for the German Language (IDS), Mannheim. This included, as subtasks, (1a) the development of a schema and conversion routine for the transformation of the XML markup and metadata in the original resource into a TEI format, (1b) the addition of a new annotation layer with part-of-speech and lemma information, (1c) a re-anonymization of the corpus data according to the recommendations given in a legal opinion. On the other hand, (2) the solutions developed to achieve goal (1) should be designed as general (and not idiosyncratic) approaches to the challenge of annotating and representing corpora of computer-mediated communication (CMC) and social media according to existing standards in the Digital Humanities / CLARIN context. The main result of goal (1) is, thus, the integrated chat corpus whereas the results of goal (2) are documented resources and practices that may be reused by other projects which aim at integrating CMC and social media resources into CLARIN.

3. The Corpus

The *Dortmund Chat Corpus* (Beißwenger, 2013) has been collected at TU Dortmund University as a resource for researching the peculiarities and linguistic variation in written CMC. The corpus comprises 478 chat documents (*logfiles*) containing 140240 user postings or 1M words of German chat discourse from heterogeneous sources representing the use of chats in a wide range of application contexts (social chats, advisory chats, chats in the context of learning and teaching, moderated chats in the media context). The corpus has been annotated using a homegrown XML format (‘ChatXML’) that describes (1) the basic structure and properties of chat logfiles and postings, (2) selected “netspeak” phenomena such as emoticons, interaction words, addressing terms, nicknames and acronyms, (3) selected metadata about the chat platforms and chat users. Since 2005, a large subset of the corpus has been available as a ChatXML resource for download and offline querying, and as an HTML version for online browsing.[3]

4. Overview of Workflow and Resources

The Dortmund Chat Corpus served as a use case to demonstrate how an integration of CMC and social media resources could be accomplished in a way that the target resource (1) conforms to established stan-

[1] http://clarin-d.de

[2] https://www.clarin.eu

[3] http://www.chatkorpus.tu-dortmund.de

dards for the representation and linguistic annotation of corpora in the Digital Humanities context and (2) can be used for comparative analyses with other types of corpus resources in CLARIN-D (text and speech corpora). A visualization of the workflow and practices developed in the project is given in Fig. 1; the steps and resources of the pipeline are described in the following subsections.

4.1 Experimental CMC Corpus with Data Samples from Heterogeneous Sources

For developing and testing the solutions for goals (1a), (1b) and (1c) (cf. Sect. 2) not only with chat data, we compiled a small experimental corpus of 38382 tokens with data also from other CMC and social media genres. The corpus included (1) two logfiles from different subcorpora of the chat corpus (12526 tokens), (2) 94 news messages from the Usenet corpus in DeReKo (Schröck & Lüngen, 2015) (9108 tokens), (3) excerpts from two Wikipedia talk pages (907 tokens), (4) donated tweets from two different twitter accounts (1412 tokens) and (5) 1907 posts from two different whatsapp conversations collected in the project "What's up, Deutschland?"[4] (14429 tokens).

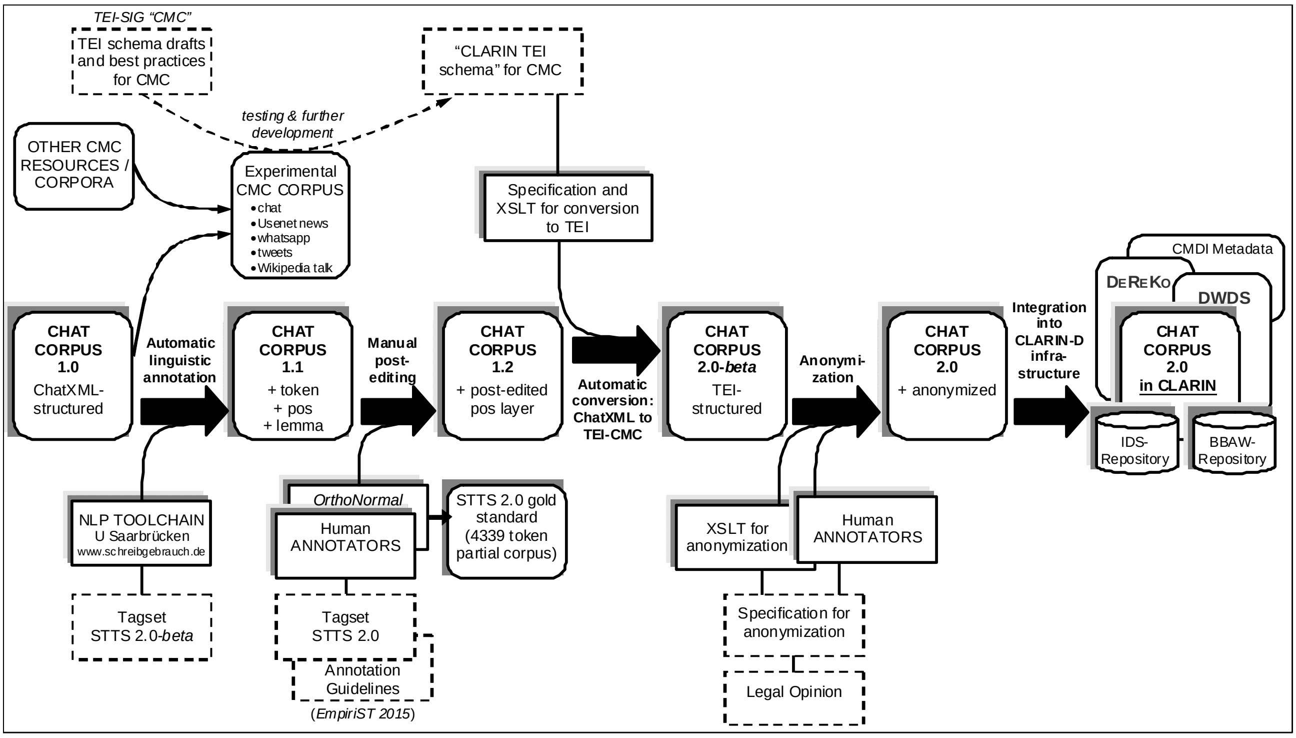


Figure 1: Workflow and resources.

4.2 The NLP Toolchain Developed in the BMBF Project www.schreibgebrauch.de

Part-of-speech (PoS) tagging was done in two stages: (1) an automatic tagging process and (2) a manual post-editing phase. Automatic tagging (including tokenization, PoS tagging and lemmatization) was done at Saarland University applying an NLP toolchain that was developed in the BMBF project *Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen*[5] (Horbach et al., 2014). The toolchain had originally been trained on annotating chat and forum data with a tag set derived from Bartz et al. (2014).

4.3 The 'STTS 2.0' Part-of-Speech Tagset and Guidelines from the EmpiriST2015 Shared Task Project

As target standard for the PoS layer, we used the STTS-IBK tag set ('STTS 2.0') developed in the GSCL shared task on automatic linguistic annotation of CMC and social media (EmpiriST2015)[6]. 'STTS 2.0' is an advanced version of the tag set suggested in Bartz et al. (2014) and builds on the categories of the "Stuttgart-Tübingen Tagset" (*STTS*, Schiller et al., 1999) which is a well-acknowledged defacto standard for PoS tagging of German written corpora. In its canonical version, STTS does not include any tags for CMC and social media genres. 'STTS 2.0' therefore introduces two types of new tags: (1) tags for phenomena which are specific for CMC and social media discourse, (2) tags for phenomena which are typical of spontaneous spoken language in colloquial registers and which can also be found in corpora of transcribed speech (e.g., in the FOLK corpus of spoken language at the IDS which uses an STTS extension which is compatible with 'STTS 2.0', Westpfahl, 2014). The resulting tag set is still downwardly compatible with STTS (1999) and therefore allows for interoperability with other corpora

[4] http://www.whatsup-deutschland.de

[5] http://www.schreibgebrauch.de

[6] http://sites.google.com/site/empirist2015/

that have been tagged with STTS. In the EmpiriST2015, several existing NLP systems have been trained on assigning the 'STTS 2.0' extensions to tokens of CMC and social media discourse (Beißwenger et al., 2016). The tag set is described in an annotation guideline (Beißwenger et al. 2015) and had previously been tested with data from several CMC genres. In the curation project, these guidelines have been used for manual post-editing the results of the automatic tagging process described in Sect. 4.2. In the post-editing process which was done using an adapted version of the tool *OrthoNormal* from the *FOLKER* tool suite (Schmidt, 2012), the whole corpus has been made compatible with the 'STTS 2.0' tag set. In addition, for a partial corpus of 4339 tokens all tags assigned in the automatic process have been post-edited independently by two human annotators who had been trained with the guidelines (agreement according to Cohens Kappa: κ = 0.92). Differing cases were decided by the project heads. The 4339 partial corpus with manually checked PoS annotation can be considered as an additional resource from the project which can be used for further retraining of tagging systems with 'STTS 2.0'.

4.4 The 'CLARIN TEI Schema for CMC' and the XSLT for Conversion

The resource was converted into a TEI representation format which builds on (1) the official TEI-P5 framework for electronic text encoding and interchange and (2) two versions of a customization of TEI-P5 for CMC genres created in the context of the TEI special interest group "computer-mediated communication" (CMC-SIG) and described in Beißwenger et al. (2012) and Chanier et al. (2014). Starting from a close evaluation of the most recent version of the customization Chanier et al. (2014), we developed the models and best practices from the TEI CMC-SIG further taking into consideration the genres available in our experimental corpus. The resulting new TEI schema draft – the 'CLARIN TEI schema for CMC' – has been made available for further use and comments in the TEI wiki[7]. The conversion of the ChatXML format into the target TEI format was done using an XSLT stylesheet.

4.5 Representation of Metadata in TEI

In contrast to the customizations needed for the markup of the primary discourse data, we did not modify the existing TEI metadata model. All metadata provided in the original version of the corpus (which was partially given as part of the ChatXML structure, partially as textual descriptions provided in the corpus-external documentation of the corpus data) could be re-modelled using their TEI equivalents within the teiHeader. Special attention was paid to the modeling of a text classification scheme which is associated with the corpus documents by means of the TEI's generic textClass/catRef mechanism. This model can be easily extended to a broader range of text and/or discourse properties to account for more detailed classifications, such as the one proposed by Herring (2007) – work that hasn't been done within the project but which is a goal for a future extension of the schema.

4.6 Legal Opinion on Republishing the Resource in CLARIN-D – and Consequences (Anonymization)

Prior to the integration of the curated resource in CLARIN infrastructures, we sought a legal opinion to get a better picture of the legal conditions for republishing the material as a whole or in parts. The legal opinion which was provided by *iRights.Law*/John H. Weitzmann (iRights.Law, 2016) carefully checked possible restrictions arising from individual property rights, copyrights and other legal statutes. One result was that the possibility to identify individuals from their utterances (with the exception of public figures) needed to be circumvented by means of an anonymization of names, nicknames, host names and IP addresses, geographical names (e. g. address data) etc. In addition, it turned out that some (minor) parts of the resource must not be made available to the public at all, notably those parts where personality rights of participants are strongly affected. This applies to a subcorpus obtained from chat-based psycho-social counseling (a subcorpus which hadn't been made available to the public even in the original version of the corpus). For this subcorpus, due to the personal context represented in the discourse, anonymization alone is unlikely to prevent the identification of individuals. Consequently, these resources (8 logfiles containing 88227 tokens) were removed from the final corpus.

The legal opinion saw no indication of concerns regarding copyright (German "Urheberrecht", specifically) as it acknowledges that the collected logfiles as well as the individual user posts in the overwhelming majority of cases do not represent works of art. Protectable under EU (and German) law however, is the work committed in the course of collection, curation and transformation of the data into the format of the intended linguistic database. Therefore and in accordance with our goal to provide the resource as openly as possible, we followed the lawyers' suggestion to provide the resource with a CreativeCommons licence (CC BY 4.0) which allows for the protection of database creator rights.

The task of anonymization could not be done completely automatically: In a first step, names that had already been annotated in the original resource could be replaced by categorized placeholders automatically. Likewise, the metadata section and the filenames were anonymized, including names and properties of participants, and the names of chat platforms. What had to be done manually was to replace all those occurrences of names that had not been annotated in the source, or that could not be matched to entries in the participant

[7] http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema

list automatically (e.g., because chatters were addressing each other using nicknames of nicknames or referring to people who were not participating in the chat themselves). This was a very time-consuming process which at the current state could only be done for the 4339 token gold standard subset of the corpus. The anonymization of the rest of the corpus is part of a follow-up work package to be finished at the end of 2016.

5. Availability

All work packages described in Sect. 4.1–4.5 have been finished. Until October 2016, a first release of the resource will present a preview in form of the 4339 token gold standard. It is planned to make the full resource available in a 2nd release in early 2017.

The corpus will be ingested into the CLARIN repositories at the IDS[8]and the BBAW[9]. At IDS, the resource will become part of the German Reference Corpus archive DeReKo and as such will be integrated in the corpus query platform COSMAS II[10]. At BBAW, the corpus will be integrated in the corpus query platform DWDS[11]. In addition, the corpus will be made accessible through CLARIN's federated content search, e.g. for NLP toolchains such as WebLicht[12].

6. Features of the Integrated Resource

Compared with the original version of the resource, the CLARIN-integrated version ('Chat Corpus 2.0', cf. Fig. 1) will allow for advanced queries using the additional linguistic annotations (sentences, tokens, PoS, lemmas). Due to the remodeling of the resource in TEI and the compatibilty of the PoS annotations with STTS the corpus will be interoperable with other TEI-/STTS-annotated language resources. The integration into the CLARIN-D corpus infrastructures at BBAW and IDS will facilitate the comparative analysis of the chat corpus with the BBAW and IDS text and speech corpora. These features will not only increase the value of the resource for language-centered CMC research and variational linguistics but also the possibilities to use it in language teaching and higher education.

7. Outlook

According to goal (2) (cf. Sect. 2), the resources and practices developed in the project were meant to function as general approaches to open issues in representing and annotating CMC and social media data which should have the potential to be useful also for other projects in the field. To assess empirically whether the current versions of the resources (the TEI schema, the 'STTS 2.0' tagset and annotation guidelines) already have this potential, it is necessary to adopt and test these resources in other CMC corpus projects. In our own work, we tested them not only with chat data but also with a selection of data from other genres (experimental corpus, cf. Sect. 4.1). We're optimistic that the availability of the resources will facilitate corpus annotation for colleagues who are building similar corpora and who are aiming to represent them on the basis of existing standards same as we did when adopting the encoding framework of the TEI and the STTS tagset for German for our purpose. We're aware of the fact that no existing schema – not even an established standard as TEI-P5 – can usually be adopted for a new project to 100%; instead, each project typically needs their own customizations and extensions when adopting an existing solution. Nevertheless, customizing and extending a given solution is usually much easier than having to start to design a solution from scratch. Especially the TEI schema for CMC is open for further changes according to experiences and results from other projects. It will be the basis for further discussions in the TEI-SIG "computer-mediated communication" which is open for the participation to everybody who is interested to bring in their own experiences and suggestions.

In our own work, we are planning to adopt the resources and practices from the project for the integration of further CMC and social media resources into the CLARIN-D corpus infrastructures at the IDS and the BBAW (starting as of autumn 2016). The TEI schema, in addition, is currently being used and tested also in projects in which none of the authors of this paper is involved – e.g., in a weblog corpus project at the University of Gießen, Germany ('Discourse-structured Blog Corpus for German', Karlova-Bourbonus et al., 2016) and for the annotation of an English Q&A corpus at the University of California, Davis, USA (Rachael Duke, Raul Aranovich).

The 'STTS 2.0' tagset for PoS tagging CMC and social media data has been used for the EmpiriST2015 shared task in which several NLP systems have been adapted for the automatic annotation of German CMC. These systems will allow corpus projects to achieve better results in tagging their data than with standard NLP tools which have typically been trained only on 'standard' genres (newspaper corpora etc.).

The results and recommendations of the legal opinion will be a useful point of reference for further inquiries into the (still difficult) legal conditions of collecting and republishing discourse from CMC and social media sources as parts of linguistic research infrastructures.

8. References

Bartz, T., Beißwenger, M., Storrer, A. (2014). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguis-*

[8] https://repos.ids-mannheim.de/

[9] http://clarin.bbaw.de/en/repo/

[10] http://cosmas2.ids-mannheim.de/

[11] http://www.dwds.de/

[12] https://weblicht.sfs.uni-tuebingen.de/weblicht/

tics 28 (1), pp. 157–198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik* 41 (1), pp. 161–164. http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

Beißwenger, M., Bartsch, S., Evert, S., Würzner, K.-M. (2016). EmpiriST 2015: A Shared Task on Automatic Linguistic Annotation of Computer-Mediated Communication, Social Media and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26), 44–56. http://aclweb.org/anthology/W16-26

Beißwenger, M., Bartz, T., Storrer, A., Westpfahl, S. (2015). *Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation.* Guideline document from the EmpiriST2015 shared task. http://sites.google.com/site/empirist2015/home/annotation-guidelines

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., Storrer, A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative (jTEI) 3*. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Longhi, J., Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: *Journal of language Technology and Computational Linguistics (JLCL) 29 (2)*, pp. 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf

Herring, S.C. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet 4 (1)*. http://www.languageatinternet.org/articles/2007/761

Horbach, A., Steffen, D., Thater, S., Pinkal, M. (2014). Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In *Proceedings of KONVENS 2014*, pp. 171–177.

iRights.Law Rechtsanwälte (2016). *Rechtsgutachten zur Integration mehrerer Text-Korpora in die CLARIN-D-Infrastrukturen.* (Legal opinion for the ChatCorpus2CLARIN project, 46 pages)

Karlova-Bourbonus, N., Grumt Suárez, H., Lobin, H. (2016). Compilation and Annotation of the Discourse-structured Blog Corpus for German. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, University of Ljubljana [this volume].

Schiller, A., Teufel, S., Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmidt, T. (2012). EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*. http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf

Schröck, J., Lüngen, H. (2015). Building and Annotating a Corpus of German-Language Newsgroups. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015)*, pages 17–22. https://sites.google.com/site/nlp4cmc2015/proceedings.

[TEI P5] TEI Consortium (eds) (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/

Westpfahl, S. (2014). STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*. Association for Computational Linguistics (ACL Anthology W14-49), 1–10. http://www.aclweb.org/anthology/W14-4901

Grammatical Frequencies and Gender in Nordic Twitter Englishes

Steven Coats
University of Oulu, Finland
English Philology, Faculty of Humanities, 90014 University of Oulu, Finland
Email: steven.coats@oulu.fi

Abstract

English is increasingly used for online communication in many contexts in which it is not the primary local language, particularly on social media platforms with global extent such as Twitter. The grammatical properties of online and Twitter Englishes, however, have mainly been considered in L1 contexts, as have correlations between gender and some grammatical features. In this study, the correlation of grammatical types (parts of speech) and gender is undertaken for English-language Twitter messages originating from the Nordic countries. A corpus of geo-located English-language Twitter messages was created by accessing the Twitter Streaming API. After disambiguating author gender and applying part-of-speech tags, the relative frequencies of grammatical items were determined and those with significant gender divergence identified. Principal components analysis shows some gender-based separation of discourse in the Nordic countries in terms of grammatical features. The analysis supports previous findings pertaining to gendered differences in English and sheds light on how English continues to evolve in online environments.

Keywords: corpus linguistics, Twitter, world Englishes, language and gender

1. Introduction and Background

Technological developments can affect the way we interact with one another, and the recent shift towards mediated, text-based communication in online environments provides opportunities for the study of English varieties in global contexts. Although the status of English as the world's principal lingua franca continues to consolidate, its use in global computer-mediated communication (CMC), especially in non-L1 environments, exhibits a diversity of orthography, lexis, and grammar that 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 sought to characterize the communicative and discourse functions of Twitter language (Page 2012; Zappavigna 2011; Squires 2015 for an overview). The extensiveness of Twitter data, its public availability, and the richness of the associated metadata have allowed for geographical analyses (Leetaru et al. 2013; Mocanu et al. 2014) and dialectological and sociolinguistic analyses of English (Eisenstein et al. 2014; Bamann, Eisenstein and Schnoebelen 2014).

Some previous studies of English-language CMC and Twitter have found different rates of use of particular word classes by males and females. For example, it has been found that females use more personal pronouns, more modal verbs, and more 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; Bamann, Eisenstein and Schnoebelen 2014). For the most part, however, analysis of Twitter English has been conducted on data without consideration of its geographical provenance, or on data gathered from Anglophone national contexts, mostly in the United States.

Knowledge of English is extensive in the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland, countries with well-developed economies and high levels of educational attainment, to such an extent that it has been suggested that their national languages are becoming linguistic systems with "restricted functional range" (Görlach 2002: 16). Although research has addressed language use on Twitter by country (e.g. Mocanu et al. 2013), and work exists on grammatical feature frequencies in Nordic non-CMC genres (e.g. for Swedish in Allwood 1998), studies of feature frequencies in English from non-L1 environments have been few, and the relationship between author gender and feature frequency in CMC language has not yet been investigated in detail in Nordic contexts, whether in local languages or English.[1]

In this study an approach based in part on multidimensional analysis (Biber 1988; 1995) is taken. 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. With a principal components analysis, the underlying association between feature frequencies and gender is established.

2. Data Collection and Processing

Data was collected in .json format from Twitter's Streaming API during May 2016 by utilizing a scripting library in Python.[2] The raw .json data was filtered for the tweet text (the "status update") and the metadata fields `author_name`, `screen_name`, `time`, `id`, `lang` (language), `country`, and the latitude and longitude `coordinates`.

[1] 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] The *Tweepy* library (Roesslein 2015) was used (https://github.com/tweepy/tweepy).

2.1. Geolocation

The collection script selected only tweets with a populated `place` object that originated within a bounding box circumscribing the territorial boundaries of the Nordic countries (longitude -26 to 32, latitude 53 to 72; see Figure 1).

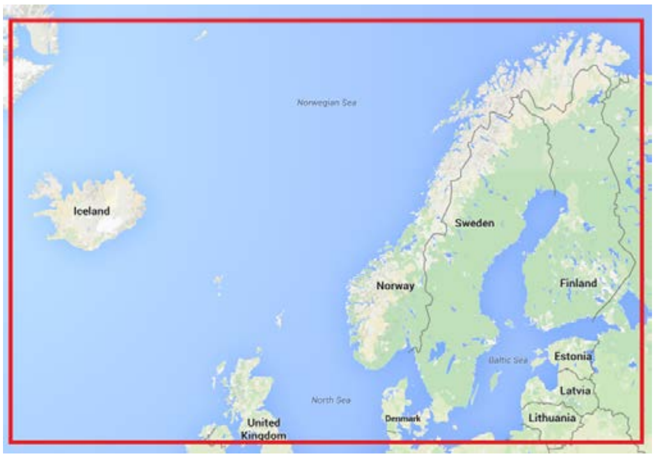


Figure 1: Area from within tweets with geographical coordinates were collected from the API.

Each tweet was assigned exact latitude/longitude coordinates.[3] From the 2.155 million tweets collected by the script, 302,737 were retained to create subcorpora from the Nordic countries of Iceland, Norway, Denmark, Sweden and Finland, based on the `country` values within the `place` field. For further analysis, two subcorpora were prepared for each country by filtering the data according to the `lang` field in the tweet object: one consisting of tweets in the principal national language, and one of tweets in English.[4] Tweets originating from outside the Nordic countries and in other languages were not further considered. The English data comprised 101,956 tweets and 1,475,553 tokens.

2.2. Gender Disambiguation

Unlike some social media platforms, Twitter does not provide a profile entry where gender is to be identified nor require users to otherwise supply gender information. Therefore, gender was disambiguated for tweets based on gender-name associations (Rao et al. 2010; Mislove et al. 2011).[5] Lists of the most frequent given names in the Nordic countries were obtained from the corresponding national statistical offices. The `author_name` field for each user was then filtered 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. Users matching both male and female names were discarded. The method assigned gender to 39% of Iceland, 50% of Norway, 61% of Denmark, 47% of Sweden, and 62% of Finland tweets.[6]

2.3. 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 subcorpora and apply part-of-speech tags using a subset of the Penn Treebank tagset (Marcus, Santorini and Marcinkiewicz 1993) and additional tags for the Twitter-specific features username, hashtag, and retweet. The tool is somewhat tolerant of the non-standard orthography typical of Twitter messages.

3. Analysis and Discussion

The linguistic profiles of the subcorpora were determined and the relationship between gender and individual grammatical features assessed using t-tests. Principal components analysis was used to gauge the extent to which males and females 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).[7]
In Iceland, Norway and Denmark, males use the national language on Twitter more than do females; Females use English more. This difference is most pronounced for Denmark. In Sweden and Finland the rates of language use by gender are similar, with males using slightly more English and females the national languages.

3.2. Correlation of Grammatical Features, Country and Gender

34 of the PoS tags were applied at least once in all of the ten gendered subcorpora. For each subcorpus, the rela-

[3] Most Twitter users select a `place` when registering with the service; the coordinates of the `place` are then automatically assigned by Twitter as a lat-long bounding box in tweet metadata. Some users additionally opt to broadcast precise GPS coordinates with each status update. For tweets without precise geographical coordinates, location was induced by calculating the center of the bounding box circumscribing the `place` field. Correlation of the precise GPS coordinates and the induced coordinates based on centering the `place` entity was 0.993, as the `place` entity is almost always populated by a bounding box circumscribing a small area such as a city. See also Leetaru et al. (2013).

[4] For Finland, corpora were also created for the country's second official language, Swedish.

[5] Latent attribute inference using Twitter data manually tagged for gender is a popular topic in machine learning (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, but both approaches can be used to investigate links between language use and gender.

[6] The differences are due in part to the somewhat different name frequency information obtained from the national statistical offices. For example, only 395 unique given names were obtained from Iceland, but 1190 from Norway, 5382 from Denmark, 1704 from Sweden, and 7899 from Finland.

[7] The Twitter automatic language detection algorithm classifies both *Riksmål* and *Nynorsk* with the language code `no`, "Norwegian". For Finland, the percentage shown includes messages messages in the national languages of Finnish and Swedish.

		Nat. lang.	English	Other
Iceland	males	80.8	9.8	9.4
	females	71.5	17.6	10.9
Norway	males	46.6	28.9	24.5
	females	37.3	40.0	22.7
Denmark	males	45.4	40.0	14.6
	females	25.7	52.5	21.8
Sweden	males	61.9	24.5	13.6
	females	63.8	23.8	12.4
Finland	males	57.2	28.8	14.0
	females	58.5	25.0	16.5

Table 1: Percent tweets by country, gender and language.

tive frequency of each tag was calculated. To determine whether features were preferred by males or females, a t-test of population means was conducted on the basis of the mean standardized value for males and for females in all subcorpora. Of the 34 features, ten exhibited significant ($p < 0.05$) differences in use between males and females: Sentence-ending punctuation, numbers or numerals, proper nouns, and gerund or present participle forms were more frequently utilized by males, while personal pronouns, possessive pronouns, adverbs, interjections, usernames, and past particples were more likely to be used by females (Table 2).

	Feature	Gender	p-value	Signif.
1	Quotation marks (”)	m	0.320	
2	Left bracket (()	m	0.080	
3	Right bracket ())	m	0.089	
4	Comma	m	0.098	
5	Period (. *? !*)	m	0.010	*
6	Other punctuation (*: ; ... + - = <> [ ]*)	m	0.245	
7	Coordinating conjunction	f	0.269	
8	Number	m	0.040	*
9	Determiner	m	0.416	
10	Hashtag	f	0.758	
11	Preposition or subordinating conjunction	m	0.502	
12	Adjective	m	0.405	
13	Comparative adjective	f	0.848	
14	Superlative adjective	f	0.213	
15	Modal verb	f	0.695	
16	Noun, singular or mass	m	0.275	
17	Proper noun	m	0.014	*
18	Plural noun	m	0.596	
19	Personal pronoun	f	0.005	*
20	Possessive pronoun	f	0.005	*
21	Adverb	f	0.036	*
22	Phrasal particle	m	0.449	
23	*to*	f	0.596	
24	Interjection	f	0.018	*
25	Username (preceded by)	m	0.168	
26	Verb, base form	f	0.007	*
27	Verb, past tense	f	0.441	
28	Verb, gerund or present particle	f	0.866	
29	Verb, past participle	m	0.022	*
30	Verb, non-3rd person singular present	f	0.001	*
31	Verb, 3rd person singular present	f	0.292	
32	Wh-determiner	m	0.094	
33	Wh-pronoun	f	0.934	
34	Wh-adverb	f	0.106	

Table 2: Grammatical features by gender

Gendered differences were also considered by country and feature. For Sweden, for example, the distribution of those features for which a significant difference by gender was detected is depicted in Figure 3. The differences between males and females are not large (Cohen's $d \leq 0.24$) , but statistically significant according to a t-test of population means: E.g. 5.83% of all words used by Swedish females in English on Twitter are personal pronouns, compared to 4.28% by Swedish males.

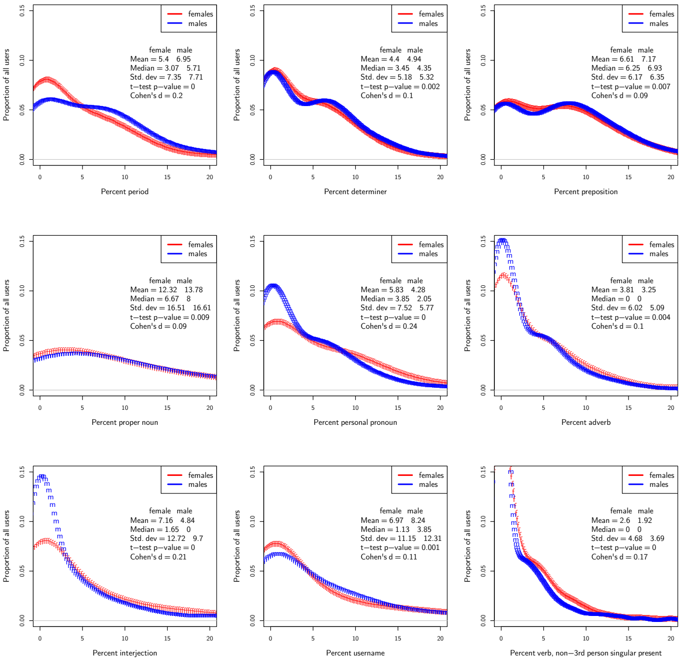


Figure 2: Percent of all tokens by feature for features that differ significantly by gender from Sweden.

3.3. Principal Components Analysis

In order to explore underlying patterning of the variance in the data, a principal components analysis was conducted on a covariance matrix of the normalized frequencies of the 34 variables for the ten English subcorpora (a male and a female subcorpus for each of the five Nordic countries). The first two components capture 70.8% of the variance in the data. The strongest loadings ($\geq |0.2|$) on the first two components are shown in Table 3.

Feature	PC1	PC2
Personal pronoun	0.60	-0.21
Interjection	0.31	0.34
Verb, non-3rd person singular present	0.21	
Period (. *? !*)	-0.28	0.28
Noun, singular or mass	-0.25	-0.51
Proper noun	-0.45	
Comma		0.38
Number		0.34
Username		0.23

Table 3: Loadings $\geq |0.2|$ on first two principal components

For the features with the strongest loadings on the first principal component, grammatical types with interpersonal interaction and stance orientation functions (personal pronouns, 1st- and 2nd-person singular present verb forms, and interjections[8]) have the strongest positive loadings, while

[8] The Carnegie-Mellon Twitter tagger also assigns the interjection tag to emoticons, word types that are often associated with the expression of emotional affect (Vandergriff 2013).

features with informational and text-organizational functions (nouns, proper nouns, and sentence-ending punctuation) have the strongest negative loadings.

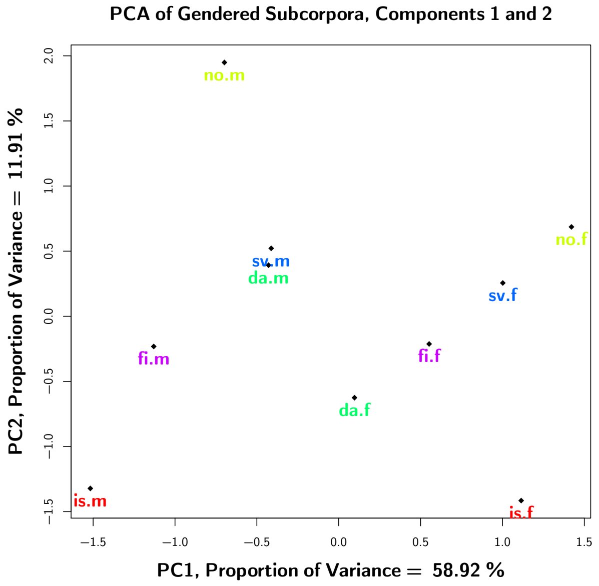


Figure 3: Loadings on components 1 and 2 of PCA for English subcorpora.

The positions of the gendered subcorpora along the first two principal components are shown in Figure 4. The analysis suggests some functional separation between males and females in Nordic Twitter Englishes as they are manifest in terms of grammatical feature frequencies: The male corpora all have negative values in the first principal component, while the female corpora have positive values. Gender separation along the second principal component is also manifest, although not as pronounced. In terms of the individual Nordic countries, the distance between males and females is larger for Iceland and Norway, while it is somewhat smaller for Denmark, Sweden and Finland.

4. Conclusion and Summary

Geographically specified and gender-induced corpora of online Englishes complied from social media sites such as Twitter 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. The results of this study bear upon research into online English varieties and the relationship between language and gender.

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 (cf. Mocanu et al. 2013). Overall, persons in Denmark and Norway send more tweets in English, and females more than males.

In the present work, gender analysis reinforces findings from previous corpus studies and research into L1 Twitter or CMC English: Females tend to use features such as personal pronouns, possessive pronouns or affect markers more than males, whereas males use features such as determiners, numbers/numerals, and nouns more than do females (Bamann, Eisenstein and Schnoebelen 2014). This patterning holds true for English used on Twitter in the Nordic countries by persons with common Nordic names, many of whom are likely non-L1 English users.

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). In this study, Nordic Twitter data that have been induced to reflect author gender exhibit differentiation by gender along a first principal component, explaining the majority of variance in the data (58.9%). 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 (1988: 107).

Most work on differences in feature frequencies by gender has been conducted on L1 English data, but there is some evidence for differential use of word classes by gender in other languages.[9] This study shows that similar differences exist for (presumable) non-L1 English users on Twitter. It has been suggested that the small differences in aggregate grammatical 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 for Nordic Twitter corpora with induced author gender, functional separation of English-language feature frequencies by gender can be observed. A tentative confirmation of some of the trends observed in CMC and Twitter data from L1 Anglophone contexts raises interesting questions as to the possible causes. Future work could further investigate this topic by exploring the extent to which gender differentiation is present in Twitter material in the Nordic languages, and whether language transfer phenomena may influence the large-scale patterning of linguistic elements in non-L1 online Englishes.

5. References

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. In *Proceedings from the XVI:th Scandinavian Conference of Linguistics*, Turku, Finland. Department of Linguistics, University of Turku.

Argamon, S., Koppel, M., Pennebaker, J., and Schler, J. (2007). Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, 12(9).

Bamann, D., Eisenstein, J., and Schnoebelen, T. (2014).

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

Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, 18(2):135–160.

Baron, N. 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, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK.

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

Blommaert, J. (2012). Supervernaculars and their dialects. *Dutch Journal of Applied Linguistics*, 1(1):1–14.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Stroudsburg, PA. Association for Computational Linguistics.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLoS ONE*, 9(1).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Stroudsburg, PA. Association for Computational Linguistics.

Görlach, M. (1995). *Still More Englishes*. John Benjamins, Amsterdam.

Gustafson-Capková, S. and Hartmann, B. (2008). *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University.

Herring, S. and Paolillo, J. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

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

Leetaru, K. H., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5/6).

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Mislove, A., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. In *Proceedings of ICWSM*, pages 554–557, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., and Vespignani, A. (2013). The Twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4).

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390. NAACL-HLT.

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

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to Twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 281–288, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, pages 37 – 44. ACM.

Roesslein, J. (2015). Tweepy. Python programming language module.

Schenk-van Witsen, R. (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.

Squires, L. (2015). Twitter: Design, discourse, and implications of public text. In Alexandra Georgakopoulou et al., editors, *The Routledge Handbook of Language and Digital Communication*, pages 239–256. Routledge, London and New York.

Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics*, 51:1–12.

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

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

Framework for an Analysis of Slovene Regional Language Variants on Twitter

Jaka Čibej
Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
E-mail: jaka.cibej@ff.uni-lj.si

Abstract

The rapid rise of computer-mediated communication has allowed regional language variation to flourish in written form, opening new doors both for dialectological studies as well as natural language processing. In this paper, we present the methodology and framework for a linguistic analysis of Slovene regional language variants on Twitter. We describe the creation and sampling of a dataset stratified by region, present a preliminary typology of non-standard Slovene language elements on Twitter, and propose an approach to measure regional specificity and dispersion of non-standard language elements in computer-mediated communication.

Keywords: regional language variants, non-standard Slovene, Twitter, computer-mediated communication

1. Introduction

In the last two decades, the rapid rise of computer-mediated communication and social media has allowed language to spread into digital communication platforms, lending a voice to a plethora of different languages traditionally present only in spoken varieties: from sociolects to dialects and everything in between. In addition, due to their ever increasing quantities, internet texts have become an important source of information, and there is an increasing demand for tools and resources to help process them, as shown by the proliferation of different areas within internet linguistics and natural language processing. One of the problematic aspects to be tackled in this regard is regional language variation.

The main goal of this paper is to present a methodology and framework for a linguistic analysis of regional language variants on Twitter. The paper is structured as follows: first, we present a brief overview of related work, which is followed by the description of our dataset and the sampling methods used. We then provide an overview of the preliminary typology of non-standard Slovene language elements on Twitter and the measures of regional specificity and dispersion to be used in further analyses, and conclude with the preliminary results of the analysis of three regional samples.

2. Related Work

Studies on regional variation of various languages on Twitter have been conducted with different purposes, mostly as part of development of NLP tools, e.g. diacritic restoration (Harrat et al., 2013) and POS-tagging (Bernhard & Ligozat, 2013), but also within sociological studies of language variation (Jørgensen et al., 2015; Eisenstein, 2015).

Slovene regional language variation on Twitter (and in social media in general), however, is currently still an under-researched area that cannot be neglected, especially considering the rich dialectal variation of Slovene (Ramovš, 1931), the numerous dialectological studies conducted on spoken Slovene (Kenda Jež, 2002), as well as the fact that regional variation has already been documented in Slovene tweets (Fišer et al., 2015b).

3. Dataset Preparation

The dataset presented in this paper consists of tweets extracted from the JANES corpus of Slovene user-generated content (Fišer et al., 2015a). The tweets were sampled by taking into account a number of criteria.

First, only tweets sent from private accounts were included, while tweets from corporate accounts (e.g. those managed by press agencies and companies) were eliminated.[1] This was done for two reasons: corporate accounts contain many automatically generated tweets, while the overwhelming majority of their original tweets are written in standard Slovene, which makes them irrelevant for our study.

Second, the dataset only includes L3 tweets, i.e. those with a high level of linguistic non-standardness (Ljubešić et al., 2015). L3 tweets contain a high degree of non-standard spelling and vocabulary and as such provide the most material for the study of regional language variants.

Third, the tweets were sampled by taking into account the metadata on the users' regional origin (Čibej & Ljubešić, 2015). This metadata was determined by collecting Slovene geotagged tweets over a period of eighth months (from January 2015 to September 2015), then assigning each user with geotagged tweets to one of 9 regions corresponding to the 7 main dialectal groups of Slovene as well as Ljubljana and Maribor, the two largest cities, which we decided to treat separately as melting pot areas. In order to exclude users with ambiguous origin, only users that sent more than 90% of their tweets from a single region were taken into account. A certain amount of noise is to be expected in the dataset despite this criterion, but should not prove too prominent and will be further penalised during the analysis (see Section 6).

[1] Twitter users included in the JANES corpus were manually annotated as corporate or private.

Regional subcorpus	Number of tokens	Number of tweets	Number of users
Gorenjska	37,683	22,070	48
Dolenjska	17,364	6,922	22
Štajerska	41,712	9,284	42
Panonska	5,020	2,512	14
Koroška	6,207	4,203	5
Primorska	13,917	5,748	31
Rovtarska	4,823	2,348	7
Ljubljana	92,104	43,018	116
Maribor	4,789	4,340	14

Table 1: Size of regional subcorpora in the JANES corpus of Internet Slovene (v0.3).

As shown in Table 1, some of the regional subcorpora are very small both in terms of the number of tokens as well as the number of users included. However, geotagged tweets are still being collected, and more users and tweets will be added when the corpus is updated. For the purposes of this paper, we focus on three of the best represented regions: Primorska, Gorenjska, and Štajerska.

3.1. Samples of Regional Subcorpora

For each region, a sample containing 500 L3 tweets was created. First, all tweets were extracted from the relevant regional subcorpus. The tweets were then shuffled and sampled by user in order to avoid overrepresentation of very prolific Twitter users. In some cases, the most active users provided more than 2,000 tweets to a regional subcorpus, while the least active provided less than 10.

The samples included all users from the relevant regional subcorpus, while the number of tweets each user contributed was limited to a maximum of 40–50 tweets (depending on the total number of users).

4. Typology of Non-Standard Slovene Language Elements on Twitter

Small subsets of 100–150 tweets were manually analysed in each sample in order to design a typology of non-standard Slovene language elements on Twitter. The typology was created with a bottom-up approach and so far includes 7 main categories: non-standard vocabulary, reductions and ellipses, non-standard morphology, spelling variants of frequent standard words, alternative graphemes, frequent transformations, and miscellaneous.[2] Currently, the typology consists of 105 different tags, but is flexible and allows for the addition of new elements as certain rare or regionally specific elements (especially those concerning morphology and syntax) may yet arise during annotation. In the following subsections, we present the main categories in further detail.

4.1. Non-Standard Vocabulary

Non-standard vocabulary includes all lexical elements that are considered non-standard, i.e. those that would not be expected in standard Slovene texts and/or are not included in existing standard language resources such as dictionaries or lexicons. Examples include regionally specific words (e.g. particles *ejga* for Gorenjska, *čuj* for Štajerska, *nanka* for Primorska), standard words with new meanings (*hudo* meaning 'awesome' instead of 'bad'), and non-standard words/phrases of foreign language origin, either in their original spelling (e.g. *web app*) or fully/partially adapted to Slovene spelling and morphology (e.g. *ekskjuz*, from English 'excuse'; *učelini*, from Italian 'uccellini').

A subcategory of non-standard vocabulary also included certain CMC-specific abbreviations, either English (*wtf*, *lol*, *omg*) or Slovene (*jbg* „fuck that‟, *bmk* „I don‟t give a fuck‟), and alphanumerical spellings (*ju3* for *jutri*, 'tomorrow').

4.2. Reductions and Ellipses

With 69 different tags, reductions and ellipses are by far the most prolific category. Most often, they involve vowel drops in different positions in a word. A common example is the ellipsis of the final *-i* in the infinitive (*delati* → *delat*, 'to work') or the final *-o* in adverbs (*čudno* → *čudn*, 'weirdly, oddly'). As for consonants, a common example is the ellipsis of *-j* in the *-lj-* or *-nj-* consonant clusters (*peljem* → *pelem*, 'I drive'; *zadnji* → *zadni*, 'the last').

4.3. Alternative Graphemes

This category encompasses alternative, non-standard spellings of graphemes, most often in cases when it is pronounced differently in spoken language. Examples include the spelling of *g* as *h* (*bog* → *boh*, 'god') or *v* as *w* (*ne vem* → *ne wem*, 'I don‟t know').

4.4. Non-Standard Morphology

This category included words that exhibited non-standard morphological characteristics such as alternative case endings (e.g. the non-standard locative ending *-i* of singular masculine nouns, *na šihtu* → *na šihti*, 'at work') or other regionally specific suffixes (e.g. the non-standard second-person plural verb suffix *-ste* instead of *-te*, *imate* → *imaste* „you have‟).

4.5. Spelling Variants of Frequent Standard Words

The category of spelling variants includes common standard (mostly function) words with numerous spelling variants that are unequally distributed between different regions. A good example is the word *toliko* ('this much, so'), which can also be spelt as *tok*, *tolk*, *tolko*, *telko*, *tuk*, *tulk*, etc. Similarly, the word *jaz* (personal pronoun, first person singular, 'I') can also be encountered as *jz*, *js*, *jst*, *jest*, *jes*, etc. Although these spelling variants often also include other non-standard elements (e.g. vowel ellipses), they are also annotated as a separate category in order to produce an exhaustive list so that their regional distribution can be tested on the entire geolocated JANES subcorpus.

[2] Initially, a syntactic category was included, but was later omitted as syntactic elements were much too scarce in the samples. However, potential regionally specific syntactic features encountered during the analysis will be researched on larger amounts of data in the JANES corpus of Internet Slovene.

4.6. Frequent Transformations

This is an additional category that included spelling transformations in non-standard word spellings that were perceived during annotation as frequently occurring. Similar to frequent spellings of non-standard words, these transformations were annotated separately in order to allow for a comparison of their distributions in different regions. A prevalent example is the transformation of *-aj-* to *-ej-* (*nekaj* → *nekej*, 'something'; *včeraj* → *včerej*, 'yesterday') or *-aj-* to *-j-* (*zdajle* → *zdjle*, 'now'; *kaj* → *kj*, 'what').

4.7. Miscellaneous

The final category included miscellaneous non-standard language elements that could not be categorised in any of the previous categories. These mainly consisted of joint spellings, i.e. instances where two words should be spelt separately in standard Slovene, but are written together in their non-standard form (e.g. *ne vem* → *nevem*; 'I don't know, I dunno'), or amalgams of two adjacent words, most often function words (*to je* → *toj*, 'this is', *če je* → *čej*, 'if it is').

5. Dataset Annotation

The samples were manually annotated in .txt format to enable flexible post-processing and analysis with Python regular expressions. Relevant tokens or phrases were annotated as shown in Figure 1.

[token/phrase]{tag 1}{tag 2}{...}

[sej]{V.saj}{Taj.ej}

Figure 1: Annotations.

The upper line shows the general pattern of annotation. A single token or phrase may be annotated with multiple tags. The bottom line shows an example of the annotated word *sej* ('because'), annotated both as a spelling variant of *saj* (*V.saj*) and as a frequent spelling transformation of *-aj-* to *-ej-*. (*Taj.ej*).

Several language elements were excluded from annotation. These included a number of CMC-specific elements (emoticons and emojis, hashtags or URLs), spelling mistakes that were perceived as obviously accidental, as well as the non-use of diacritics, which is often a consequence of technical limitations and rarely voluntary. Code-switching, although relatively common, was also omitted. If foreign language words or phrases were used as part of a Slovene sentence, they were annotated as non-standard vocabulary. Entire sentences or independent units in foreign language, however, were disregarded. The same was true of non-standard variants of proper nouns (e.g. phoneticised versions of Twitter and Facebook – *Tviter*, *Fejsbuk*).

6. Measures of Regional Specificity and Dispersion

In addition to statistical tests, we also propose a method to determine the level of regional specificity of a certain language element based on a number of criteria described in the following subsections. In addition, these measures should help to reduce the effect of potential noise in the dataset (e.g. users that are originally from a different region and have permanently moved to a different one, but continue to use language elements typical of their region of origin).

6.1. Relative Frequency

Relative frequency (f_R) is the ratio of the frequency of a language element and the total number of occurrences in its category. The greater the relative frequency, the more frequent the language element within the region.

6.2. User Ratio

The user ratio (u) is the ratio of the number of users using a language element and the number of all users from the region in question. The greater the number of users that use a language element, the greater the user ratio. This value thus measures how widespread the element is among the users of the region. It penalises idiosyncratic elements (especially with prolific users) or elements used by users that have been misclassified as pertaining to a specific region.

6.3. Type/Token Ratio

The type/token ratio (t) is the ratio of the number of types and the number of tokens used with a language element. The greater the t-ratio, the greater the number of words it occurs with, and the greater the likelihood that the element will arise in text. This value penalises frequent language elements that only occur in a limited number of words.

6.4. Annotation Ratio

Similar to the type/token ratio, the annotation ratio (a) is the ratio of the number of different tags the element occurs with and the number of all tags in its category. The greater the annotation ratio, the greater the number of tags it occurs with and the greater the likelihood it occurs.

6.5. Coefficient of Regional Dispersion

The coefficient of regional dispersion (δ_R) is meant as a simple summarisation of all other measures of regional specificity and dispersion. It is calculated as follows:

$$\delta_R = f_R \times u \times t \times a \times 100$$

The greater the coefficient of regional dispersion, the more widespread and frequent the element in question.

7. Annotation Results

In this section, we provide some of the preliminary results of the annotated dataset and demonstrate the use of the abovementioned f_R, u, t, a and δ_R values to measure regional specificity and dispersion for a particular language element.

The annotation results for the Gorenjska, Štajerska and Primorska regions are shown in Table 2. As all samples are of comparable size (they consist of 500 tweets each), the absolute frequencies are given.

Category	Primorska	Gorenjska	Štajerska
Non-standard vocabulary	394	347	371
Spelling variants of frequent standard words	233	322	183
Alternative graphemes	40	54	34
Reductions and ellipses	588	1122	648
Non-standard morphology	90	99	67
Frequent transformations	120	181	68
Miscellaneous	39	59	24
Total	1504	2184	1395

Table 2: Quantitative Analysis of Annotated Samples.

The regions do not differ to a great extent in terms of the frequency of non-standard vocabulary, although we expect that a detailed qualitative analysis will show differences in the type of non-standard words used (e.g. we expect to find more words originating from Italian in the Primorska region, which lies next to the border with Italy).

As far as the frequencies of other categories are concerned, the differences between the three regions are more pronounced. What is particularly interesting to note is that while reductions and ellipses are the most prolific category in all three regions, they are especially frequent in the Gorenjska region. The most frequent type of ellipsis in all three regions was the *-i* ellipsis. The frequencies of final and non-final *-i* ellipses are shown in Table 3, along with χ^2 p-values and Cramer's V effect sizes.

	Gorenjska vs. Primorska		Gorenjska vs. Štajerska		Primorska vs. Štajerska	
Final *-i* ellipsis	254	140	254	174	140	174
Non-final *-i* ellipsis	231	156	231	118	156	118
χ^2 p-value	>0.05		>0.05		0.037	
Cramer's V	0.05		0.07		0.12	

Table 3: χ^2 p-values and Cramer's V effect sizes for distributions of final vs. non-final *-i* ellipses.

The only statistically significant difference in the distribution of final vs. non-final *-i* ellipses is the one between Primorska and Štajerska, with a small, but not entirely negligible effect size. It would appear Štajerska slightly prefers final *-i* ellipsis to non-final *-i* ellipsis.

Table 4 shows the measures of regional specificity and dispersion for the final *-i* ellipsis for all three regions.

	Gorenjska	Štajerska	Primorska
f_R	0.52	0.60	0.47
u	0.75	0.42	0.54
t	0.61	0.53	0.67
a	0.43	0.41	0.24
δ_R	10.25	5.41	4.08

Table 4: Measures of regional specificity and dispersion for the final *-i* ellipsis.

As can be deduced from Table 4, the final *-i* ellipsis has a significantly greater user ratio in Gorenjska, as well as a significantly higher coefficient of regional dispersion, which would indicate that the language element is much more widespread in this region compared to Štajerska and Primorska.

8. Conclusion

In the paper, we described the creation of a dataset for the analysis of Slovene regional language variants on Twitter and presented a method for the analysis of regional language variants on Twitter.

In our future work, we will perfect the typology of non-standard language elements in Slovene CMC and make a comparison with phenomena presented in existing Slovene dialectological studies. We will also extend the annotated dataset to other Slovene regions and analyse all encountered language elements in terms of their regional specificity and dispersion, then compare the results with the results obtained through other statistical methods. In addition, elements that rarely occur in the samples (e.g. non-standard syntactic constructions) will be tested on larger text samples in the JANES corpus of Internet Slovene. The results of the analysis will be used to design features to be used in the development of a model for the automatic recognition of Slovene regional language variants on Twitter.

9. Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Non-standard Internet Slovene” (J6-6842, 2014–2017).

10. References

Fišer, D. Ljubešić, N. & Erjavec, T. (2015a). The JANES corpus of Slovene user generated content: construction and annotation. *International Research Days: Social Media and CMC Corpora for the eHumanities: Book of Abstracts, 23–24 October 2015*. Rennes, France, p. 11.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. & Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*. Hissar, Bulgaria, pp. 371–378.

Jørgensen, A. K., Hovy, D. & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Beijing, China, July 31, 2015, pp. 9–18.

Harrat, S., Abbas, M., Meftouh, K. & Smaili, K. (2013). Diacritics restoration for Arabic dialect texts. *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*. France.

Bernhard, D. & Ligozat, A.-L. (2013). Hassle-free POS-Tagging for the Alsatian Dialects. Zampieri, M. & Diwersy, S. (eds.), *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag, pp. 85–92.

Čibej, J. & Ljubešić, N. (2015). "S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. Fišer, D. (ed.), *Proceedings of Konferenca Slovenščina na spletu in v novih medijih*. Ljubljana, ZIFF, pp. 10–14.

Kenda Jež, K. (2002). *Cerkljansko narečje: teroetični model dialektološkega raziskovanja na zgledu besedišča in glasoslovja*. PhD dissertation. Ljubljana: Faculty of Arts.

Eisenstein, J. (2015). Written dialect variation in online social media. Boberg, C., Nerbonne, J. & Watt, D. (eds.): *Handbook of Dialectology*. Wiley.

Ramovš, F. (1931). Dialektološka karta slovenskega jezika. Ljubljana: Rektorat univerze kralja Aleksandra I. in J. Blaznika nasl. – Univerzitetna tiskarna.

Fišer, D., Erjavec, T., Čibej, J. & Ljubešić, N. (2015). Gradnja in analiza korpusa spletne slovenščine JANES. Smolej, M. (ed.): OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis. Ljubljana: ZIFF, pp. 217–223.

Analysis of Sentiment Labeling of Slovene User-Generated Content

Darja Fišer,[*†] **Tomaž Erjavec**[†]
[*] Department of Translation, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana
[†] Department of Knowledge Technologies, Jožef Stefan Institutute, Jamova cesta 39, 1000 Ljubljana
E-mail: darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si

Abstract

The paper takes a close look at the results of sentiment annotation of the Janes corpus of Slovene user-generated content on 557 texts sampled from 5 text genres. A comparison of disagreements among three human annotators is examined at the genre as well as text level. Next, we compare the automatically and manually assigned labels according to the text genre. The effect of text genre on correct sentiment assignment is further investigated by investigating the texts with no inter-annotator agreement. We then look into the disagreements for the texts with full human inter-annotator agreement but different automatic classification. Finally, we examine the texts that humans and the automatic model struggled with the most.

Keywords: sentiment analysis, quantitative and qualitative evaluation, user-generated content, non-standard Slovene

1. Introduction

Sentiment analysis or opinion mining detects opinions, sentiments and emotions about different entities expressed in texts (Liu, 2015). It is currently a very popular text-mining task, especially for social networking services, where people regularly express their emotions about various topics (Dodds et al., 2015). A sentiment analysis system for Slovene user generated content (UGC) was developed by Mozetič et al. (2016) and has been, *inter alia*, used to annotate the Janes corpus of Slovene UGC (Erjavec et al., 2015). The first results are encouraging but the results vary both in inter-annotator agreement and accuracy of the system across genres (Fišer et al., 2016), suggesting further improvements of the system are needed. One of the steps towards this goal is a qualitative analysis of (dis)agreement among the annotators and an error analysis of the incorrectly classified texts, which is the goal of this paper.

The paper is organized as follows. In Section 2 we give a brief presentation of the corpus and its sentiment annotation. In Section 3 we present the results of a quantitative analysis of manual and automatic sentiment annotation on a sample collection of texts. In Section 4 we follow with a qualitative analysis of the texts and their features that make the task difficult for humans as well as those that the algorithm struggles with. The paper ends with concluding remarks and ideas for future work.

2. Sentiment Annotation of Janes

The Janes corpus (Erjavec et al., 2015) is the first large (215 million tokens) corpus of Slovene UGC that comprises blog posts and comments, forum posts, news comments, tweets and Wikipedia talk and user pages. Apart from the standard corpus processing steps, such as tokenization, sentence segmentation, tagging and lemmatization (Ljubešić and Erjavec, 2016) as well as some UGC-specific processing steps, such as rediacritization (Ljubešić et al., 2016), normalization (Ljubešić et al., 2014) and text standardness labeling (Ljubešić et al., 2015), all the texts in the corpus were also annotated for sentiment (negative, positive, or neutral) with a SVM-based algorithm that was trained on a large collection of manually annotated Slovene tweets (Mozetič et al., 2016).

We also produced a manually annotated dataset. This evaluation dataset comprised 600 texts, which were sampled in equal proportions from each subcorpus (apart from blog comments as they have been found to behave very similar to news comments) in order to represent all the text genres included in the corpus in a balanced manner.

The sample was then manually annotated for the three sentiment labels by three human annotators. The annotators marked some texts as out of scope (written a foreign language, automatically generated etc.), so the final evaluation sample consists of 557 texts.

In the following sections the labels assigned by the annotators were compared to each other while the automatically assigned scores were compared to the annotators' majority class, i.e. the sentiment label assigned to each text by the most annotators. In cases of complete disagreement the neutral sentiment is assigned as the majority class.

3. Quantitative Analysis of Sentiment Annotation

In our quantitative analysis we first analyze the difficulty of the task for humans and the algorithm on the evaluation sample. We also compare annotation results with respect to text genres. Finally, we measure the degree of disagreements of the assigned labels in order to measure the severity of the annotation incongruences.

3.1. Comparison Between Manual and Automatic Annotations

First, a comparison of disagreements among the human annotators was computed as well as that of the automatic system with the majority class. Since we are investigating sentiment annotation accuracy from the perspective of the difficulty of the task, measured with the dispersion of annotations by human annotators, we are operating with percentage agreement in this paper. While we have measured inter-annotator agreement with Krippendorff's alpha, which is 0.563 for human annotations and 0.432 for automatic annotations with respect to the human majority vote (cf. Fišer et al., 2016), this measure reports inter-annotator agreement for the entire annotation task and is as such not informative enough for the task at hand in this paper.

The results in Table 1 shows that the task was easier for some texts in the sample both for humans and for the system as annotators' labels range from perfect agreement to an empty intersection. While at least two human annotators provided the same answer on nearly 97% of the sample, all three annotators agreed on less than half of the texts, which is a clear indication that the task is not straightforward and intuitive for humans, suggesting that better guidelines and/or training are needed to obtain consistent and reliable results in the annotation campaign. As could be expected, texts that were difficult to annotate for humans also proved hard for the system. Namely, the system chose the same label as the annotators in the majority of the cases (65 %) only for those that humans were in complete agreement. Where the annotators disagreed partially or completely, there is substantially less overlap with them and the system (46% - 33%).

Manual / *Automatic*	*All annotators agree*		*2/3 annotators agree*		*All annotators disagree*	
identical	160	**65%**	133	46%	6	33%
different	87	35%	159	**54%**	12	**67%**
total	247	44%	292	52%	18	3%

Table 1: Comparison between automatic (to majority class) and manual annotations.

3.2. Comparison Between Text Genres

In order to better understand which text types are easy and which difficult for sentiment annotation, we compared the labels assigned by the annotators and the system according to the genre of the texts in the sample. In texts for which annotators are in perfect agreement, the biggest overlap between the system and the majority vote of the annotators is achieved on news comments. These are followed by blog posts which, together with the news comments, represent over half of all the texts receiving the same sentiment label by both humans and the model.

The effect of text genre on the difficulty of correct sentiment assignment was further investigated by looking at the genre of those texts for which there was no agreement among the human annotators, i.e. texts which were annotated as negative by one annotator, positive by another and neutral by the third. The results of this analysis are presented in Table 3 and are consistent with the previous findings in that sentiment in forum posts and tweets is the most elusive while being the least problematic on Wikipedia talk pages and in news comments.

	All annotators agree				*2/3 annotators agree*				*All annotators disagree*			
	Different		*Identical*		*Different*		*Identical*		*Different*		*Identical*	
Type	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
blog	14	16	34	21	38	24	27	20	3	25	1	17
forum	23	**26**	29	18	42	**26**	20	15	4	**33**	1	17
news	12	14	48	**30**	21	13	32	**24**	2	17	0	0
tweet	23	**26**	24	15	25	16	21	16	2	17	4	**67**
wiki	15	17	25	15	33	21	33	24	1	8	0	0
total	87		160		159		133		12		6	18

Table 2: Comparison between automatic and majority vote per text genre.

Since the system was trained on tweets, one would expect them to receive the highest agreement, which is not the case. A possible reason for this is that sentiment is more explicitly expressed in news comments than in tweets, whereas blogs might be easier because they are longer which again makes them easier for sentiment identification. Forum posts, on the other hand, seem to be the hardest overall, which is addressed in more detail in Section 4.

Text type	*Disagreement*	
blog	4	21%
forum	6	**32%**
news	2	11%
tweet	6	**32%**
wiki	1	5%
total	19	100%

Table 3: Disagreement among the annotators per text genre.

3.3. Comparison of the Degree of Disagreements

Since not all incongruences between the system and the true answer are equally bad from the application point of view, we looked into the degrees of disagreements for the texts receiving the same label by all three annotators and a different one by the system. As can be seen from Table 4, the automatic system has a clear bias towards neutral labels, i.e. more than half of the mislabeled opinionated texts were marked as neutral by the algorithm. Mislabeling neutral texts as opinionated is seen in about a third of the cases. The worst-case scenario, in which negative texts are labeled as positive or vice versa and therefore hurts the usability of the application the most, is quite rare (12%). The behavior of the system on texts with partial human agreement is consistent with the findings above in assigning sentiment of opposite polarities which again represents the smallest part of the sample (8%). Neutralizing negative and positive texts occurs on 40% of the sample, which is slightly lower than for the texts on which all the annotators agree. The most prevalent category are neutral texts mislabeled as negative which is seen in 34% of the cases, substantially more than above.

Differences	*Annotators agree, system disagrees*		*2/3 annotators agree, system disagrees*	
neg → neut	29	**33%**	36	23%
neg → pos	7	*8%*	7	*4%*
neut → neg	14	16%	54	**34%**
neut → pos	13	15%	29	18%
pos → neut	20	**23%**	6	17%
pos → neg	4	*5%*	27	*4%*
Total	87	100%	159	100%

Table 4: Discrepancies between automatic and majority human vote.

4. Qualitative Analysis of Sentiment Annotation

In this section we present the results of qualitative analysis of the biggest problems in sentiment annotation observed in the evaluation sample. We first examine all the texts for which there was no agreement among the human annotators and then focus on the texts that humans found easy to annotate consistently but the system failed to annotate correctly.

4.1. Toughest Sentiment Annotation Problems for Humans

By examining the texts which received a different label by each annotator we wished to investigate the difficulty of the task itself, regardless of the implementation of an automatic approach. In the evaluation sample of 557 texts there were 18 such cases: 6 tweets, 5 forum posts, 4 blog posts, 2 news comments, and 1 Wikipedia talk page.

As can be seen from Table 5, there are significant discrepancies in annotator behavior. While Annotators 1 and 2 chose positive and negative labels equally frequently (A1: 9 negative, 8 positive, 1 neutral; A2: 8 negative, 8 positive, 2 neutral). Annotator 3 was heavily biased towards the neutral class (A3: 1 negative, 2 positive, 15 neutral). The automatic system lies in between these two behaviors (S: 5 negative, 7 positive, 6 neutral), sharing the most equal votes on individual texts with Annotator 1 (44%) and the fewest with Annotator 2 (22%). This suggests that annotators did not pick different labels for individual texts due to random/particular mistakes but probably adopted different strategies in selecting the labels systematically throughout the assignment. While Annotators 1 and 2 favored the expressive labels even for the less straightforward examples, Annotator 3 opted for a neutral one in case of doubt. These discrepancies could be overcome by more precise annotation guidelines for such cases.

Source	*Ann1*	*Ann2*	*Ann3*	*System*	*Note*
blog	-	+	0	-	mixed
blog	-	+	0	-	mixed
blog	-	+	0	-	mixed
blog	+	-	0	0	mixed
forum	-	+	0	+	mixed
forum	+	-	0	-	context
forum	+	-	0	0	context
forum	+	-	0	+	context
forum	+	0	-	+	context
news	-	+	0	+	context
news	+	-	0	-	mixed
tweet	-	+	0	0	sarcasm
tweet	-	+	0	0	mixed
tweet	-	+	0	0	mixed
tweet	0	-	+	0	short
tweet	+	-	0	+	mixed
tweet	+	-	0	+	sarcasm
wikip.	-	0	+	+	mixed

Table 5: Analysis of the difficult cases for the human annotators.

A detailed investigation of the 18 problematic texts showed that 3 out of 6 tweets contain mixed sentiment in the form of message and vocabulary distinctive for one sentiment, which is then followed by an emoticon of a distinctively opposite sentiment. 2 tweets were sarcastic and 1 simply too short and informal to understand what the obviously opinionated message was about (“*prrrr za bič :P / prrrr for the whip :P*”).

4 out of 5 forum posts are lacking a wider context (the entire conversation thread) which is needed in order to find out whether the post was meant as a joke or was sarcastic. Some annotators annotated it as is, others assumed sarcasm or opted for a neutral label. 1 forum post contained mixed sentiment.

All 4 blog posts were relatively long and contained mixed sentiment. For example, a post that contains a description of a blogger’s entire life starts off with very positive sentiment that then turns into a distinctly negative one after some difficult life situations. While some annotators treated this text as neutral as it contained all types of sentiment, others treated it as negative since negative sentiment is the dominant one in terms of amount of text it appears in with respect to other parts, in terms of strength with which it is expressed, and/or in terms of the final position in the text, suggesting it to be the prevailing sentiment the author wished to express.

1 news comment was lacking context and 1 contained mixed sentiment, which is also true with the Wikipedia talk page that is complaining about a plagiarized article but in a clearly constructive, instructive tone that is trying not to complain about the bad practice but teach a new user about the standards and good practices respected by the community.

4.2. Toughest Sentiment Annotation Problems for Computers

In the second part of the qualitative analysis we focus on the 87 texts from the sample which were labeled the same by all three annotators but differently by the system. With this we hope to see the limitations of the system when trying to deal with the cases most straightforward for humans. The sample consisted of 23 forum posts and 23 tweets, 15 Wikipedia talk pages, 14 blog posts and 12 news comments. As said in Section 3, almost all of the discrepancies (87%) were neutral texts that were mislabeled as opinionated by the system or vice versa. Serious errors, i.e. cross-spectrum discrepancies were rare (4.6% true negatives mislabeled as positive and 8% true positives mislabeled as negative).

Problematic feature	*No.*	*%*
no feature identified	22	25.29
neg. vocabulary	18	20.69
+ vocaulary	10	11.49
cynical	10	11.49
emoticons	7	8.05
too short	5	5.75
quote	5	5.75
foreign/specialized vocabulary	5	5.75
non-standard text	2	2.3
names	2	2.3
mixed sentiment	1	1.15
Total	87	100.00

Table 6: Analysis of the problematic text features for the sentiment annotation algorithm.

We performed a manual inspection of the erroneously annotated texts and classified them into one of 10 the categories representing possible causes for the error. As Table 6 shows, in over a quarter of the analyzed texts, no special feature was identified and it really is not clear why the system made an error there as the sentiment in them is obvious. The most common characteristics of the mislabeled texts, which occurred in 43% of the analyzed sample, were lexical features, i.e. the vocabulary typical of negative/positive messages, foreign and specialized vocabulary, proper names and non-standard words that are most likely out-of-vocabulary for the model and therefore cannot contribute to successful sentiment assignment. E.g. a perfectly neutral discussion on Wikipedia was labeled as negative due to the topic of the conversation (*quisling, invader, traitor*). Similarly, many posts with objective advice to patients on the medical forum which contain a lot of medical jargon were mislabeled as negative.

The second common source of errors were the inter- and hyper-textual features that are typical of user-generated content, such as quotes from other sources, parts of discussion threads, fragmentary, truncated messages, URL links and emoticon and emoji symbols. The remaining issues include cynical texts and texts with mixed sentiment that have already been discussed in Section 4.1.

5. Conclusions

In this paper we presented the results of a quantitative and qualitative analysis of sentiment annotation of the Janes corpus. These insights should enable better understanding of the task of sentiment annotation in general as well as facilitate improvements of the system in the future. The results of the first analysis show that overall, blogs have proven to be the easiest to assign a sentiment to as both humans and the automatic assignment achieve the highest score here. The sentiment of the blog posts we examined was straightforward to pin down by the annotators due to text length and informativeness, through which it becomes clear which sentiment is expressed by the author.

For humans, the second easiest are tweets, whereas the automatic system preforms worse on them than on news comments and Wikipedia talk pages. This is especially interesting as the automatic system was trained on tweets and would therefore be expected to perform best on the same type of texts. A detailed examination of the problematic tweets shows they are extremely short, written in highly telegraphic style or even truncated and therefore do not provide enough context to reliably determine the sentiment. Furthermore, messages on Twitter are notoriously covertly opinionated, often sarcastic, ironic or cynical, making it difficult to pin down the intended sentiment.

The results of the second analysis are consistent with the first in that texts which contain vocabulary that is typically associated with a particular sentiment but used in a different context or communicative purpose makes the sentiment difficult to determine. As for the forum posts which are much harder for the system to deal with than for humans, highly specialized vocabulary on the medical, science and automotive forums (which in addition to terminology is full of very non-standard orthography and vocabulary) would most likely be beneficial in the training data for the model to learn on. Based on the analysis reported on in this paper, we plan to improve inter-annotator agreement by providing the annotators with more comprehensive guidelines that will inform the annotators about how to treat the typical problematic cases. We will try to improve the automatic system by providing it with training material from the worst performing text types. It is less clear how to improve the quality of the automatic labeling of sarcastic, ironic and cynical tweets that are a very common phenomenon.

6. Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842. 2014-2017).

7. Bibliography

Sheridan Dodds, P., Clark, E. C., Desu, S., Frank, M. R., Reagan, A. J., Ryland Williams, J., Mitchell, L., Decker Harris, K., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F. and Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proc. of the National Academy of Sciences*. 112(8): 2389–2394.

Erjavec, T., Fišer, D., and Ljubešić, Nikola (2015). Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. *Zbornik konference Slovenščina na spletu in v novih medijih*. 20–26. Ljubljana, Znanstvena založba Filozofske fakultete.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I., and Grčar, M. (2016). Sentiment Annotation of Slovene User-Generated Content. Proc. of the *Conference Language Technologies and Digital Humanities*. Ljubljana, Faculty of Arts.

Kilgarriff, A. (2012). Getting to Know Your Corpus. *Proc. of 15th International Conference on Text. Speech and Dialogue (TSD'12)*. Brno, Czech Republic. September 3-7 2012, 3–15, Springer Berlin Heidelberg.

Liu, B. (2015). *Sentiment Analysis: Mining Opinion,. Sentiments, and Emotions.* Cambridge University Press.

Ljubešić, N., Erjavec, T., and Fišer. D. (2014). Standardizing tweets with character-level machine translation. *Computational Linguistics and Intelligent Text Processing*. LNCS 8404, 164–175, Springer.

Ljubešić, N., Erjavec, T., (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proc. of 10th International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association (ELRA).

Ljubešić, N., Erjavec, T., and Fišer, D. (2016). Corpus-Based Diacritic Restoration for South Slavic Languages. *Proc. of 10th International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association (ELRA).

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. *Proc. of 10th International Conference on Recent Advances in Natural Language Processing Conference (RANLP'15). 7–9 September 2015*, 371–378. Hissar. Bulgaria.

Martineau, J., and Finin, T., (2009). Delta TFIDF: An improved feature space for sentiment analysis. *Proc. of 3rd AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 258–261.

Mozetič, I., Grčar, M., and Smailović, J., (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE*. 11(5):e0155036.

Vapnik, V. N., (1995). *The Nature of Statistical Learning Theory*. Springer.

Compilation and Annotation of the Discourse-structured Blog Corpus for German

Holger Grumt Suárez, Natali Karlova-Bourbonus, Henning Lobin
Department for German Linguistics and Literature
Applied and Computational Linguistics
Justus-Liebig-University Giessen, Germany
Holger.H.Grumt-Suarez@germanistik.uni-giessen.de, natali.karlova-bourbonus@zmi.uni-giessen.de,
henning.lobin@germanistik.uni-giessen.de

Abstract

The present paper reports the first results of the compilation and annotation of a blog corpus for German. The main aim of the project is the representation of the blog discourse structure and relations between its elements (blog posts, comments) and participants (bloggers, commentators). The data included in the corpus were manually collected from the scientific blog portal SciLogs. The feature catalogue for the corpus annotation includes three types of information which is directly or indirectly provided in the blog or can be construed by means of statistical analysis or computational tools. At this point, only directly available information (e.g., title of the blog post, name of the blogger etc.) has been annotated. We believe, our blog corpus can be of interest for the general study of blog structure or related research questions as well as for the development of NLP methods and techniques (e.g. for authorship detection).

Keywords: CMC, blog corpus, corpus compilation, corpus annotation, TEI

1. Introduction

In our opinion, two views on computer-mediated communication (CMC) – linguistic and structural – have so far been established. According to the linguistic view, the language of CMC represents a distinct type of language form besides written and spoken language. Moreover, it combines characteristics of these two traditional language forms thus constituting a bridge between them. The structural view in its turn concentrates on building up of CMC. Two different kinds of CMC structure can be distinguished – external and internal. External structure relates to the representation, or layout, of CMC by means of HTML mark-up language which may be an individual decision of a developer. External structure most of the blogs includes for example a header (title), content, a footer (contact information) and a sidebar (site navigation). Internal structure in its turn relates to the generic structure of the CMC content. It describes a set of structural elements (e.g., post, comment, thread, word cloud etc.), properties and principles a CMC is constructed of and built on to function as a holistic construct and to match its purpose.

The identification of the full spectrum of CMC characteristics – linguistic or structural – still faces some major challenges primarily as a result of lacking valid annotated data. Storrer (2014: 189) claims that for this purpose a special – third - kind of corpus besides the written and spoken corpora is needed. She also adds that appropriate standards, methods and quality criteria for the study of CMC are crucially important as well.

In the present study, the structural nature of the weblog (henceforth blog) as a representative genre of CMC is of interest. We describe the genre blog as a dynamic, "living" construct of interrelated and interacting elements. The dynamics of a blog arise from its constant expansion as a result of ever more comments and blog posts as well as on the account of new blog participants. Additionally, the author of the blog (henceforth blogger) can edit his post any time and add new information on request. The interrelatedness and interaction between elements (blog post, comments) and agents (blogger, commentators) of the blog contribute to the dynamics of the blog as well.

To demonstrate this idea, we compiled the first version of an annotated blog corpus in German using the scientific blog portal SciLogs (SciLogs, 2016) as a data source. The corpus includes both blog posts and related comments. The catalogue of features for the annotation of the corpus is based on three types of information directly or indirectly available from the data source. The typology of information is proposed in Section 3.2.1.

The structure of the paper is as follows. Section 2 provides an overview of the studies related to the topic of the present project. Section 3 describes the main steps conducted for the purpose of the blog corpus compilation and annotation. Some observed challenges for the automation of the task and possible solutions are also included in this section. Finally, Section 4 reports the results of the project and outlines the next steps.

2. Related Work

Currently, there is a limited number of publicly-available, large-scale blog corpora. This is surprising given the great influence of blogs on the web in general.

An example for one of the few large-scale blog corpora is the Birmingham Blog Corpus compiled at Birmingham City University. The corpus consists of more than 630 million words, including a 180 million words sub-section separated into posts and comments (Kehoe, 2012). One objective of this corpus was to analyze if "comments could be used to improve document indexing on the web" (ibid.). The online tool (WebCorp, 2016) of the Birmingham blog corpus allows the querying of words and phrases, but there is no possibility to either search the comment structure, a specific time period, keywords or a specific blogger respectively commentator.

Another example of a blog corpus is the bilingual (German, French) corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne), which is part of the LETEC (Learning and Teaching Corpus) (Abendroth-Timmer, 2014). This corpus is included in the CoMeRe (Communication médiée par les réseaux) project, which "aims to build a Kernel corpus assembling existing corpora of different CMC […] genres and new corpora build on data extracted from the Internet" (Abendroth-Timmer, 2014). The INFRAL blog corpus consists of posts from two groups: a group of ten francophone learners of German as a foreign language from l'Université de Franche-Comté and a group of nine German-speaking learners of French as a foreign language from the University of Bremen who e.g. had to discuss various intercultural topics. One task of this corpus was the modeling of the structure of interactions. Therefore, every comment has been given a reference to the ID of the post, but the links between the comments themselves are not included. The TEI schema developed for the CoMeRe project – this project is also part of the TEI special interest group (SIG) "computer-mediated communication" (CMC) – will be an important basis for our own schema.

Finally, the German language wordpress blog corpus by Barbaresi and Würzner (Barbaresi & Würzner, 2014) is another example of a blog corpus worth mentioning. The corpus consists of a total of 158,719 German wordpress blogs. The collected data is released under the Creative Commons license. The corpus can be used for example in the lexicography for the purpose of dictionary building. Moreover, it can be a good source "to test linguistic annotation chains for robustness" (Barbaresi & Würzner, 2014).

3. Methodology

3.1 Data Collection

To date, we have compiled a test corpus in the German language, which contains 21 blog posts and 195 comments related to those blog posts. For this test corpus, we wanted to cover a whole week and we therefore randomly choose week 49 in 2015 (from November 30, 2015 to December 6, 2015). The source of the data is the scientific blog portal SciLogs (SciLogs, 2016). SciLogs is subdivided into different sections (BrainLogs, ChonoLogs, KosmoLogs, WissenLogs) where scientists – and those interested in science – can interact in interdisciplinary discussions about science. For the test corpus, we did not focus on a particular section; we extracted the blogs from different sections.

The result of analyzing the SciLogs source code is that the different sections store the data using the same template. In the source code, we can find the information for title, category, keywords, date, name of the blogger / commentators, the comments and their different levels of indentation, the permalinks of the comments and the blogpost itself. The data collection for the test corpus was done manually and in the process of the work we discovered some complications that will have to be dealt with later during the automation phase. We will discuss some of these complications in Section 3.2.3.

Our next step will be to complete our corpus with the data appeared in 2015 considering all SciLogs sections. According to our current knowledge, the SciLogs data of 2015 includes about 1.200 blog posts and 12.000 comments. Retrieval of the blog data from the web will be conducted semi-automatically. For this purpose, an open source program HTTrack Website Copier (Roche, 2016) will be used. HTTrack enables the download of all kinds of the website data stored on the server including HTML pages, images and other files to a local directory on a computer. After the retrieval step, the data will be cleaned from the noise in the data and represented in form of HTMl pages (external structure). Finally, the relevant content will be extracted from the HTML pages and annotated with TEI annotation standard (internal structure). The programming language Python and its packages for XML parsing will be used for this purpose.

3.2 Data Annotation

3.2.1 Types of Blog Information

We distinguish between three types of information provided in the blog based on how the former is made available. The first type (**type A**) incorporates information which is directly available in the blog or from the source code of the blog site. In the blog post structure, it includes the blog post itself along with the meta information such as the title of the blog post, date of creation, the name of the blogger, the categories the entry belongs to and main keywords. In the structure of the comments, type A information is represented by the total number of comments as well as the name of the commentator, date and comment ID. The second type (**type B**) includes information which is not directly available but can be inferred from **type A** information, e.g. usual activity time of a commentator (at what time a particular commentator usually writes his comments). Finally, the third type (**type C**) is an interpretative information type. This kind of information is neither directly nor indirectly provided in the blog but is rather the result of statistical (basic statistics), linguistic (e.g., part-of-speeches) and discourse (e.g., topic identification with topic modeling) interpretation and analysis of the blog entries. The interpretative information type can either be collected manually or by use of computational tools.

3.2.2 Annotation Standard

To date, no standard exists for representing CMC data. One option could be to design an XML schema for CMC from scratch, which would perfectly fit the needs of our project. The main reason as to why we are not going along with XML is that the schema would be idiosyncratic and the corpus would not interoperate without causing difficulty with other resources. When searching for a standard for the representation of texts in digital form, one will take a look at the Text Encoding Initiative (TEI). However, none of the modules in the current version of the TEI Guidelines (P5) can be adopted for our project. Fortunately, the SIG CMC group under the direction of Beißwenger (TU Dortmund) has been working on the adaption of TEI guidelines to the presentation of genres of CMC since 2012 (Beißwenger, 2015). Given that no module for CMC is so far ready to use, we have started to look for schema drafts by the SIG CMC group and up to now, a couple of corpora have been released by the SIG CMC group. Among them are CMC genres like tweets, email, text chat, wiki discussions and weblogs (Chanier, 2014; Beißwenger, 2013; Storrer, 2015). The schema that fits our needs best, is the one released in 2014 by the French network CoMeRe (Communication médiée par les réseaux) (Hriba, 2013). The CoMeRe schema is based on the previous schema draft by DeRiK (Beißwenger, 2013) and includes e.g. the metadata schema for CMC. But still, there is no possibility for representing the full structure of a blog and especially the related comments. Our goal is to take the latest schema draft provided by the SIG CMC (Beißwenger, 2016) and not to try to change the main characteristics of the schema. The status of that schema is that of a "core model for the representation of CMC" (Beißwenger et al. 2012: 6). And so we will need to redefine some elements while also introducing some new ones.

3.2.3 Challenges and Possible Solutions

A number of aspects are challenging since the task of blog corpus annotation is in some cases the result of the particularities of the content management system (CMS) functionality used by our blog data source. Most of the challenges deal with the structure of the comments. As we are at an early stage of our project, only a limited number of challenges and solutions will be described here.

The first challenge is due to the absence of an editing function for the comments. The commentator who edits the text of the comment creates a new entry which appears in the timeline as an autonomous comment. Thus, the comments structure of our blog corpus includes both original comments and their edited versions appeared to the time of the data collection. Though, this aspect does not have an impact on the difficulty of the automation of the annotation task. However, it first impacts the accuracy of the total number of distinct comments (type A information). Second, it creates confusing linkages in the comments structure.

The latter problem also arises as the result of the second challenge – the possibility that one comment refers to more than one previous comment. Unfortunately, the CMS of our blog source does not offer any special options to mark or highlight multiple comment references. In some cases, the commentators use constructions such as [@name]* to overcome this problem. In other cases, an additional analysis of the comment content is required. For the purpose of the study, only explicit references are taken into consideration. No deeper content analysis has been conducted. The identification of multiple references and their annotation with TEI was processed automatically and then manually checked for mistakes in order to achieve accurate and reliable results. We believe that it is be less time- and cost-consuming than fully manual processing of the data. The automatic part is conducted based on explicit marks of multiple reference such as [@name]*. In the TEI blog annotation the multiple references are specified by enumeration of the ids of their comments (<replyTo>).

Finally, the third challenge is the task of the correct assignment of the comments to the level in the hierarchical structure of the comments. At present, the number of possible level assignments is limited to five. All comments appearing after the first comment on the fifth level are (wrongly) assigned to the fifth level. In order to solve this problem, we developed a simple algorithm to compute the correct level of the comments. The algorithm first takes the person reference ("@name", "[name] schrieb (engl.: wrote)" etc.) included in the text of the analyzed comment as the input. In the case of multiple references, only the first reference is taken into consideration. The algorithm then searches backwards for the matches between the person reference and the name of the commentator in the previous comments. Through matches, level of the analyzed comment is computed as the sum of the level assignment of the comment which the person reference belongs to and 1. By absence of the references, the level of the comment is counted subsequently.

4. Results

The main steps conducted for the purposes of a scientific blog corpus compilation as well as challenges faced during this process were described in the present study. The current version of the corpus contains 21 blog posts and 195 related comments written in the period of one week. We are convinced that comments are an essential part of a blog corpus. On their own or in connection with the correspondent blog post, they provide valuable information for processing diverse research questions on the language of the blog and its structure. For example, based on the name of the commentator and the time of his comments, we can compute at what time a particular commentator is active in the blog.

The data for our blog corpus was manually collected and annotated according to the TEI schema drafts developed by the TEI special interest group. For the annotation, three types of information (direct, indirect and interpretative) based on the availability of the latter have been identified. The present version of the corpus includes annotation of the first type - directly retrieved information (e.g., the name of the blogger, title of the blog entry, the name of the commentator etc.). The next objective of the project is an expansion and full annotation of the corpus as well as the automation of the data collection and annotation task. At the final stage of the present project, our annotated corpus will be made available to the interested community to perform diverse kinds of research and experiments. Our aim is to enable the access to the corpus through a searchable online database. Additionally, we plan to make a part of the corpus to be available upon request. For the legal aspects of the SciLogs data usage and publication an external competent institution will be consulted.

5. Acknowledgements

We would like to thank our anonymous reviewers for their insightful comments and suggestions. Following the feedback, we included several improvements in our paper.

6. References

Abendroth-Timmer, D. et al. (2014). Corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne). Banque de corpus CoMeRe. *Ortolang.fr: Nancy*. https://hdl.handle.net/11403/comere/cmr-infral (last retrieved 23 August 2016).

Barbaresi, A., Würzner, K.-M. (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In KONVENS 2014, NLP4CMC workshop proceedings, p. 2–10.

Beißwenger, M. et al. (2012). A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI), Issue 3.

Beißwenger, M. et al. (2013). DeRiK: A German reference corpus of computer-mediated communication. pp. 531-537. In: M. A. Finlayson (Eds.), LLC. The Journal of Digital Scholarship in the Humanities, Volume 28, Number 4. Oxford, OUP, pp. 531-537.

Beißwenger, M. (2015). Computer-Mediated Communication SIG. In TEI Website. http://www.tei-c.org/Activities/SIG/CMC/ (last retrieved 20 April 2016).

Beißwenger, M. (2016). SIG:Computer-Mediated Communication. In TEI Website. http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication (last retrieved 20 April 2016).

Chanier,T. et al. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Journal of Language Technology and Computational Linguistics. Berlin, GSCL, pp1-31.

Hriba, L., Chanier, T. (2013). Projet européen TEI-CMC. Comere: Corpuscomere. Communication médiée par les réseaux. In Comere Website. https://corpuscomere.wordpress.com/tei/ (last retrieved 20 April 2016).

Kehoe, A., Gee, M. (2012). Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. In Studies in Variation, Contacts and Change in English 12: Aspects of corpus linguistics: compilation, annotation, analysis. http://www.helsinki.fi/varieng/series/volumes/12/kehoe_gee/ (last retrieved 20 April 2016).

Roche, X. (2016). HTTrack. Website Copier. http://www.httrack.com/ (last retrieved 20 April 2016).

SciLogs (2016). SciLogs. Tagebücher der Wissenschaft. Spektrum der Wissenschaft Verlagsgesellschaft mbH. http://www.scilogs.de/impressum/ (last retrieved 20 April 2016).

Storrer, A. (2014). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In: A. Plewina & W. Andreas (Eds.), Sprachverfall? Berlin, De Gruyter, pp. 171-196.

Storrer, A. (2015). ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D. In CLARIN-D Website. http://de.clarin.eu/en/curation-project-1-3-german-philology (last retrieved 20 April 2016).

WebCorp (2013). Birmingham Blog Corpus. WebCorp: Linguist's Search Engine. Birmingham City University. http://wse1.webcorp.org.uk/cgi-bin/BLOG/index.cgi (last retrieved 20 April 2016).

Expressiveness in Flemish Online Teenage Talk: A Corpus-Based Analysis of Social and Medium-Related Linguistic Variation

Lisa Hilte, Reinhild Vandekerckhove, Walter Daelemans
CLiPS, University of Antwerp
E-mail: lisa.hilte@uantwerpen.be, reinhild.vandekerckhove@uantwerpen.be, walter.daelemans@uantwerpen.be

Abstract

We analyze linguistic expressiveness in an extensive corpus (2 million tokens) of Flemish online teenage talk, focusing on the use of typographic chatspeak features, an onomatopoeic and a lexical variable and its correlation with the chatters' profile and the online medium. General quantitative findings are that girls outperform boys in the expression of emotional involvement, and younger adolescents outperform the older group. However, medium has the largest impact: much more expressive markers are used in asynchronous social media posts than in synchronous instant messaging. On a qualitative level, utterances written by girls, by younger teenagers and on the asynchronous platform contain more expressive markers related to love or friendship.
Apart from the medium's (a)synchronicity and its public or private character, the nature of the interaction appears to be a determining factor too. The asynchronous social media posts involve a lot of flirting or pleasing, which drastically increases linguistic expressiveness.

Keywords: computer-mediated communication, adolescents, computational sociolinguistics

1. Introduction

Since the rise of informal computer-mediated communication (CMC), both laymen and linguists have been fascinated by the prototypical features that they identified in several forms of digital writing (see Crystal, 2001). Androutsopoulos relates these features to three dimensions or themes: "orality, compensation, and economy" (2011: 149). While orality refers to the use of spoken language features in written discourse and economy covers all strategies to shorten messages, the "semiotics of compensation" "includes any attempt to compensate for the absence of facial expressions or intonation patterns" (Baron, 1984: 125; Androutsopoulos, 2011: 149). The latter dimension is at issue in the present paper, which examines the use of expressive markers in Flemish online teenage talk.

2. Goal of the Paper

We examine social and medium-related linguistic variation concerning expressiveness in a corpus of Flemish online teenage talk. The linguistic variables include several typographic features that are generally associated with chat discourse (e.g. emoticons), an onomatopoeic variable (rendition of laughter) and a lexical variable (intensifiers[1]). All features will be discussed more elaborately in section 3. We investigate the potential (quantitative and qualitative) correlations between the use of the selected expressive markers and the profile of the chatters (in terms of age and gender) as well as the impact of the synchronicity and (largely) public versus private character of the medium on which the utterances were written.

3. Expressive Markers

First of all, the present study includes six typographic expressive markers:

- flooding (i.e. deliberate, expressive repetition) of letters
  e.g. *s**uuu**per*
- flooding of punctuation marks
  e.g. *nice**!!!***
- combinations of exclamation and question marks
  e.g. *wtf**?!?***
- capitalization of words or entire utterances
  e.g. ***FAIL***
- emoticons
  e.g. *dude **:P***
- typographic rendering of kisses or hugs and kisses
  e.g. ***Xxxx***

The onomatopoeic marker studied in this research is the rendering of laughter in CMC, which includes all variants of *haha* and *hihi*.

e.g. ***hahahaha***

Finally, we added a lexical variable, i.e. the use of intensifiers: "items that amplify and emphasize the meaning of an adjective or adverb" (Stenström, Andersen & Hasund, 2002: 139). In Dutch, these items can either be adverbs or intensifying prefixes.

e.g. ***Super**mooie t-shirt* '**super** nice T-shirt'

4. Corpus and Methodology

4.1. Corpus

Our corpus consists of 400 808 online messages or 2 066 521 tokens[2]. The messages were produced between 2007 and 2013 by adolescents from Dutch-speaking northern Belgium (Flanders), all aged between 13 and 20 years old. The utterances were written on both a synchronous electronic medium (private instant messaging) and an asynchronous electronic medium (private and public messages on a social media site). Table 1 shows the distribution of the tokens over the age and gender groups

[1] We sincerely thank Jens Vercammen for the data processing for this variable.

[2] These tokens are the result of splitting the text on whitespace. A token can be a word, but also an emoticon or isolated punctuation marks.

and the two media. We note that, although there is an imbalance for all three social variables (e.g. more male than female material), the smaller subcorpora are always sufficiently large and thus do not exclude valid testing for the three variables.

	GIRLS		BOYS		
	YOUNGER	OLDER	YOUNGER	OLDER	**total**
SYNC.	118 694	176 233	29 146	973 061	1 297 134
ASYNC.	463 277	67 257	162 077	76 776	769 387
total	581 971	243 490	191 223	1 049 837	**2 066 521**

Table 1: Distribution of variables in the corpus.

4.2. Methodology

The typographic and onomatopoeic expressive markers were automatically detected and counted using Python scripts. The coverage of the software was evaluated and judged accurate on a test set of 1000 randomly chosen posts from the corpus by comparing a human annotator's feature extraction to the software's output. The intensifiers were automatically extracted using a predefined list[3] (which covered most of the intensifiers used in the corpus) and a frequency cutoff to not take into account very infrequent variants. The software's output was manually screened and filtered. To evaluate the human judgment, finally, a test set of 700 utterances was screened by two annotators, who obtained a low error rate (1.57%).

5. Results and Discussion

To verify the statistical significance of our quantitative findings, we combined chi square tests with a bootstrapping approach (with Monte Carlo resampling), to obtain more solid results than when performing one single chi square test on the entire data set[4]. The statistical values we report in the next paragraphs (p-values, Cramer's V scores and odds ratios) are the mean of the values for all samples.

5.1. Quantitative Findings

We quantified the degree of expressiveness by counting all markers in the subcorpora and dividing these counts by the number of tokens in the subcorpora. This approach led to relative expressiveness scores or ratios. The entire data set contained 295 127 expressive markers, which is a ratio of 14.28% (in terms of tokens – in terms of types: 21 427 markers, or a ratio of 11.88%). An overview of the ratios per independent variable is shown in Table 2. The asynchronous posts contain the highest relative number of expressive markers (28.35%), followed by the younger participants' texts (25.23%) and the girls' texts (21.77%).

Female	Male
21.77%	9.30%
Younger (13-16)	Older (17-20)
25.23%	7.74%
Asynchronous posts	Synchronous posts
28.35%	5.94%

Table 2: Overview of expressiveness ratios per subcorpus.

General tendencies for the social variables are that the girls use significantly more expressive markers than the boys (p < .001), that younger teenagers use significantly more expressive features than older ones (p < .001) and that significantly more expressive writing is used on asynchronous media (p < .001). These general tendencies also hold for each of the analyzed expressive markers: the female (resp. younger, resp. async.) texts contain *each* expressive marker significantly more often than the male (resp. older, resp. sync.) texts.

As for the strength of the correlation between the linguistic and independent variables, the strongest correlation can be found for medium (Cramer's V = 0.31), followed by age (Cramer's V = 0.24) and gender (Cramer's V = 0.17). The same order can also be found for effect size: medium has the largest effect size (odds ratio = 6.27), followed by age (odds ratio = 4.02) and gender (odds ratio = 2.71). These scores should be interpreted as follows: the odds that a token contains an expressive marker are 6.27 times higher if the token is produced on the asynchronous platform than when produced on the synchronous platform[5]. Medium seems to be the most interesting independent variable when it comes to expressiveness, as the correlation with the linguistic variables is very high and the actual effect size is large as well.

Some expressive features both heavily correlate with the social variables and are used very differently (quantitatively) by the subgroups of the same social variable. This is the case for letter flooding (i.e. deliberate, expressive letter repetition) and the rendition of kisses (e.g. 'xxx'), especially with regards to medium. The odds ratios are respectively 51.85 (kisses – medium) and 16.33 (letter flooding – medium): for each occurrence of kisses (flooding letters, resp.) in the synchronous chat messages, 51.85 occurrences (16.33, resp.) can be expected in the asynchronous posts.

5.2. Qualitative Findings

On a qualitative level, some constants could be found among all different subgroups. The most popular expressive markers in all groups are emoticons and punctuation flooding (deliberate repetition of question and exclamation marks). These features' popularity could be

[3] In alphabetical order: (1) *bere*, (2) *echt*, (3) *echt wel*, (4) *erg*, (5) *fucking*, (6) *gans*, (7) *heel*, (8) *kei*, (9) *kweetniehoe*, (10) *loei*, (11) *mass(as)*, (12) *massiv*, (13) *mega*, (14) *muug*, (15) *over*, (16) *overdreven*, (17) *so*, (18) *super*, (19) *vies*, (20) *vree*, (21) *zeer*, (22) *zo*, (23) *zot*.

[4] We thank Giovanni Cassani and Dominiek Sandra for their help and advice in the statistical aspect of the research.

[5] Note that these numbers differ from the ratios reported in Table 2. Although both numbers express a similar concept, the calculation behind them is different, as sample sizes of both subcorpora are taken into account to calculate odds ratio and not to calculate the straightforward percentages.

due to their ‘explicit’ expressive nature: many emoticons represent facial expressions and question and exclamation marks are the most expressive punctuation marks. Apparently, because of the explicit nature of these features, they are very obvious and favored markers.

As for letter flooding, we note that in all subgroups, mainly vowels are repeated, and hardly ever plosives. This supports the hypothesis that flooding is the orthographic representation of an oral phenomenon (Darics, 2013: 144), i.e. the lengthening of sounds, which is easiest for vowels and impossible for plosives.

A third general tendency is the top position of the Dutch first person singular pronoun ‘ik’ (I) among the lexemes written in capital letters. As pronouns are function words, they are automatically used more frequently (Newman et al., 2008: 216; Pennebaker, 2011: 27). However, the top position of ‘ik’ could also be symptomatic of the fact that when the teenagers write in a very expressive way, they often talk about something personal. This finding also suggests that quite often entire utterances are written in capitals, as merely capitalizing function words would make less sense (although the chatters could, of course, only emphasize the word ‘I’ in their utterance to stress its importance).

Finally, the qualitative in-depth analyses for each of the expressive markers also lay bare correlations between the independent variables. Strikingly, similar tendencies could be noted for texts written by female participants, by younger teenagers, and on the asynchronous medium. These texts contain a lot more expressive markers related to love and friendship. The most popular emoticons were related to love (e.g. heart-emoticons: <3) and many of the top lexemes that were written in allcaps concerned love or friendship (e.g. ‘LOVEYOU’, ‘BFF’: *best friend forever*). These results are incongruent with male texts, the texts written by older adolescents or the synchronous posts. E.g.: While heart-emoticons were much favored by girls, they were at the bottom of the list of the emoticons produced by boys.

However, some caution might be needed when interpreting these correlations, as there is an imbalance in our dataset which could (partially) influence our results: many of the female participants are also younger adolescents, often writing on the asynchronous medium, whereas many of the male participants are also older teenagers, often writing on the synchronous chat platform. Still, linguistic correlations between gender and age have been reported on before (Argamon et al., 2007; Pennebaker, 2011; Schwartz et al., 2013). Stylistic correlations concern the use of function words: men and older people use more articles and prepositions, whereas younger people and women use more pronouns, conjunctions and auxiliary verbs (Pennebaker, 2011: 66; Argamon et al., 2007: n.pag.; Schwartz et al., 2013: 8-9). On a content-related note, Argamon et al. report that men and older people prefer topics like politics, religion and business, whereas women and younger people prefer discussing home, romance and fun (2007: n.pag.). These findings correspond to the younger and female teenagers’ preference for expressive markers related to love and friendship. As for medium, however, no correlations have been reported between the way people write on certain platforms and their gender or age. This could thus be an artefact of the imbalance in our dataset. Another possible explanation lies in the nature of our asynchronous texts. Although many posts on the asynchronous medium are public, the interaction often has a largely personal character. Many comments on this social medium involve flirting and/or pleasing (e.g. in positive reactions to other users’ pictures). In this respect, our asynchronous medium differs from other social media, like Twitter, where the writing is less personal and more targeted at informing a wider audience, rather than at bonding or pleasing[6]. The latter focus prevails in our asynchronous data, which could explain the higher rate of love-related expressive markers in this subcorpus.

6. Conclusion

This paper discussed linguistic expressiveness in (Belgian) Dutch informal computer-mediated messages. We included typographic CMC features (e.g. emoticons), an onomatopoeic variable (the rendition of laughter) and a lexical feature (the use of intensifiers) and looked for possible correlations between these linguistic variables and the authors’ profile (gender, age) versus the CMC medium. Girls appeared to outperform boys in the use of expressive markers, and so did the younger adolescents compared to the older ones. The results were extremely consistent in this respect: the same tendencies could be observed for each of the expressive markers. Quite strikingly however, medium appeared to have the largest impact (more expressive writing in asynchronous and largely public than in synchronous and mainly private posts). The qualitative analyses show that girls and younger teenagers produce more love-related expressive markers than boys and older adolescents. And again, remarkably, these types of correlations were found for medium too (with more love-related markers used in the asynchronous than in the synchronous posts).

The present research differs from previous research into expressive markers in CMC in that it includes a wider range of expressive markers (both lexical and typographic) and combines three independent variables (age, gender and medium). While gender and to a minor extent age have received ample attention in related research, the present findings highlight the importance of the variable medium. They call for refinement of this variable, since apart from (a)synchronicity and the public versus private character of the medium, the character and goal of the interaction seem to be determinant factors too and consequently need to be operationalized in future research.

7. References

Androutsopoulos, J. (2011). Language Change and Digital Media: A Review of Conceptions and Evidence. In: T.

[6] We thank Lieke Verheijen for pointing out this difference.

Kristiansen & N. Coupland (Eds.), *Standard Languages and Language Standards in a Changing Europe.* Oslo: Novus, pp. 145--161.

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2007). Mining the Blogosphere: Age, Gender and the Varieties of Self-Expression. *First Monday,* 12(9), n.pag.

Baron, N.S. (1984). Computer Mediated Communication as a Force in Language Change. *Visible Language*, 18(2), pp. 118--141.

Crystal, D. (2001). *Language and the Internet.* Cambridge: Cambridge University Press.

Darics, E. (2013). Non-verbal Signalling in Digital Discourse: The Case of Letter Repetition. *Discourse, Context and Media,* 2, pp. 141--148.

Newman, M.L., Groom, C.J., Handelman, L.D. & Pennebaker, J.W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3), pp. 211--236.

Pennebaker, J.W. (2011). *The Secret Life of Pronouns. What Our Words Say About Us*. New York: Bloomsbury Press.

Schwartz, A.H., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M. et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), e73791.

Stenström, A.B., Andersen, G., & Hasund, I.K. (2002). Non-Standard Grammar and the Trendy Use of Intensifiers. In: A.B. Stenström, G. Andersen & I.K. Hasund, *Trends in Teenage Talk. Corpus Compilation, Analysis and Findings.* Amsterdam: John Benjamins, pp. 131--163.

French Wikipedia Talk Pages: Profiling and Conflict Detection

Ho-Dac L.-M.(*), Laippala V.(**), Poudat C.(***) and Tanguy L.(*)
(*) CLLE, University of Toulouse, CNRS, UT2J, 5 allées A. Machado, 31058 Toulouse CEDEX 9, France
(**) TIAS, University of Turku, 0014 Turun yliopisto, Finland
(***) BCL, University of Nice Sophia Antipolis, 24, avenue des Diables bleus, 06357 Nice CEDEX 4, France
E-mail: hodac@univ-tlse2.fr, mavela@utu.fi, celine.poudat@unice.fr, tanguy@univ-tlse2.fr

Abstract

Wikipedia is a popular and extremely useful resource for studies in both linguistics and natural language processing (Yano and Kang, 2008; Ferschke et al., 2013). This paper introduces a new language resource based on the French Wikipedia online discussion pages, the WikiTalk corpus. The publicly available corpus includes 160M words and 3M posts structured into 1M thematic sections and has been syntactically parsed with the Talismane toolkit (Urieli, 2013). In this paper, we present the first results of experiments aiming at classifying and profiling the talk pages and threads in order to determine criteria for selecting discussions with conflicts.

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

1. Introduction

With the exponential development of the Internet, new communicative situations and new genres have come about. The new web genres, which are not yet fully characterized, are complex objects challenging the existing methodologies and analysis tools: the Wikipedia encyclopedic project is one of these new textual objects that can be studied under the umbrella term Computer-Mediated Communication (CMC, (Herring et al., 2013)). Wikipedia, which celebrates its 15th birthday this year, is an open and collaborative project, available in numerous languages. The success of the web encyclopedia is indisputable, as evidenced by its huge size (5M articles in the English Wikipedia / 1.7M articles in the French Wikipedia as of June 2016). In addition, Wikipedia is one of the 10 most consulted websites in the world (Alexa, June 2016).

Over the last decade, Wikipedia has become a wealth of information which is more and more used by natural language processing (NLP) and text mining applications (Ferschke & al. (2013) propose an overview of the use of Wikipedia in NLP). It has also been the subject of many studies in social sciences. After the quality of the encyclopedia has been established by (Giles, 2005), a large number of studies use Wikipedia for describing human coordination and collaboration processes (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 edition, 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, certainly because of the sprawling structure of Wikipedia (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, an interesting result for our project is that article reverts (when users restore a previous version) are proven significant features to detect 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, etc.

Such criteria serve as the basis for the automatic detection of quality articles (Wilkinson and Huberman, 2007), conflictual pages (Kittur et 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 edition, the linguistic characteristics of Wikipedia pages remain little explored. In particular, talk pages are specifically interesting to observe as they are at the heart of the Wikipedia device. Each article is associated with a talk page, where most of the coordination work is done, and where the 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 on the article, including sections to be rewritten or suppressed (Ferschke et al., 2012).

Wikipedia talk pages 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 that respect, they might be compared to forum discussions except that they rely on a specific Wiki device 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 notably to editing actions, conflict management and Wikipedia procedures (e.g. NPOV, i.e. Neutrality of Point of View, relevance, source, quality etc.). Conflicts are particularly interesting to observe in 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 usual 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 signaling that a conflict is ongoing on the article or talk pages (e.g. NPOV or relevance disputes, ”Keep calm” banner). Examples of pages with such labels are quite numerous: *Abortion in Iran*, *Bengali cuisine*, *List of Volvo trucks* to cite just a few.

The aim of the present study is twofold: at a descriptive level, we would like to contribute to the linguistic description of Wikipedia talk pages, which have been little explored using linguistic criteria. In particular, few linguistic studies have been conducted on French Wikipedia (see (Denis et al., 2012) on the detection of conflicting threads or (Poudat and Loiseau, 2007) on the exploration of Wikipedia categories). We will first perform an automatic classification on the entire set of French Wikipedia talk pages, which were gathered within the WikiTalk Corpus, making the most of the French ”Appel au calme” (keep calm) label, signaling ongoing conflict(s) on the talk page. In order to have a broader view of the linguistic characteristics of the French Wikipedia talk pages, We will then propose a profiling of the genre, using a mutidimensional analysis enabling us to highlight key features and oppositions at a global level. Conflicting threads and pages will be characterized within this global generic profile.

2. 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 them, even if these talk pages are basically online discussions. Only 24% of the remaining talk pages contained more than two words[1].

The 365,612 remaining talk pages were segmented into threads and posts based on the wikicode. Threads correspond to divisions delimited by (sub)headings signaled by the wiki markup: /==.*?==/. Posts are delimited according to

1. timestamp and an optional user signature, such as: *Viking59 10 mai 2009 à 17:16 (CEST)*; or

2. a change in the interactional level indicated by the number of semi-colons (:) in the beginning.

Once threads and posts were delimited, all discussions were formatted according to the TEI-P5 guidelines. Metadata are encoded in the teiHeader as illustrated below with the <classDecl> element.

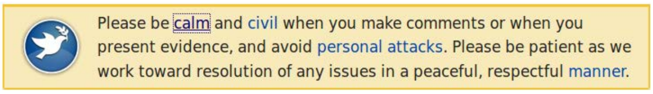
```
<category type="discipline">
   <catDesc>Politique</catDesc>
   <catDesc>France</catDesc>
</category>
<category type="avancement">
   <catDesc>Featured</catDesc>
</category>
<category type="interaction">
   <catDesc>{{calm}}</catDesc>
</category>
```

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

1. ”discipline” indicates associated thematic portals,

2. ”avancement” corresponds to article’s quality scale based on Wikipedian assessments[2],

3. ”conflictness” gives information about possible conflicts in the discussion. Such information may be manually inserted by Wikipedians via the template {{keep calm}} which adds the following banner at the top of the talk page[3]).


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.

Discussion structure is encoded according to the following TEI elements:

- <div> for threads
- <head> for topic titles and
- <post who="user" when="timestamp" interactionalLevel="#"> for posts.

Table 1 gives a quantitative overview of the WikiTalk corpus[4].

discussions	sections	posts	words
365,612	1,023,841	2,406,514	161,833,298

Table 1: Quantitative overview of the WikiTalk corpus.

Eight of the extracted talk pages, amounting to 413 posts and 47,284 tokens, were manually inspected to evaluate the extraction process. 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.

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

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

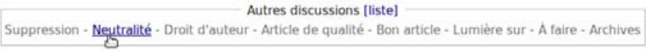
[3] https://en.wikipedia.org/wiki/Template:Calm

[4] Soon available at http://redac.univ-tlse2.fr/

3. Classification of Conflicting vs. Peaceful Talk Pages

The first tested method consisted in a data-driven comparison of the global linguistic characteristics of two classes of talk pages, distinguished according to an experimental classification of ”conflicting” vs. ”peaceful” talks. The selection criteria used for distinguishing between these two classes are based on the Wikipedians’ assessment of the article’s quality and the Wikipedians’ alert regarding conflict or impoliteness in a talk page. Moreover, only talk pages containing more than 100 words were taken into account. Among those, 2,028 a priori ”conflicting” talks (11M words) were selected according to the following criteria:

- `<category type="interaction">` in teiHeader indicates that the ”keep calm” template was inserted;
- a parallel talk page was created for discussing the article’s neutrality[5];



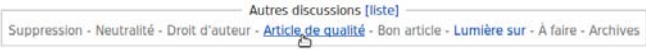
Autres discussions [liste]

Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives

- the page itself is a parallel talk page created for discussing the article’s neutrality.

Criterion for selecting 4,569 a priori ”peaceful” talks (8.8M words) are the following:

- `<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.



Autres discussions [liste]

Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives

For the purpose of evaluating our distinction between these two classes while also determining features that may be used for selecting talk pages where conflicts may occur, we trained a text classification model using the Vowpal Wabbit linear classifier (Agarwal et al., 2011). In addition to being fast and easily adjustable to large corpora, it has the advantage of generating a list of the most significant features and their relative weights.

Two feature sets were tested for the classification task: lexical features and syntactic features. Classification based on lexical features which considers texts as bags-of-words or bags-of-lemmas is the traditional approach, as for example (Scott et al., 2006) which propose a keyword analysis for reflecting thematic and stylistic features. Classification based on syntactic features which considers texts as bags-of-syntactic N-grams more or less lexicalized is less common (Kanerva et al., 2014; Goldberg et al., 2013). This method enables a more robust analysis on text characteristics that does not depend on the text topic but attempts to generalize the level of description beyond individual lexical topics to typical structures (Laippala et al., 2015).

The classification is performed using the stochastic gradient method with two-thirds of the corpus used for training and the remaining for testing. As lexical features we use lemmas; as syntactic features we use unlexicalized *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 Fig. 1.

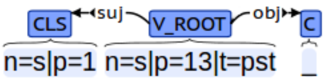


Figure 1: A delexicalized syntactic bi-arc describing a clitic+verb+conjunction as in the clause ‘I find that’.

Syntactic analysis and lemmatisation were provided by the Talismane toolkit (Urieli, 2013). Two levels of text segments were considered: threads and posts. Entire pages were not taken into account because a conflict usually happens inside a thread. In addition, our previous experiments on the page-level have already shown higher scores for the bag of words method (Ho-Dac and Laippala, 2015). In the analysis, we consider, however, that all the posts and threads in a page labeled as conflicting / peaceful are in the same category. Table 2 gives the precision (P) and recall (R) for detecting the ”conflict” category by using the two feature sets on threads and posts.

	threads		posts	
features	P	R	P	R
lemmas	0.84	0.60	0.79	0.69
bi-arcs	0.55	0.48	0.63	0.59
units	46,690		194,289	

Table 2: Comparison of lexical vs. syntactic approaches for the automatic classification of conflicting threads and posts.

Results show that the best method for detecting conflict seems to be a classification of threads by using a lexical approach. A closer look on the threads classified with high probability and on typical bi-arcs used by the classifier is necessary for better understanding.

Even if the precision of more than 80% seems encouraging, we must admit that these results lead us to question both the features used for classification and our *a priori* definition of a conflicting talk. Next sections begin to address these questions by proposing a range of new features for profiling Talk pages in a bottom-up approach and presenting a current project of conflict manual annotation in the WikiTalk corpus.

4. A Bottom-Up Approach to Talk Page Profiling

The automatic classification was supplemented by a second approach which uses statistical techniques based on linguistic features and portals information for discovering talk pages and thread profiles in a bottom-up approach, without a focus on conflict. This method considered all the 366,612 talk pages and used the R package FactoMineR dedicated

[5] This possibility seems specific to the French Wikipedia

to multivariate exploratory data analysis[6]. Each talk page and thread was automatically described with four types of features:

- **THEMA**: portal sections of the associated article page knowing that an article may be categorized as belonging such as *Art, History, Sport*[7] up to 7 of the 11 possible Wikipedia sections (these 11 variables were binarised);
- **GLOBAL**: general quantitative characteristics (number of words and posts) and, for entire talk pages, amount of threads and different contributors, proportion of anonymous posts;
- **INTERACT**: the frequency of a wide range of interaction and politeness cues per talk pages and threads (social deixis, marks of agreement and disagreement);
- **DISCREL**: the frequency of connectives for each discourse relations as defined in the LEXCONN, "a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey" (Roze et al., 2012).

A Principal Components Analysis on talk pages and threads extracted 5 dimensions that explain around 30% of the total variance (29.2% for entire talk pages, 32.4% for threads). The first dimension is simply related to the size of the text units. The second dimension is more interesting and the correlated features differ between talk pages and threads. As for talk pages, it opposes

- talk pages with politeness cues (*thanks*, *hello*, *cheers*, *please*, etc.), formal *you* (*vous*) and *we* (*nous*) and discourse relations expressing concession, condition and temporal relations; to
- talk pages with more discourse relations expressing contrast, background/narration and causality.

As for threads, dimension 2 opposes

- threads with agreement cues (*ok*, *agree*, *of course*, *yes*, *no*, etc.), formal *you* and discourse relations expressing alternation, consequence, goal and temporal relations; to
- threads with more *I*, informal *we* (*on*) and discourse relations expressing contrast.

A third dimension that may be relevant gathers together talk pages (as threads) in which more connectives expressing narrative relations (*then*, *later*, *once*, *before*, etc.) and consequence relations (*in this case*, *in this respect*, etc.) occur. We may also notice that no THEMA features are significant for any dimensions.

More precise details defining these profiles will be presented during the presentation, with a focus on extreme talk pages and threads on each dimension. Our next goal is to locate conflicting threads in this 5 dimensional space.

5. Perspective: Exploring Conflicts at the Thread Level

In this paper, we have proposed different ways to explore Wikipedia talk pages; CMC genres are indeed complex objects that challenge our traditional methods and 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 conflicting talk pages are still fairly inaccurate. In addition our definition of a conflict discussion must be revised. Several paths are currently being followed, including (i) using other criteria, starting with the dimensions with identified in the profiling step; (ii) using more detailed categories, combining the article labels signaling conflicts, and the talk page labels; and (iii) using a dataset of manually annotated talk pages. We are currently annotating the threads of 30 talk pages extracted from the WikiTalk corpus in terms of conflicts (degree, intensity, type) thanks to a CORLI grant[8]. We just led a first annotation experience, following the example of (Denis et al., 2012), which enabled us to bring interesting contrasts to light (Poudat et al., 2016).

For the moment, two talk pages have been annotated, totalling 255 threads for which coders have just to indicate if the thread is conflict or not with a very basic definition. As Table 3 shows, around one thread on 2 was annotated as conflicting.

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

Table 3: Conflicting annotated threads in two talk pages.

6. References

Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, volume Part II of *CICLing'11*, pages 277–288, Berlin, Heidelberg. Springer-Verlag.

Agarwal, A., Chappelle, O., Dudik, M., and Langford, J. (2011). A reliable effective terascale linear learning system. *JMLR*, 15:1111–1133.

Brandes, U. and Lerner, J. (2007). Revision and co-revision in wikipedia: Detecting clusters of interest. In *Proceedings of International Workshop Bridging the Gap Between Semantic Web and Web 2.0, 4th European Semantic Web Conference (ESWCÂ´07)*, Innsbruck, Austria.

Denis, A., Quignard, M., Fréard, D., Détienne, F., Baker, M., and Barcellini, F. (2012). Détection de conflits

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

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

[8] TGIR Huma-Num CORLI (Corpus, Languages and Interactions, French National Consortium)

dans les communautés épistémiques en ligne. In *TALN-Actes de la Conférence sur le Traitement Automatique des Langues Naturelles-2012*.

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012). Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics.

Ferschke, O., Daxenberger, J., and Gurevych, I. (2013). A survey of nlp methods and resources for analyzing the collaborative writing process in Wikipedia. In *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Goldberg, Y., , and Orwant, J. (2013). A dataset of syntactic-n grams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), 1. Association for Computational Linguistics*.

Herring, S., Stein, D., and Virtanen, T. (2013). *Pragmatics of computer-mediated communication*, volume 9. Walter de Gruyter.

Ho-Dac, L.-M. and Laippala, V. (2015). Les discussions wikipedia : un corpus pour caractériser le genre "discussion". In *International Research Days Social Media and CMC Corpora for the eHumanities*, Rennes, France, october.

Kanerva, J., Luotolahti, J., Laippala, V., , and Ginter, F. (2014). Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Proceedings of the Sixth International Conference Baltic HLT*.

Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM.

Kittur, A. and Kraut, R. E. (2010). Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224. ACM.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007). He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462. ACM.

Kittur, A., Chi, E. H., and Suh, B. (2009). What's in wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1509–1512, New York, NY, USA. ACM.

Laippala, V., Kanerva, J., and Ginter, F. (2015). Syntactic ngrams as keystructures reflecting typical syntactic patterns of corpora in finnish. *Procedia - Social and Behavioral Sciences*, 198:233 – 241.

Miller, N. (2012). Characterizing conflict in wikipedia. *Mathematics, Statistics, and Computer Science Honors Projects*.

Potthast, M., Stein, B., and Gerling, R. (2008). Automatic vandalism detection in wikipedia. In *Advances in Information Retrieval*, pages 663–668. Springer.

Poudat, C. and Loiseau, S. (2007). Représentation et caractérisation lexicale des sciences dans wikipédia. *Revue française de linguistique appliquée*, 12(2):29–44.

Poudat, C., Vanni, L., and Grabar, N. (2016). How to explore conflicts in french wikipedia talk pages? In *JADT*, pages 645–656.

Roze, C., Danlos, L., and Muller, P. (2012). Lexconn: A french lexicon of discourse connectives. *Discours*, 10.

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

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2008). Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, April.

Suh, B., Chi, E. H., Pendleton, B. A., and Kittur, A. (2007). Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 163–170. IEEE.

Sumi, R., Yasseri, T., Rung, A., Kornai, A., and Kertész, J. (2011). Characterization and prediction of wikipedia edit wars. In *Proceedings of the ACM WebSci'11*, pages 1–3, Koblenz, Germany, June 14-17 2011.

Urieli, A. (2013). *Analyse syntaxique robuste du français : concilier methods syntaxiques et connaissances linguistiques dans l'outil Talismane*. Ph.D. thesis, Université de Toulouse - Jean Jaurès.

Viégas, F. B., Wattenberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM.

Viegas, F., Wattenberg, M., Kriss, J., and van Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, pages 78–78, January.

Vuong, B.-Q., Lim, E.-P., Sun, A., Le, M.-T., Lauw, H. W., and Chang, K. (2008). On ranking controversies in wikipedia: Models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 171–182, New York, NY, USA. ACM.

Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and Quality in Wikipedia. In *Proceedings of the 2007 International Symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA. ACM.

Yano, T. and Kang, M. (2008). Taking advantage of wikipedia in natural language processing term project report. *Language and Statistics*, II:11–762.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertész, J. (2012). Dynamics of conflicts in wikipedia. *PloS one*, 7(6):e38869.

Slovene Twitter Analytics

Nikola Ljubešić,[*‡] **Darja Fišer**[†*]
[*] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[‡] Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
† Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
E-mail: nikola.ljubesic@ijs.si, darja.fiser@ff.uni-lj.si

Abstract

The paper presents the results of metadata analysis in a corpus of 7.5 million Slovene tweets. In our analyses we primarily focus on the weekly and daily posting dynamics, their dependence on the account type (corporate vs. private) and user gender, as well as the dependence of the mentioned variables on retweeting, favoriting, text standardness and text sentiment. Through these analyses we gain insight into both user behaviour on social networks and the available linguistic material.

Keywords: Twitter corpus, meta-data analysis, Slovene language

1. Introduction

The large volumes of content generated by Twitter users as well as Twitter's proactive policy have sparked a new venue of research that is attractive for a wide range of disciplines, including information and computer science, media and communication studies, and linguistics. Twitter analytics has been successfully employed to discriminate between different types of users (Mislove et al., 2011) and behaviour (Pennacchiotti and Popescu, 2011; Rao et al., 2010). With state-of-the art techniques, a number of latent user attributes can be identified, such as their location (Hecht et al., 2011), gender (Burger et al., 2011), age (Nguyen et al., 2013), occupation (Hu et al., 2016), social class (Borges et al., 2014) and personality type (Quercia et al., 2011).

This paper is our first attempt at twitter analytics of the Slovene JANES Tweet v0.4 corpus (Fišer et al., 2016a) which contains 7.5 million tweets or 107 million tokens that were posted by nearly 9,000 different users between June 2013 and January 2016. Our goal is to gain insight into user behaviour on social networks and their language characteristics. In addition to the automatically harvested metada during tweet collection, such as posting time, no. of favourites and retweets, the corpus was enhanced with a set of manually and automatically assigned metadata at both user and tweet level. At user level, account type (private / corporate) and user gender (male / female) were manually assigned, while at tweet level text standardness (completely standard / slightly non-standard / very non-standard) and sentiment scores (positive / negative / neutral) were automatically computed.

2. Related Work

Rios and Lin (2013) have used tweet timestamps to visualize annual tweeting dynamics in different cities all over the world, discovering some interesting cultural differences. Scheffler and Kyba (2016), on the other hand, have examined the morning routine of German Twitter users and have found it to be bound to the social norms of working life.

While gender studies on Twitter predominantly focus on gender classification, (Bamman et al., 2012) give a detailed overview of the commonly attributed characteristics of male and female language and behaviour relevant for our study: language standardness (women more standard than men), communication style (men more *informative*, women more *involved*), and characteristic vocabulary (with women exhibiting more distinct features than men, such as frequent use of emoticons, expressive lengthening of words, repeated exclamation marks, etc.).

The typology and granularity of user types varies greatly in the literature. While they typically exceed the two classes used in our corpus, most researchers distinguish *organizations*, such as news media outlets and public institutions from other users. Arakawa et al. (2014) have the closest reading to our *corporate users* in their *organizations* category, which they were able to classify with the highest accuracy. They report that tweets from organizations posted the highest number of tweets the objective of which is to transmit information, which is characterized by a distinctly high use of nouns, polite language, hashtags, URLs and retweets.

The relationship between gender and subjective language in tweets has been explored for English, Spanish and Russian by Volkova et al. (2013) who have shown that there are substantial differences in the use of subjective words (e.g. *weakness*, which is used to express positive sentiment by women and negative by men), hashtags (e.g. *baseball*, which expresses positive sentiment by men and negative by women) and emoticons (with women using more emoticons overall than men in English and Spanish but, interestingly, not in Russian) and that these differences can improve sentiment classification.

3. Posting Dynamics

The first part of our statistical analyses focuses on the volume of posts, retweets and favourites. We inspect the

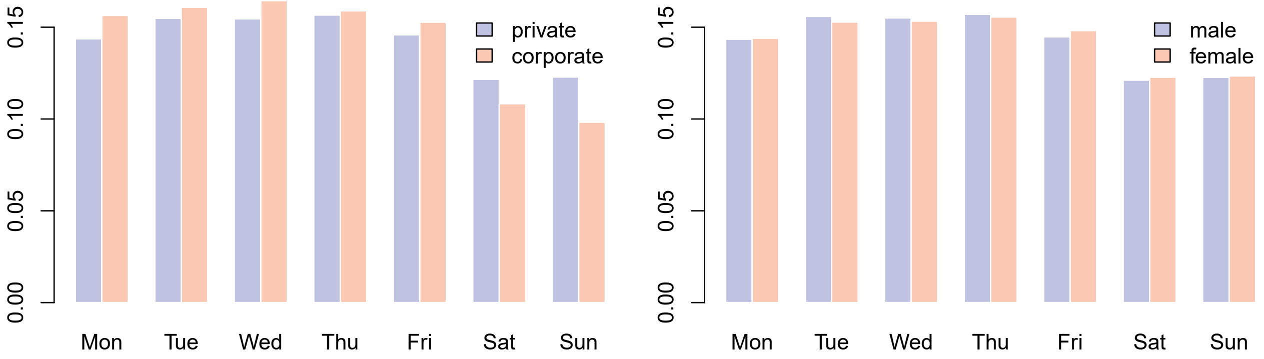


Figure 1: Probability distribution of tweets by day of week, separate by source (left) and gender (right).

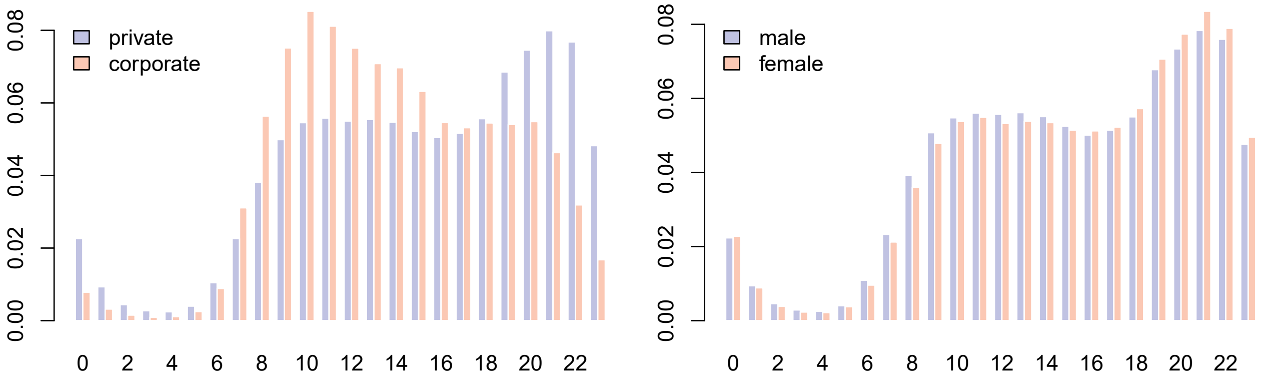


Figure 2: Probability distribution by hour in day, separate by source (left) and gender (right).

weekly and daily posting cycles and their dependence on the account type (private vs. corporate) and user gender (male vs. female). Finally, we inspect the dependence of the two last variables and post retweeting and favouriting.

3.1. Weekly Posting Cycle

The weekly posting cycle is presented in Figure 1 where the graph on the left shows distributions for private and corporate accounts while distributions for male and female users of private accounts are displayed on the right. We can see that while the overall volume of tweets posted is higher on weekdays, corporate users are dominant during the week and private ones on weekends which is not surprising but does have important implications on the topics and the language of the tweets published during the week vs. on weekends. Genderwise the distributions are very similar to the type of user, with male users prevailing mid-week and females on weekends.

3.2. Daily Posting Cycle

Figure 2 shows the daily posting cycle with user behaviour per account type displayed on the left and behaviour per user gender limited to private accounts on the right. As expected, tweeting volume of corporate users peaks during morning hours (11 a.m.) while private users are most active in the evening (9 p.m.). Interestingly, both types of users have a secondary peak that coincides with the period of the major peak of the other group. In terms of user gender, male users dominate slightly from 1 a.m. to 3 p.m. after which female users take over and are more active throughout the afternoon and evening, suggesting that male users display behaviour a bit similar to corporate accounts while females display a distinct private-use behaviour tweeting in their spare time after work.

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

Table 1: Probabilities of tweets to be retweeted, i.e. favorited, given account type and user gender variables.

3.3. Retweets and Favorites

Next we make comparisons between the retweeted and favorited variables on one side and the source and gender variables on the other. We operationalise the retweet and favorite variables as binary variables that are true if a tweet was retweeted or favorited, respectively. We present the percentages of the retweeted or favourited tweets given the account type or user gender in Table 1.

We begin by inspecting the dependence of the source variable and the retweet variable. The probability of a corporate tweet to be retweeted is twice as high as for private tweets, which was to be expected as the primary function of most corporate tweets is information dissemination. Running the chi-square test of independence proves for the vari-

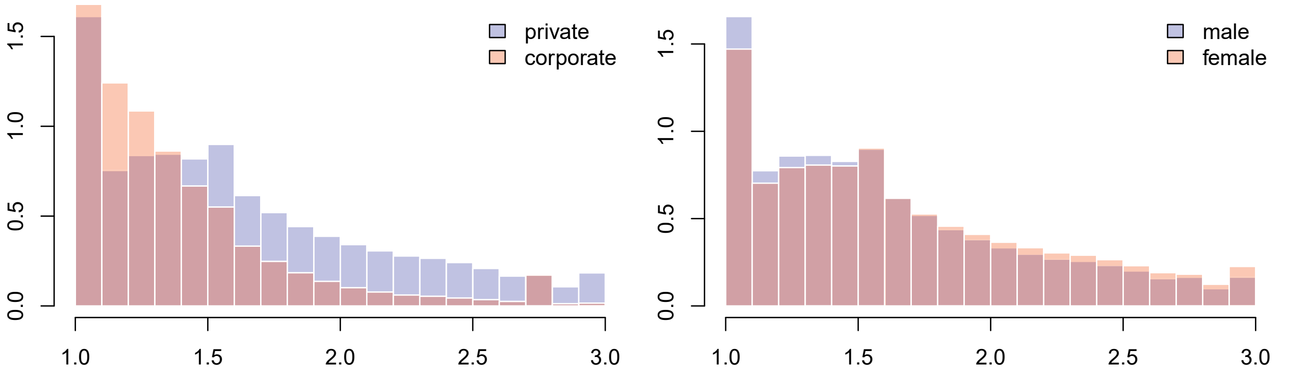


Figure 3: Distribution of the three standardness levels by account type (left) and user gender (right). Lilac represents distribution overlap.

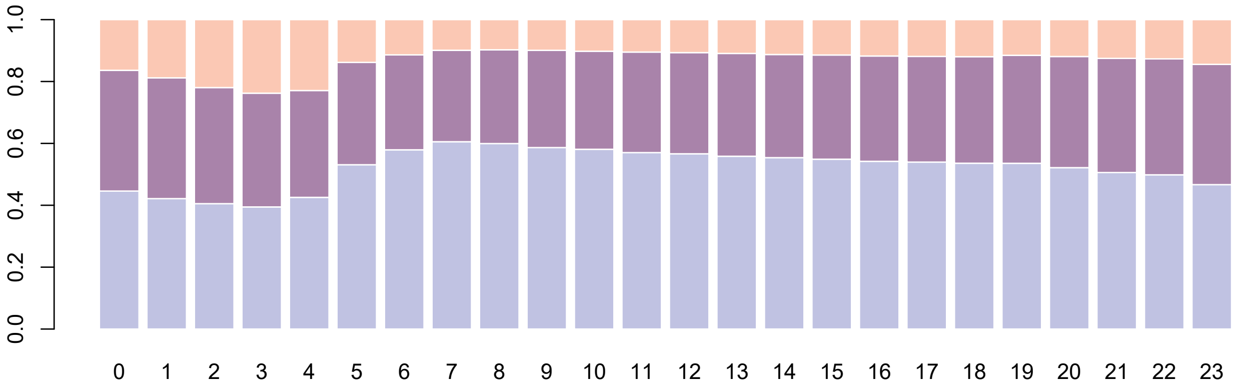


Figure 4: Standardness by hour of day, standard represented with blue, slightly non-standard with lilac, very non-standard with red.

ables of source and retweets not to be independent with $X^2(1, N = 7503200) = 74308, p < .001$.

Similarly, analysing the dependence of the source variable and the favorite variable, we measure that private tweets tend to be almost twice as frequently favorited as corporate tweets which is again consistent with the communicative role of private posts that have a strong community- and relationship-building role. The chi-square test of independence shows a relationship between the source and favorite variable with $X^2(1, N = 7503200) = 80215, p < .001$.

Moving to the comparison with the gender variable, we first inspect the dependence of the gender and the retweets variable. Male tweets are 38% more probable to be retweeted than female tweets. Calculating the chi-square test of independence shows a relationship between these two variables with $X^2(1, N = 7503200) = 11714, p < .001$.

By comparing the gender and favorited variables, we calculate that it is 13% more likely for a female tweet to be favorited than a male tweet. The chi-square test of independence shows a relationship between the gender and favorite variable with $X^2(1, N = 7503200) = 8913.4, p < .001$.

The presented results again suggest that male Twitter users behave more like corporate users and females are more aligned with the private Twitter accounts.

4. Language Standardness

The second part of statistical analyses inspects the linguistic characteristics of tweets posted by the different groups of users. Due to space constraints, we only present the results for language standardness scores assigned to each tweet in the corpus via a regression model (Ljubešić et al., 2015) while the behaviour of tweets according to the percentage of normalised tokens via CSMT (Ljubešić et al., 2016) that was also computed is consistent with the text standardness results.

We inspect the relationship of the account type and the user gender variable on one hand and the standardness continuous variable (ranging from 1 to 3) on the other. The resulting plot is presented in Figure 3. We can see that tweets posted by private and corporate users differ significantly regarding linguistic standardness, corporate users showing a much stronger tendency towards standard language, which is not surprising given their communicative goal. Male and female users are much more similar in this respect, but male users tend to produce more standard tweets, while female ones produce more semi- and non-standard ones, which is an interesting finding that deserves a closer examination in future work.

Given that the difference in text standardness by user gender presented in the right plot of Figure 3 is minor, we perform the chi-square test of independence showing a relationship of user gender and tweet standardness with

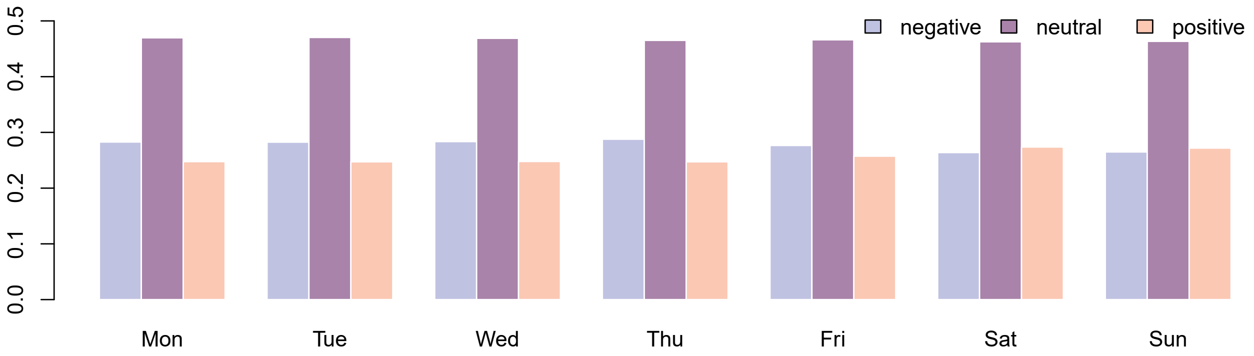


Figure 5: Distribution of the sentiment by day of week among private users.

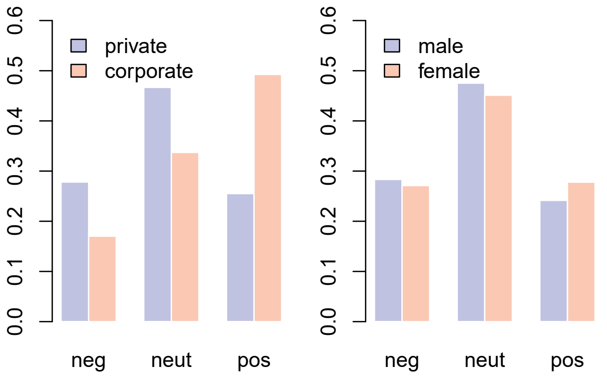


Figure 6: Distribution of the sentiment by source (left) and gender (right).

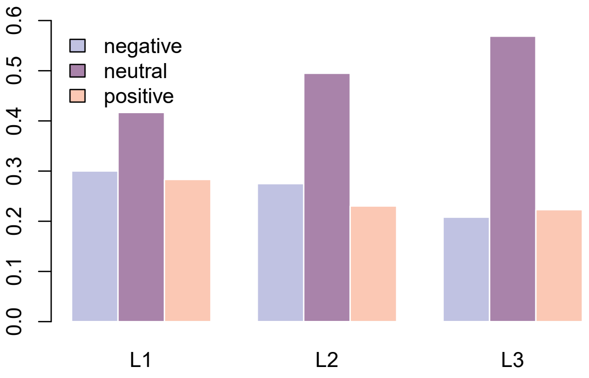


Figure 7: Distribution of the sentiment by language standardness (L1 - completely standard, L2 - slightly non-standard, L3 - very non-standard).

$X^2(1, N = 7503200) = 9740.9, p < .001$. For this test we operationalise the standardness variable as a binary variable, discarding tweets that are by the discrete standardness variable (three levels) slightly non-standard. While 24% of the remaining female tweets are estimated as being non-standard, for males the percentage is 19.6%.

Next, we plot the distribution of the discrete standardness variable (three levels) in the daily posting cycle in Figure 4. As expected, tweets are the most standard in the early morning hours (7 a.m.), which is probably an effect of corporate accounts of newspapers and other media posting links to new content for the day. As the day progresses, the proportion of slightly non-standard tweets rises steadily as does the proportion of very non-standard ones but they go up only slightly until late evening hours (after 11 p.m.) when they pick up and peak at around 3 a.m.

5. Sentiment Analysis

Finally, we look into the relationship of the account type and user gender variables with the sentiment score automatically assigned to each text in the Janes corpus using SVM (Fišer et al., 2016b). The three variables are compared in Figure 6. While corporate users post predominantly positive tweets and private users more neutral and negative ones, male users post slightly more negative posts and female users take the lead in the positive ones.

Again, given the close results on user gender, we perform the chi-square independence test of the user gender variable and the binary positive / negative text sentiment variable, discarding thereby neutral tweets. The test shows a relationship between the gender and the sentiment variables with $X^2(1, N = 7503200) = 6179.8, p < .001$.

In order to gain more insight into how the sentiment of Slovene tweet users varies throughout the week, we plotted the relationship of posting day and tweet sentiment in Figure 5. Disregarding the neutral tweets which prevail every day of the week, we can see that users start the week with a distinctly negative attitude which peaks on Thursday and then starts decreasing on Friday so that positive sentiment prevails during the weekend, peaking on Saturday.

The relationship between sentiment and standardness among private users is examined in Figure 7. Disregarding the neutral tweets that are prevalent across the board, positive sentiment prevails in very non-standard posts while the opposite is true at the other end of the spectrum. Our plan is to investigate this dependence in more detail in future work.

6. Conclusions

In this paper we carried out an analysis of a series of extralinguistic and linguistic variables in a large corpus of Slovene tweets. Among many of our findings, the most interesting ones are that there are big differences between tweeting behaviour, content and treatment of corporate and private tweets that are aligned with the primary communicative functions of the two types of Twitter users. Private male users tweet more than female users during weekdays while female users dominate on weekends. Male users tweet more in the morning hours while female users take the lead in the afternoon and evening. Male users use more standard language than female users, which is most frequently used in the early morning hours overall. Female users express more positive sentiment in their posts than their male counterparts, which is the prevalent sentiment overall while both tend to be more positive on weekends than during the week.

While the results are difficult to compare directly with the related work, the results obtained for the communication behaviour and styles of private and corporate users closely resemble the ones reported by Arakawa et al. (2014), Scheffler and Kyba (2016) and Volkova et al. (2013). The most striking difference between our results and related work is the language standardness level, which is higher in male users, contrary to what Bamman et al. (2012) have observed.

In the future we plan to extend our work with comprehensive statistical content and linguistic analyses. We also wish to compare the results with other text genres in the JANES corpus, such as blog posts, forum messages, news comments and Wikipedia talk pages. Finally, we envisage to compare the results with similar languages, such as Croatian and Serbian.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project ”Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017).

8. References

Arakawa, Y., Kameda, A., Aizawa, A., and Suzuki, T. (2014). Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets. *Journal of the Association for Information Science and Technology*, 65(7):1416–1423.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2012). Gender in Twitter: Styles, stances, and social networks. *CoRR*, abs/1210.4567.

Borges, G. R., Almeida, J. M., Pappa, G. L., et al. (2014). Inferring user social class in online social networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, page 10. ACM.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fišer, D., Erjavec, T., and Ljubešić, N. (2016a). The compilation, processing and analysis of the JANES corpus of Slovene user-generated content. *Slovenščina 2.0*, 4(2):67–100.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I., and Grčar, M. (2016b). Sentiment Annotation of the Janes Corpus of Slovene User-Generated Content. In *Proceedings of the 10th conference on language technologies and digital humanities*.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber’s Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’11, pages 237–246, New York, NY, USA. ACM.

Hu, T., Xiao, H., Luo, J., and vy Thi Nguyen, T. (2016). What the Language You Tweet Says About Your Occupation.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., and Škrjanec, I. (2015). Predicting the Level of Text Standardness in User-generated Content. In *Proceedings of Recent Advances in Natural Language Processing*, pages 371–378.

Ljubešić, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of KONVENS 2016*.

Mislove, A., Jørgensen, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J., (2011). *Understanding the Demographics of Twitter Users*, pages 554–557. AAAI Press.

Nguyen, D.-P., Gravel, R., Trieschnigg, R., and Meder, T. (2013). ” how old do you think i am?” a study of language and age in twitter.

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC ’10, pages 37–44, New York, NY, USA. ACM.

Rios, M. and Lin, J. (2013). Visualizing the ”pulse” of world cities on twitter.

Scheffler, T. and Kyba, C. C. (2016). Measuring social jetlag in twitter data. In *Tenth International AAAI Conference on Web and Social Media*.

Volkova, S., Wilson, T., and Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1815–1827.

A Textometrical Analysis of French Arts Workers “fr.*Intermittents*” on Twitter

Julien Longhi, Dalia Saigh
Cergy-Pontoise University, AGORA
E-mail: julien.longhi@u-cergy.fr, dalia-saigh@hotmail.com

Abstract

The term "social media" is increasingly used and tends to replace the term Web 2.0. Through social networks, people create various relationships. The aim of this paper is to describe how communities of users interact with each other on a specific subject, especially on Twitter. The theme that we will study is about the controversy concerning French arts workers (*fr.intermittents*. We will conduct a textometrical analysis using the software Iramuteq and then explain the statistical results.

Keywords: social media, Twitter, *intermittents*, textometrical analysis, Iramuteq

1. Introduction

The term "social media" is increasingly used and tends to replace the term Web 2.0. Through social networks, people interact and create various relationships. In their exchanges, they establish content, organize, modify, and combine it with personal creations. Despite authors’ freedom of expression and drafting, the content structure must obey rules of writing that are specific to each medium.

The aim of this paper is to analyze and describe how communities of users interact with each other on a specific subject. In our study, the theme is the controversy concerning French arts workers on Twitter: a microblogging service that is a hugely successful in spite of its particular working principle: blogging through ultra-short messages containing 140 characters. This feature allows the information flow faster but requires authors to be very concise when writing the tweet.

We will first describe the context and methodology for building our corpus. Then, we will introduce the method that we adopted for the textual analysis of this corpus entitled #intermittent (arts workers). We will also present Iramuteq, an analytical software tool that we have selected for this purpose and explain certain statistical results achieved.

2. Corpus Building: Background and Methodology

In March 2014, social partners signed a new agreement concerning the unemployment benefits for French arts workers. This text that became the convention of 14 May 2014 on unemployment benefits aroused concerns and opposition among the arts workers. A protest movement and mass demonstrations took place in Paris and in other French cities and lasted for several days.

These reactions rapidly invaded social networks especially Twitter. Millions of tweets were written as soon as the first information about this controversy emerged.

2.1 The Project Goal

The finalization of this corpus was made possible thanks to financial support from Ortolang[1]. The funding request centred around the finalization of the corpus-building process. The corpus is composed of tweets formed from the word hashtag (#) followed by the word “arts workers” then listed in a database of 13 074 tweets with #*intermittent*(s) and distributed in 4 617 twittos (Twitter users) over the period of June to September 2014, when tensions stepped up a notch and movements intensified.

Through the constitution of the corpus #*intermittent*, we hope to obtain a corpus which enables us to work on this kind of discourse (tweets related to a controversial topic), to characterize it and understand it under different forms in order to extend previous research (Longhi 2006, 2008) that focused on French arts workers in 2003/2004.

2.2 Data Building: the Choice of Data

After having contacted Twitter and having obtained confirmation that we had the right to collect and use information available on the site[2], we started tweet collection. This step was guided by the following process:

In 2014: retrieval of 13 074 tweets with #*intermittent* posted by 4 617 people.

In 2015: we established a threshold of at least 10 tweets with #*intermittent*: we obtained 215 accounts that had produced at least 10 tweets explicitly referenced as belonging to this theme (in order to have representative accounts). By collecting all the tweets from these 215 people, we gathered 586 239 tweets that included 10 876 tweets with #*intermittent*. The corpus #*intermittent* corresponds to these 10 876 tweets.

For the proper conduct of this process, we made, in collaboration with project participants from the field of Computing (Boris Borzic and AbdulhafizAlkhouli) a selection of data and metadata. For this, our colleagues developed a customized application. The application

1) uses the Twitter API: using ten functions of the API according to our needs, and recovering all the information in JSON format that we then convert;

2) allows the database to be enriched with a clean basic design (ten tables, fifty fields). Then we have programs that calculate indices for enriching additional fields;

3) allows customized export, with the information stored in a range of data formats. The challenge for a linguistic approach is to use this material to develop the #intermittent corpus.

These tweets were then formatted in TEI (with CMC formats extension tracks offered by a European group) to

[1] https://repository.ortolang.fr

[2] http://scinfolex.com/2009/06/14/twitter-et-le-droit-dauteur-vers-un-copyright-2-0/

become a corpus in order to meet the institutional requirements of the CoMeRe[3] project, and allow us to carry out a discourse analysis with word-processing tools on the corpus #*intermittent* or future corpora.

3. Textometrical Analysis of the Corpus #*Intermittent*

Textometry offers an instrumented approach to corpus analysis, articulating quantitative syntheses and analyzes including text (Lebart & Salem, 1994). Functionally, textometry implements differential principles. The approach highlights similarities and differences observed in the corpus according to the representation dimensions considered (lexical, grammatical, phonetic, or prosodic ones, etc). In addition to provide sorting procedures and statistical calculations for the study of digital corpora of texts, textometry establishes contextual and contrastive modeling. Thus, the text is characterized by its words in relation to their use in the corpus, the word is characterized by its co-occurrences, etc. (Pincemin, 2011).

Textometry is particularly relevant to corpus exploitation in human and social sciences. It simultaneously enables a detailed and global observation of different texts while remaining close to them, and highlights the fact that language is an important observation field for human and social sciences.

3.1 Iramuteq: the Text Analysis Tool

The Iramuteq software offers a set of analysis procedures for the description of a textual corpus. One of its principal methods is Alceste. This allows a user to segment a corpus into "context units", to make comparisons and groupings of the segmented corpus according to the lexemes contained within it, and then to seek "stable distributions" (Reinert, 1998). In addition to the Alceste method, Iramuteq provides other analysis tools including prototypical analysis, similarities analysis, and word clouds analysis. All of these methods allow the users of this tool to map out the dynamics of the discourses of the different subjects engaged in interaction (Reinert, 1999).

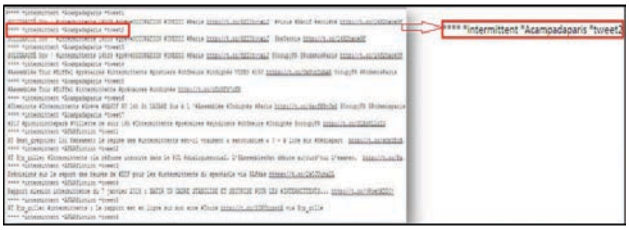
3.2 The Corpus Structure

Input files for *Iramuteq* must be in text format (.txt) and observe the following formatting rules:

The basic unit is called "text". A text can represent an interview, an article, a book or any other type of documents. A corpus may contain one or more texts (but at least one). The texts are introduced by four stars (****) followed by a series of starred variables separated by a space. It is possible to put the starred variables within the text by introducing the beginning of the line by a hyphen followed by a star (- *). This is known as "themes". The line should contain only this variable.

For our corpus format, we have chosen a format with three representative variables: we called the first "*intermittent*", because it constitutes the key word of this corpus. The second is about the "usernames", it's why this variable will change from a tweet to another depending on who posted the tweet. The third variable allows "the number" of tweets sent by a twittos to be counted as well as the re-tweets.

The figure below shows the formatting of the corpus #*intermittent*:



```
**** *intermittent *Acampadaparis *tweet2
```

Figure 1: The format of the corpus #*intermittent*.

4. Methods and Results of the Analysis

4.1 The Word-Cloud

Iramuteq contains an option that makes a kind of a lexical compendium of a document in which the discussed key concepts are represented by a size unit (in the sense of the used typography weight). This allows their importance within the corpus to be highlighted. Specifically, the more a keyword is quoted in an article, the bigger it will appear in the cloud of words. This technique will allow us to put forward the keywords used by twittos.

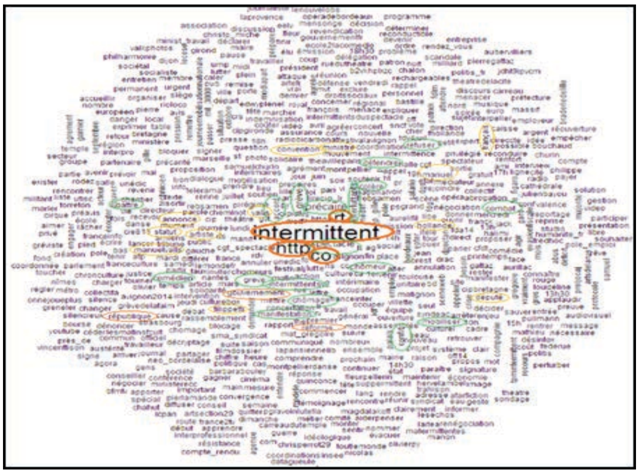


Figure 2: The word-cloud of the corpus #*intermittent*.

This word-cloud highlights the most common occurrences in tweets. These lexical items are positioned centrally in the cloud. The occurrence "*intermittent*" is the largest in size because it constitutes the key word of our corpus; this is why its frequency is higher. That word is followed by specific markers such as "co" and "http" that refer to links shared on Twitter. Indeed, these links are automatically abbreviated http: // co to allow long URLs to be shared without exceeding the maximum number of characters allowed when writing a tweet. There is also the sign "rt" which means "retweet". This has the function of reposting the tweet of another person enabling users to quickly share it with all subscribers.

[3] http://corpuscomere.wordpress.com

Around those keywords are others which have more or less the same frequency and thus appear the same size. Among them, those that refer to the semantic field of the republic and the French government such as *Manuel Valls*, *Republique* (republic), *député* (deputy), *français* (French), *F.Hollande*, *Fillipetit, minister* (minister). Other lexical items evoke either movements or activities such as *accord* (agreement), *grève* (strike), *mobilisation* (mobilization), *manifestation* (protest), *convention* (convention), *combat* (fight). There are also names or adjectives referring to French arts workers, and describing their situation as *chômeurs* (unemployed workers), *précaires* (precarious), *interluttents*, *comédiens* (actors).

Despite the interest of this method, the resulting description remains very general. For a more detailed analysis, *Iramuteq* offers another graphical representation of a corpus' words, a significant method called "similarities analysis", which retains the idea of size proportional to the frequency, but introduces the relations of co-occurrences between words.

4.2 Similarities Analysis

Similarities analysis is a technique based on graph theory (Flament, 1962). It presents in a graphical format the structure of a corpus, distinguishing between the shared parts and the specificities of coded variables. This allows the link between the different forms in the text segments to emerge (Marchand & Ratinaud, 2012).

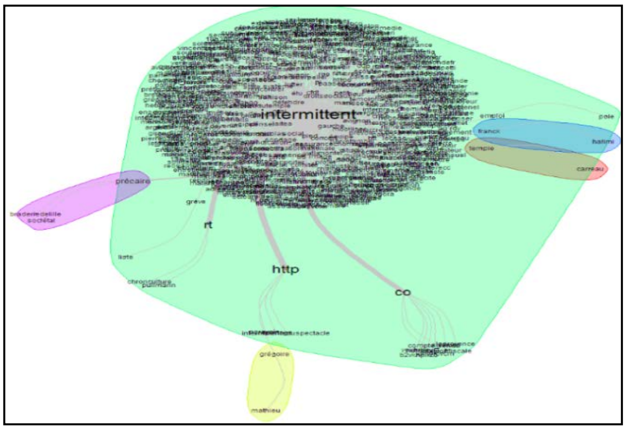


Figure 3: The similarities analysis of the corpus *#intermittent*.

The first observation that we can make is that this corpus is very homogeneous with one central idea around which revolves the greatest part of the lexicon of our corpus. This figure shows a single main cluster, with some others which are very small and not relevant. This cluster consists of a word cloud which contains the key word "*intermittent*" at its center and around it, are grouped a very dense and related lexicon.

We notice the presence of some small groups, which are in the main cluster, directly related (with edges) to "*intermittent*", the most important one. Among these groups, there is: "*http"* in which we find the term *intermittentdespectacle* (arts workers) and a little further, a small cluster containing the name *Gregory Mathieu*, a sociologist who wrote a book with the title "*Les intermittents du spectacle. Enjeux d'un siècle de luttes"*. So, in this group, we understand that the majority of links mentioned in tweets refer users to web pages where the name of the sociologist is mentioned.

There is also the "*rt"* group which includes the following terms: *chronculture, pullmarin, dinamopress, angelin...* which refer to the names of people who have retweeted the most. The "co" group is, as explained above, the abbreviated form of links on Twitter.

We can already understand from this figure that the *#intermittent* corpus contains a lot of links, retweets related to French arts workers, and it describes their various actions and their status (highlighted by the cluster *précaires* (precarious).

That being said, as the lexicon related to the keyword "*intermittent*" is very dense, the function similarities analysis has simply helped us to describe the nature and the main topic of tweets (tweets with links, retweets, arts workers status ...). To further clarify the corpus structure, we will use the HDC "Hierarchical Descending Classification" function (a method established by Max Reinert).

4.3 The Hierarchical Descending Classification

One method used by Alceste is the hierarchical descending classification. This method offers a global approach to a corpus. The *HDC* after partitioning the corpus, identifies statistically independent word classes (forms). These classes are interpreted through their profiles, which are characterized by specific correlated forms. The *HDC* shows that using a dendrogram.

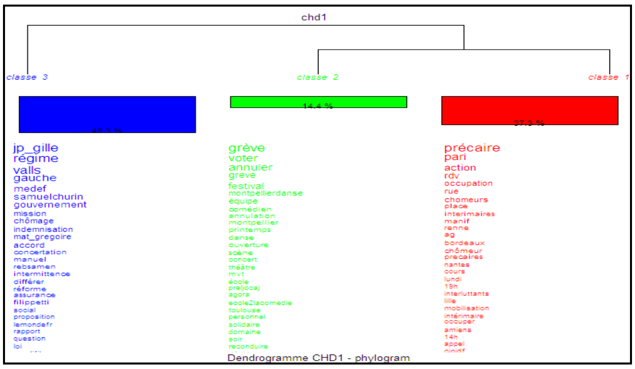
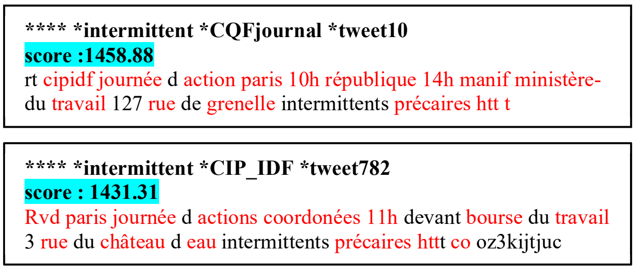


Figure 4: The result of the Hierarchical Descending Classification.

Two groups are distinguished in this figure, the first with two related classes (class 1 and class 2), and the second where there is only one class (class 3).

The class 1 includes forms associated with the different protest movements of French arts workers such as the occupation of streets, theaters and other places, the demonstrations in Paris and elsewhere.

Here is an extract of characteristic segments (with a high score), which contain the most common words associated with class 1 like *manif* (event), *cipdfjournée* (cip-idf day), *action* (action), the common words are highlighted in red:

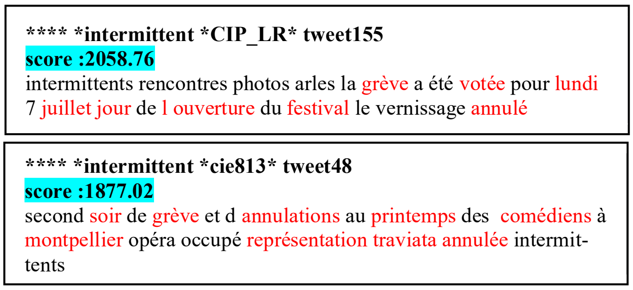


**** *intermittent *CQFjournal *tweet10
score :1458.88
rt cipidf journée d action paris 10h république 14h manif ministère-du travail 127 rue de grenelle intermittents précaires htt t

**** *intermittent *CIP_IDF *tweet782
score : 1431.31
Rvd paris journée d actions coordonées 11h devant bourse du travail 3 rue du château d eau intermittents précaires httt co oz3kijtjuc

Figure 5: The characteristics segments of class 1.

Class 2 refers to strikes held by the French arts workers and their different concerts and show cancellations. This class contains words such as: *grève* (strike), *festival* (festival), *annulé* (cancelled). The following figure shows the characteristics segments of this class:

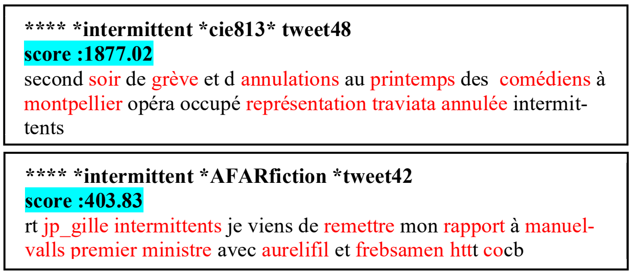


**** *intermittent *CIP_LR* tweet155
score :2058.76
intermittents rencontres photos arles la grève a été votée pour lundi 7 juillet jour de l ouverture du festival le vernissage annulé

**** *intermittent *cie813* tweet48
score :1877.02
second soir de grève et d annulations au printemps des comédiens à montpellier opéra occupé représentation traviata annulée intermittents

Figure 6: The characteristics segments of class 2.

Class 3 concerns the tweets that talk about the unemployment insurance system related to the French arts workers and political entities involved in this affair. Here is a characteristics segment summarizing the words associated with this class, including *medef*, *valls*, *samuelchurin*, *aurelifil*:



**** *intermittent *cie813* tweet48
score :1877.02
second soir de grève et d annulations au printemps des comédiens à montpellier opéra occupé représentation traviata annulée intermittents

**** *intermittent *AFARfiction *tweet42
score :403.83
rt jp_gille intermittents je viens de remettre mon rapport à manuel-valls premier ministre avec aurelifil et frebsamen httt cocb

Figure 7: The characteristics segments of class 3.

These results demonstrate that unlike the written press which showed a plurality of views concerning the semantic representation of the word "*intermittent*" (see Longhi, 2006) which was seen whether as a status (*statut*), a profession (*métier*) or in the dynamics of these two semantic components. Here, the word "*intermittent*" is presented using three different senses "system" (*régime*), "status" (*statut*) and "fight" (*lutte*). This indicates that Twitter focuses on the status side and declines it by introducing the French arts workers insurance system (one way of looking at the status) or the consequence of this status (fight).

5. Conclusion

A Textometrical analysis of this corpus has allowed us to see how twittos have reacted to the announcement of the new unemployment insurance system related to French arts workers. Through the analysis of similarities, we have found that there were a lot of links pointing to this topic with references to the sociologist Mathieu Grégoire and his various texts, and also newspaper names and publications including *Le Monde*. There were also various retweets and thanks to this, the issue has become in a short time a "trending topic" on Twitter. This is due to the various markers such as #, URLs, the @ sign ... The Reneirt method (HDC) taught us that discourse around this subject is divided into two different sets. On the one hand, tweets that describe the precariousness of French arts workers and their various protest movements against the new regime. On the other hand, tweets denouncing the impartiality of the agreement, with links providing information about that act and citing various political personalities who were involved in the controversy.

6. References

Flament, C. (1962). L'analyse de similitude. *Cahiers du centre de recherche opérationnelle*, 4, pp. 63--97

Lebart, L., Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Longhi, J. (2006). De intermittent du spectacle à intermittent: de la représentation à la nomination d'un objet du discours. *Corela*, 4 (2). URL : http://corela.revues.org/457.

Longhi, J. (2008). Sens communs et dynamiques sémantiques : l'objet discursif intermittent. *Langages*, 170, pp. 109--124.

Marchand, P., Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles. JADT, 2012*, pp. 687--699.

Pincemin, B. (2011). Sémantique interprétative et textométrie. *Corpus*, 10. URL: http://corpus.revues.org/2121.

Reinert, M. (1998). Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste. *Actes des 4èmes journées Internationales d'Analyse Statistiques des Données textuelles*. URL : http://lexicometrica.univ-paris3.fr/jadt/jadt1998/reinert.htm.

Reinert, M. (1999). Quelques interrogations à propos de l'objet d'une analyse de discours de type statistique et de la réponse « Alceste ». *Langage et société*, 90 (1), pp. 57--70.

Corpus CoMeRe: https://corpuscomere.wordpress.com

Iramuteq: www.iramuteq.org

Ortolang: https://www.ortolang.fr/market/home

The Use of Alphanumeric Symbols in Slovene Tweets

Dafne Marko
Faculty of Arts, University of Ljubljana
E-mail: dafne.marko@gmail.com

Abstract

This paper deals with the use of alphanumeric symbols in Slovene tweets. We use the JANES corpus, a large corpus of internet Slovene containing tweets, forum and blog posts, and comments on news articles and on Wikipedia discussion and user pages. We analyze the use of words consisting of alphabetic and numeric symbols as a means of both creative writing and as a word-shortening strategy. We investigate which alphanumeric features are most frequently used in Slovene tweets and identify the numerals substituting the letters. The results are compared with other subcorpora in the JANES corpus as well as with the Kres corpus, a collection of standard Slovene texts. Furthermore, we compare the distribution of alphanumeric features according to user type and text standardness.

Keywords: alphanumeric symbols, letter/number homophones, Slovene tweets, computer-mediated communication

1. Goal of the Paper

The main goal of the paper is to research the occurrence of a CMC-specific linguistic feature – words consisting of alphabetic and numeric characters. We focus on the subcorpus of Slovene tweets, but also compare the distributions with other subcorpora in the JANES corpus (forum posts, blog entries, comments on news articles, Wiki talk) as well as with the Kres corpus, a corpus of standard Slovene with a balanced genre structure. We predict to find no or very few occurrences in the Kres corpus, proving that it is, indeed, a CMC-specific linguistic feature. We also predict the Twitter subcorpus to be most abundant with this sort of writing, whereas no significant difference between other subcorpora is expected. We research whether gender (male vs. female) or user type (corporate vs. private) influences the use of alphanumeric characters in words. Furthermore, our goal is to carry out a detailed analysis of the most frequently used words with alphanumeric symbols in Slovene tweets. We try to investigate which numeric symbols (numerals) are used to substitute the letters and whether they are used phonetically (e.g., *ju3*, translated as *2morrow*, with a number 3 pronounced as /tri/, the same as in the word *jutri*) or graphically (e.g., *g33k*, where the number 3 represents the letter "e"). With our analysis, we present a linguistic phenomenon which could be described as a type of creativity in writing, and – according to the fact that the length of a single tweet is limited to 140 characters – also as a word-shortening strategy.

2. Related Work

So far, little research has been done on the use of alphabetic and numeric characters in tweets. Most researchers deal with text messaging or *texting* as a relatively new writing medium. It has been pointed out that "/t/here are a great deal of apparent similarities between Twitter and text messaging" since "they are both a medium via which friends and acquaintances can communicate with one another, and they both fall under the broad banner of technology mediated communication" (Denby, 2010). Another shared characteristic is character limitation – Twitter imposes an explicit message length limit of 140 characters, and in text messaging, the limitation is 160 characters. Although there are several differences between the aforementioned writing media (public vs. private, cost of specific service, device used for text messaging and Twitter posts, etc.), research on texting could help us get a deeper insight into the characteristics of another CMC phenomenon – specific linguistic features in Twitter posts.

Most of the researchers claim that "shortenings are presented to be the one major characteristic of text messaging that is assumed to be technologically determined by the limited number of permitted characters and the cumbersome input via the small cellular phone keypad" (Bieswanger, 2006). Language used in texting, or what Crystal (2001) refers to as *Netspeak*, is assumed to be "heavily abbreviated" (Thurlow, 2003), although Thurlow reports "relatively few (n = 73) examples of language play using letter-number homophones (e.g. Gr8 'great', RU 'are you'), which, in popular representations at least, have become the most definitive feature of text-messaging".

Some authors claim that Twitter posts, which fall into the category of *microblogs* (Moseley, 2013) or *microtexts* (Gouws et al., 2011), are rich with abbreviations "solely to conserve space within a text" (Alkawas, 2011), when others observe that "SMS language seems to have evolved into a fashionable and stylish way of writing where the way of writing is as important as the content" (Kirsten Torrado, 2014).

The linguistic phenomenon discussed in this paper is often referred to as *letter/number homophones* (comp. Bieswanger, 2006; Kirsten Torrado, 2014; Frehner, 2008; Kadir et al., 2012; Elizondo, 2011; Farina and Lyddy, 2011; Thurlow, 2003; Kul, 2007; Alkawas, 2011) or (*alphanumeric*) *rebus writing* (Halmetoja, 2013; Danet and Herring, 2007), but we can also find wider, more generic expressions, e.g., *complex abbreviations*

(Filipan-Žignić et al., 2012) or *textism* (Grace et al., 2012; Bushnell et al., 2011). Crystal (2001) refers to this phenomenon as a "rebus-like potential" of letters and numbers "whose pronunciation is identical with words or parts of words" and "are used to replace words or letter sequences". Frehner (2008) differentiates between *letter homophones* – the use of a single letter whose phonological content is equated with a word, e.g., "u" for "you", *number homophones* – the use of a numeral whose phonological content is equated with a word, e.g., "4" for "for", and a combination of letters and numerals, forming *letter/number homophones*, e.g., "b4" for "before".

Such shortenings can also be observed in Slovene. Michelizza (2008) talks about a special group of abbreviations, "typical for the language of SMS, where parts of a word are substituted by a mathematical symbol or numeral, pronounced the same or at least similar to the part of the word it substitutes (e.g., *ju3*)"[1]. Dobrovoljc (2008) lists the most frequently used letter/number homophones in Slovene (ju3 = "jutri", pr8 = "prosim", 5er = "Peter", 1x ="enkrat") and English (4yeo = "for your eyes only", j4f = "just for fun", 2 much = "*too much*"), but concludes that there is a low percentage of such writings in Slovene texting language. Logar (2006) names this kind of linguistic feature a "combination of various writing symbols", which are commonly observed in Slovene SMS. In her research, she showed that "/a/mong the more than 450 examples of SMS abbreviations that had been submitted to the site by 11 January 2002, more than 60% were some type of abbreviation, while the rest of the material (160 examples) was made of, for example, the following: :-) 'zadovoljen', :) 'veselje', :(... 'jočem', :x 'poljubček', :D 'širok nasmešek', mi2 'midva', ju3 'jutri', 2mač 'preveč', sk8ar 'skejtar', 8-) 'Nosim očala', <>< 'ribica', {*} 'objemček, poljubček', *+* 'vidim te', @x@ 'maš mačka?', @->-- 'vrtnica', _/0 'A greš na kavo?', =:x 'zajček'." Since most of the researches mentioned above focused only on the use of alphanumeric symbols in texting, we will investigate how often they occur in Slovene tweets.

3. Dataset and Methodology

For our research, we used the JANES v0.4 corpus[2], a large corpus of Slovene tweets, forum posts, blog texts, comments on news articles and on Wikipedia pages and users, which contains over 175 million words or 9 million documents, published between 2002 and 2016 (Fišer et al., 2016). We focused on the biggest subcorpus, the Slovene tweets, which consists of 90.180.337 words from 7.503.199 different Twitter posts.

Using the concordancer SketchEngine[3], we searched for all occurrences of words consisting of alphanumeric symbols, where the numerals appear at the beginning, in the middle, or at the end of the word. To achieve that, appropriate regular expressions were used for querying the corpus: [word="[0-9]+[a-zA-ZščžŠČŽ]+"], [word="[a-zA-ZščžŠČŽ]+[0-9]+[a-zA-ZščžŠČŽ]+"], and [word="[a-zA-ZščžŠČŽ]+[0-9]+"]. Prior to further analysis, irrelevant results had to be manually selected and excluded from the list. This was done because there are numerous examples of words which consist of alphanumeric symbols but represent a proper name/part of a proper name, a chemical symbol, a unit of measurement, or some other abbreviation (e.g., *A4*, *CO2*, *C4*, *TEŠ6*, *m2*, etc.). With these examples, no transformation from numerals to letters can be made in the written form (e.g. A4 → *A-štiri, CO2 → *CO-dva, etc.). After a quick overview of the concordances, we created a frequency list of all the words with alphabetic and numeric symbols which appear in Slovene tweets.

To investigate which users (corporate or private; male or female) incorporate such shortenings into their tweets, we used the appropriate filters and compared the results. We also compared the frequency of letter/number homophones in tweets according to their technical and linguistic standardness[4].

Furthermore, the distributions of letter/number homophones in all JANES subcorpora were compared to that of the Kres corpus[5], a collection of standard written Slovene with a balanced genre structure (Logar Berginc et al., 2012).

In the second part of our study, the letter/number homophones found in Slovene tweets were analyzed in more detail. Among the most frequently used numerals in the shortenings discussed, we investigated:

- where they appear (at the beginning, in the middle, or at the end of a word);
- what they substitute (a string of letters, a single letter);
- whether they are used phonetically or graphically (*2morrow* vs. *g33k*).

4. Results

Surprisingly, the first query returned no results, which means that there are no words with numerals at the beginning of a word appearing in Slovene tweets represented in the JANES corpus. Thus, we used only

[1] The paragraph was translated into English by the author.

[2] Description available at http://nl.ijs.si/janes/ (access: June 12, 2016).

[3] Available at https://www.sketchengine.co.uk/ (access: June 10, 2016.

[4] All texts in the JANES corpus are annotated according to their level of standardness. "The score for technical text standardness focuses on word capitalisation, the use of punctuation, and the presence of typos or repeated characters in the words. The score for linguistic standardness, on the other hand, takes into account the knowledge of the language by the authors and their more or less conscious decisions to use non-standard language, involving spelling, lexis, morphology, and word order" (Ljubešič et al., 2015. Tweets are annotated "using a score between 1 (standard and 3 (very non-standard, with 2 marking slightly non-standard texts" (Ljubešič et al., 2015).

[5] Description available at http://www.slovenscina.eu/korpusi/kres (access: August 20, 2016).

the remaining two regular expressions for our further research.

4.1 Numeral at the End of the Word

The total number of concordances for all tokens ending in a numeral is 58.794. However, as mentioned before, words which represent a part of a proper name, a chemical symbol, a unit of measurement, etc., had to be manually selected and excluded from the list to get the actual result. The remaining tokens and their absolute frequencies are represented in Table 1.

Token	Frequency
ju3	1173
Mi2	593
mi2	371
Ju3	337
s5[6]	292
MI2	119
vi2	110
hi5	97
tr00	77
zju3	50
Hi5	47
na1	36
gr8	36
Tr00	31
Mi3	31
me2	27
str8	26
Vi2	20
Gr8	17
Me2	11
h8	11
u3	10
sk8	7
Zju3	5
mi3	5
H8	3
TR00	1

Table 1: Words with numerals at the end of the word.

As seen in the table above, 27 different tokens with 15 different lemmas[7] were found in the corpus of Slovene tweets, altogether representing a relative frequency of 33.1 per million tokens. Nine out of 27 tokens, which are written in bold, represent alternative written forms of different personal pronouns. The relatively high frequency can be explained by the fact that the numeral in the personal pronoun does not only have the same phonetic content as the letters (e.g., *s5* → /spet/, where 5 is pronounced as /pet/), but emphasizes the number of people a specific personal pronoun is denoting (*mi2* → two people; *mi3* → 3 people, etc.).

It is also interesting that 11 tokens are actually English homophones, frequently used in Slovene tweets as well.

4.2 Numeral in the Middle of the Word

Interestingly, the list of different letter/number homophones with numerals appearing in the middle of the word is significantly longer, whereas the relative frequency in much lower (9.97 per million tokens). This kind of shortening technique proves to be very productive, but still appears less frequently than the one with numerals at the end of the word. After excluding all the proper names and other irrelevant words, 117 tokens with approximately 50 different lemmas were found in Slovene tweets.

In some cases, it was difficult to identify whether we were dealing with a typographical error or whether the word was intentionally written in that form. We used the proximity of the letters and numbers on the keyboard to identify and exclude possible typographical errors (e.g., *v0lilcev* → number 0 and the letter "o" are very close on the keyboard, so we identified it as a typographical error vs. v8dja → probably intentionally written as such). As expected, the majority of homophones represent English words, where the preposition "to" is typically substituted by number 2 (e.g., *B2B*, *p2p*, *coffee2go*, *up2date*, etc.). There are, however, also numerous Slovenian words written with both alphabetic and numeric symbols. It is important to emphasize that we did not exclude the homophones which represent a phrase consisting of two or more words, e.g., *mi3je* = mi trije; *še1x* = še enkrat, etc. We decided not to consider them as typographical errors, but as a decision of Twitter users to write them as one word, similar to the English multi-word phrases listed above (*coffee2go*, *up2date*, etc.).

Token	Frequency
B2B/b2b	205/41
w00t/W00t	66/39
d00h/d0h/D0h/d000h	51/48/26/4
pr0n/Pr0n	49/6
g33k/g33ki/g33kov/ g33ka/G33k	35/9/6/5/4
na1x	30
n00b/n00be	24/4
B2C	21
s3ksi/S3ksi	19/4
p2p/P2P	19/18
B4B	19
p0rn[8]	18

[6] Token *S5* was excluded from the list because it was almost exclusively used in the proper name *Galaxy S5*.

[7] Since the texts in the JANES corpus are normalized, the tokens *Ju3* and *ju3* would both have the same lemma – *jutri*.

[8] The motivation for writing specific words with numerals

Za1x	13
mi3je	12
še1x	11
ju3šnji/ ju3snji/ Ju3šnji/ ju3šnjega/ ju3snjem	11/4/4/3/3

Table 2: Most frequently used words with numerals in the middle of the word.

4.3 The Use of Alphanumeric Symbols According to User Gender

The comparison of frequency of letter/number homophones according to user gender shows that the use of words with alphanumeric symbols is far more frequent among male users (roughly 80% of all the occurrences were written by male users). However, we must take into account that the comparison was made using all the occurrences returned by the regular expressions, since we could not filter out all irrelevant examples.

4.4 The Use of Alphanumeric Symbols According to User Type

The comparison according to user type (corporate vs. private) shows a strong tendency of private users to incorporate such writing into their tweets. From 65.042 occurrences, 45.682 (≈ 70%) were written by private users.

4.5 The Use of Alphanumeric Symbols According to the Level of Text Standardness

Regarding the level of text standardness, an assumption can be made that less standard tweets (from both linguistic and technical perspective) contain more letter/number homophones than those written according to grammatical and orthographic rules. We compared all 9 possibilities of text standardness available in the JANES corpus (from L1T1 = linguistic 1, technical 1 to L3T3 = linguistic 3, technical 3 with all median possibilities). Words with alphabetic and numeric symbols are most frequently used in tweets annotated as very non-standard (L1T1) or linguistically very non-standard and technically slightly non-standard (L1T2).

4.6 Comparison with the Kres Corpus

In the figure bellow, relative frequencies of all letter/number homophones (regardless of the position of the numeral) in the JANES subcorpora and the Kres corpus are presented. As evident from the chart, letter/number homophones are far most frequent in Twitter posts (43.07 per million), followed by the forum posts (18.19 per million). Blog entries, comments on news, and wiki talk show a fairly similar distribution (2.61, 3.37, and 2.94 per million, respectively). Surprisingly, letter/number homophones can also be found in the Kres corpus (1.36 per million). A total of 12 different examples were found, 10 of them with numerals in the middle of the word (e.g., *l33t*, *cig4ni*, *za1x*, *pr0n*), one with numerals ending a word (*ju3*), and also one with numerals starting a word (*4ever*). However, all of these examples were found in the texts obtained from the web pages and from the Slovenian magazine *Joker*, which is primarily a computer gaming magazine with a distinctive writing style.

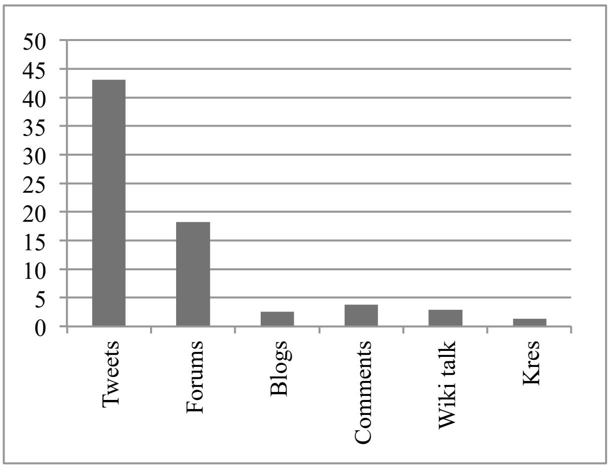


Figure 1: Relative frequencies of letter/number homophones in the JANES subcorpora and the Kres corpus.

5. Qualitative Analysis of Extracted Letter/Number Homophones

In this section, a qualitative analysis of the most frequently used letter/number homophones is presented, along with interpretations of the numeric features and their functions.

5.1 Most Frequent Numerals Used in Letter/Number Homophones

If we analyze the extracted words with alphanumeric symbols in more detail, it is evident that numerals are used only in the middle of the word (117 tokens) or at the end of a word (27 tokens). As mentioned before, no tokens with numerals starting a word were found in the corpus. Numerals used in letter/number homophones are definitely not randomly picked by the users. Each numeral has a specific meaning or interpretation and substitutes either a single letter or a string of letters in a word. In our corpus, 9 numerals were identified in letter/number homophones, i.e. 0, 1, 2, 3, 4, 5, 7, and 8. Among them, numerals 2, 3, 8, and 0 are the most frequent ones. Since the same numeral can appear in different words and even have different functions, it is interesting to identify which letter/letters are substituted

instead of letters could be to avoid censorship carried out by the moderators of forums, blogs, or comment sections. The same pattern appears in the word *cig4n* (= cigan) found in the Kres corpus.

by them. In Table 3, all numerals and their interpretations are presented, together with the examples from the corpus.

Numeral	Interpretation	Example
1	“ena” “i”	*na1* = “na ena” *BRA71L* = “Brazil”
2	”dva” “dve” “to”	*mi2* = “midva” *me2* = “medve” *up2date* = “up to date”
3	“tri” “e”	*ju3 = “jutri”* *s3njam = “strinjam”* *g33k = “geek”*
4	“for” “a”	*t4t = “training for trainers”* *G4ME = “game”*
5	“pet” “five”	*s5 = “spet”* *hi5 = “high five”*
7	“z”	*BRA71L = “Brazil”*
8	“eat” “aight” “ate”	*gr8 = “great”* *str8 = “straight”* *h8 = “hate”* *l8r = “later”*
0	“o”	*n00b = “noob”* *p0rn = “porn”* *w00p = “woop”*

Table 3: Numerals used in letter/number homophones with their interpretations and examples.

5.2 Phonetic vs. Graphic Function of Numerals

As evident from the table above, there is a striking difference between numerals that are used phonetically (e.g. *s5* = “spet”, where their pronunciation is identical with a part of the word, enabling them to replace the letter sequence) and graphically (e.g. G4ME = “game”, where the numeral 4 has a similar visual appearance as the letter “A”). All numerals used at the end of the words (see Table 1) are used phonetically; the only exception is the word *tr00* (= “true”). This example is especially interesting because numerals do not only substitute a string of letters (*00* = “oo”), but the pronunciation of the substituted letters is similar or the same as the original pronunciation of the word string (“ue” in *true*). In other words, 3 transformations are needed to identify the “original” word, namely *tr00* → troo → /tru:/ → “true”. Numerals which appear in the middle of the word can be used either graphically or phonetically. Numerals which substitute letters based on their appearance rather than their pronunciation tend to be duplicated (e.g. *g33k*, *w00p*, *n00b*, etc.)[9].

According to that, we can undoubtedly claim that letter/number homophones are not only used as a word-shortening strategy, but as a form of creative writing and a specific stylistic feature as well. Furthermore, graphically used numerals also prove that CMC is “essentially a mixed modality” which “resembles speech” but “looks like writing” (Baron, 2008), since the words, such as *g33k*, *tr00*, or *n00b*, have little significance if not visually represented.

6. Conclusion

This paper presents the use of so-called letter/number homophones in Slovene tweets as presented in the JANES corpus. The results show that a considerable amount of alphabetic and numeric symbols are used both in Slovene and in English words. This phenomenon, also described as a type of “neography” (Danet and Herring, 2007), proved to be characteristic for CMC, especially microtexts, such as Twitter and forum posts, since no letter/number homophones were found in the Kres corpus apart from the texts obtained from web pages. As expected, Twitter proved to be the richest subcorpus regarding this phenomenon, followed by the forum subcorpus with a relatively high frequency of the shortenings discussed. Numerals used in the middle or at the end of specific words substitute letters or strings of letters and make the texts either shorter or more interesting to read. Since a certain numeral can be used both graphically (*g33k*) and phonetically (*u3nek*), a creative writing style has emerged among new generations, which definitely deserves linguistic attention. For a more precise description of the phenomenon, a more detailed comparison with other CMC media (SMS, blogs, forums, etc.) would be necessary. Apart from that, an analysis of usernames would be very useful for investigating language creativity as observed in computer-mediated communication.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017).

8. References

Alkawas, S. (2011). *Textisms: The Pragmatic Evolution among Students in Lebanon and its Effect on English Essay Writing*. Master Thesis, Lebanese American University.

Baron, S. (2008). *Always On: Language in an Online and Mobile World*. Oxford University Press, Oxford.

Bieswanger, M. (2006). 2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space-and time-saving strategies in English and German text messages. In Hallett, T., Floyd, S., Oshima, S. and Shield, A. (Eds.), *Texas Linguistics Forum Vol. 50*, Austin.

[9] This type of writing is also refered to as “l33t speak” or “l33t”, used mostly by players of video games “where numbers and symbol combinations are used to represent letters” (Sherblom-Woodward, 2002).

http://studentorgs.utexas.edu/salsa/proceedings/2006/Bieswanger.pdf

Bushnell, C., Kemp, N. and Heritage Martin, F. (2011. Text-messaging practices and links to general spelling skills: A study of Australian children. In *Australian Journal of Educational & Developmental Psychology*. Vol 11, pp. 27–38.

Crystal, D. (2001). *Language and the Internet*. Cambridge, Cambridge University Press.

Danet, B. and Herring, S. (Eds.). (2007). *The Multilingual Internet. Language, Culture, and Communication Online.* Oxford University Press, Oxford.

Denby, L. (2010). *The Language of Twitter: Linguistic Innovation and Character Limitation in Short Messaging.* Undergraduate dissertation, University of Leeds.

Dobrovoljc, H. (2008). Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. In Košuta, M. (Ed.), *Slovenščina med kulturami*, Slavistično društvo Slovenije, Celovec, pp. 295–314.

Elizondo, J. (2011). *Not 2 Cryptic 2 DCode: Paralinguistic Restitution, Deletion, and Non-standard Orthography in Text Messages*. Ph.D. thesis, Swarthmore College.

Farina, F. and Lyddy, F. (2011). The Language of Text Messaging: "Linguistic Ruin" or Recource? In *The Irish Psychologist*, Vol. 37, Issue 6, pp. 145–149.

Filipan-Žignić, B., Velički, D. and Sobo, K. (2012). SMS communication – Croatian SMS language features as compared with those in German and English Speaking Countries. In *Revija za elementarno izobraževanje*, št. 1. Pedagoška fakulteta, Maribor.

Fišer, D., Erjavec, T. and Ljubešić, N. (2016). Janes v0.4: korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* (to appear).

Frehner, C. (2008). *Email, SMS, MMS: The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Peter Lang.

Gouws, S., Metzler, D., Cai, C. and Hovy, C. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the workshop on language in social media* (*LSM 2011*), pp. 20–29. http://aclweb.org/anthology/W/W11/W11-0704.pdf.

Grace, A., Kemp, N., Martin, F. H. and Parrila, R. (2012). Undergraduates' use of text messaging language: Effects of country and collection method. In *Writing Systems Research.* Taylor & Francis Online.

Halmetoja, T. (2013). *Gender-Reated Variation in CMC Language: A Study of Three Linguistic Features on Twitter*. BA thesis, Göteborgs Universitet.

Kadir, Z. A., Maros, M. and Hamid, B. A. (2012). Linguistic Features in the Online Discussion Forums. In *International Journal of Social Science and Humanity*, Vol. 2, No. 3, May 2012, pp. 276–281.

Kirsten Torrado, U. (2014). Development of SMS language from 2000 to 2010. In Cougnon, L. and Fairon, C. (Eds.), *SMS communication: A linguistic approach*. Benjamins Current Topics.

Kul, M. (2007). Phonology in text messages. In *Poznań Studies in Contemporary Linguistics 43(2)*, pp. 43–57.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *Proceedings*, pp. 371–378, Hissar: [s.n.]. http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. and Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Logar, N. and Smith, J. (trans.). (2006). Stilno zaznamovane nove tvorjenke: tipologija = Stylistically marked new derivates: a typology. In Vidovič-Muha, A. (Ed.), *Slovensko jezikoslovje danes*, Slavistično društvo Slovenije, Ljubljana, pp. 87–101.

Michelizza, M. (2008). Jezik SMS-jev in SMS-komunikacija. In *Jezikoslovni zapiski: zbornik Inštituta za slovenski jezik Frana Ramovša*, Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana, pp. 151–166.

Moseley, N. (2013). *Using word and phrase abbreviation patterns to extract age from Twitter microtexts*. Thesis, Rochester Institute of Technology.

Sherblom-Woodward, B. (2002). *Hackers, Gamers and Lamers: The Use of l33t in the Computer Sub-Culture.* http://www.swarthmore.edu/SocSci/Linguistics/papers /2003/sherblom - woodward.pdf.

Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. In *Discourse Analysis Online*, Sheffield.

A Multilingual Social Media Linguistic Corpus

Luis Rei*,†, Dunja Mladenić*,†, Simon Krek*
*Jožef Stefan Institute, †Jožef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: luis.rei@ijs.si, dunja.mladenic@ijs.si, simon.krek@ijs.si

Abstract

This paper focuses on multilingual social media and introduces the xLiMe Twitter Corpus that contains messages in German, Italian and Spanish manually annotated with Part-of-Speech, Named Entities, and Message-level sentiment polarity. In total, the corpus contains almost 20K annotated messages and 350K tokens. The corpus is distributed in language specific files in the tab-separated values format. It also includes scripts that enable to convert sequence tagging tasks to a format similar to the CONLL format. Tokenization and pre-tagging scripts are distributed together with the data.

Keywords: social media, Twitter, part-of-speech, named entities, named entity recognition, sentiment Analysis

1. Overview

High-quality newswire manually annotated linguistic corpora, with different types of annotations, are now available for different languages. Over the past few years, new social media based linguistic corpora have begun appearing but few are focused on classical problems such as Part-of-Speech tagging and Named Entity Recognition. Of these few, most are English corpora.

It has been documented that social media text poses additional challenges to automatic annotation methods with error rates up to ten times higher than on newswire for some state-of-the-art PoS taggers (Derczynski et al., 2013a). It has been shown that adapting methods specifically to social media text, with the aid of even a small manually annotated corpus, can help improve results significantly (Ritter et al., 2011; Derczynski et al., 2013a; Derczynski et al., 2013b).

While there exist social media sentiment corpora for twitter messages in the languages we annotated, the corpus we are presenting also includes message level sentiment labels. One motivation for this is the potential contribution of annotations, such as PoS tags, to sentiment classification tasks (Zhu et al., 2014).

The xLiMe Twitter Corpus provides linguistically annotated Twitter[1] social media messages, known as ”tweets”, in German, Italian, and Spanish. The corpus contains approximately 350K tokens with POS tags and Named Entity annotations. All messages, approximately 20K, are labeled with message level sentiment polarity. We further explain the composition of the corpus in § 3.

2. Related Work

An early effort in linguistically annotating noisy online text was the NPS Chat Corpus (Forsyth and Martell, 2007) which contains more than 10K online chat messages, written in English, manually annotated with POS tags.

The Ritter twitter corpus (Ritter et al., 2011) was the first to introduce a manually annotated Named Entity recognition corpus for twitter. It contains 800 English messages (16K tokens) which also contain Part-of-Speech and chunking tags.

The (Gimpel et al., 2011) corpus contains almost 2K twitter messages with POS tags while (Owoputi et al., 2013) annotated 547 twitter messages. Tweebank drawn from the latter boasts a total of 929 tweets (12,318 tokens) as well as providing clear guidelines which the previously mentioned twitter annotation efforts had not.

While there are many English social media sentiment corpora, the most well known is probably the Semeval corpus (Rosenthal et al., 2014) which contains over 21K Twitter messages, SMS, and LiveJournal sentences. All messages are annotated with one of three possible labels: Positive, Negative, or Objective/Neutral. For Spanish sentiment classification, the TASS corpus (Villena Román et al., 2013) contains 68K Twitter messages labeled semi-automatically with one of five labels: the three Semeval labels plus Strong Positive and Strong Negative. Smaller corpora with at least three labels exist for many other languages including German and Italian. We decided to add sentiment polarity to our multilingual corpus because it is a popular task, challenging for automated methods, and the cost (annotator time) of adding this additional annotation is mostly marginal when compared to the cost of PoS and NER annotations.

3. Description

The developed multilingual social media corpus includes document level and token-level annotations. There is one document level annotation, Sentiment polarity and two (2) token-level annotations, PoS and NER. The corpus details are shown in table 1, namely the distribution of annotated tweets and tokens per language. The Italian part of the corpus is the largest with 8601 annotated tweets, followed by Spanish with 7668 tweets, and German containing 3400 tweets.

3.1 Data Collection

The tweets were randomly sampled from the twitter public stream from late 2013 to early 2015. Tweets were selected based on their reported language. Some rules were automatically applied to discard spam and low information tweets (”garbage”) tweets:

[1] Twitter: http://twitter.com

1. Tweets with less than 5 tokens were discarded;
2. Tweets with more than 3 mentions were discarded;
3. Tweets with more than 2 URLs were discarded;
4. Automatic language identification with langid.py (Lui and Baldwin, 2011) was used on the tweet text without twitter entities and if didn't match the reported language, the tweet was discarded.

Language	Tweets	Tokens	Annotators
German	3400	60873	2
Italian	8601	162269	3
Spanish	7668	140852	2

Table 1: Number of annotated tweets and tokens per language.

3.2 Preprocessing

URLs and Mentions were replaced with pre-specified tokens. Tokenization was performed using a variant of twokenize (O'Connor et al., 2010) that was additionally adapted to break apart apostrophes in Italian as in "l'amica" which becomes "l'", "amica".

3.3 Annotation Process

There were two annotators for Spanish, two for German, and three for Italian. A small number of tweets for each language were annotated by all the annotators working on the language in order to allow estimation of agreement measures as described in § 4.. POS tags were pre-tagged using Pattern (De Smedt and Daelemans, 2012) and some basic rules for twitter entities such as URLs and mentions.

We built an annotation tool optimized for document and token level annotation of very short documents, i.e. tweets. The annotation tool included the option to mark tweets as "invalid" since despite the automatic filtering performed in § 3.1 it was still possible that tweets with incorrectly identified language, spam, or incomprehensible text might be presented to the annotators. This feature can be seen in fig. 1.

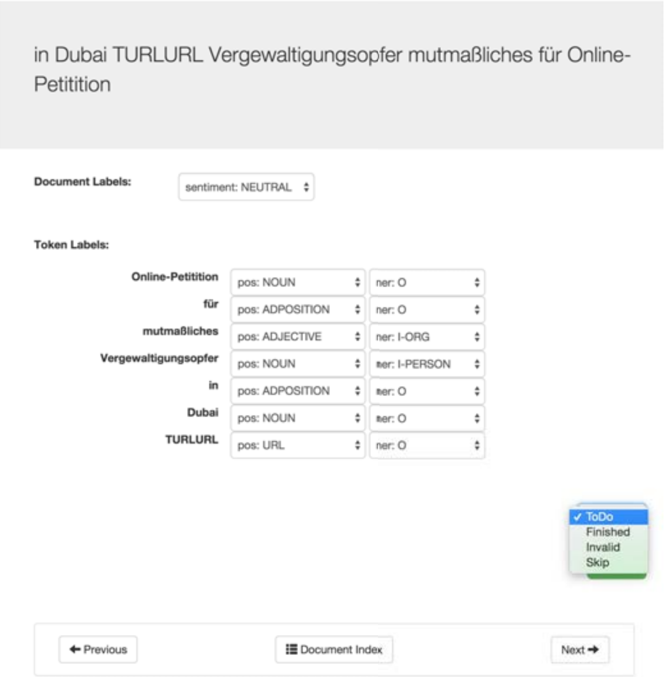
3.4 Part-of-Speech

The part of speech tagset consists of the Universal Dependencies tagset (Petrov et al., 2012) plus twitter specific tags based on Tweebank (Owoputi et al., 2013). We present the full tagset and the number of occurrences, per language, of each tag in table 2.

3.41. Twitter Specific Tags

While most tags will be easily recognizable to most readers, we believe it is useful to provide here a description of the tags which are specific to social media and twitter. Further details about these tags can be found in our guidelines.

Continuation indicates retweet indicators such as "rt" and ":" in "rt @jack: twitter is cool" and ellipsis that mark a truncated tweet rather than purposeful ellipsis;



in Dubai TURLURL Vergewaltigungsopfer mutmaßliches für Online-Petitition

Document Labels: sentiment: NEUTRAL

Token Labels:

Online-Petitition	pos: NOUN	ner: O
für	pos: ADPOSITION	ner: O
mutmaßliches	pos: ADJECTIVE	ner: I-ORG
Vergewaltigungsopfer	pos: NOUN	ner: I-PERSON
in	pos: ADPOSITION	ner: O
Dubai	pos: NOUN	ner: O
TURLURL	pos: URL	ner: O

Figure 1: Screenshot of the annotation tool interface. The text of a tweet is at the top followed by the sentiment label dropdown menu. Below there is a column with the tokens and rows for each annotation (PoS and NER). Annotators manually fix the errors inherent in the automatic pre-tagging step previously described. Finally, a dropdown menu allows marking the annotation of the document as "To Do", "Finished", "Invalid", or "Skip". Note that in this example, the labels have not yet been manually corrected.

Tag	German	Italian	Spanish
Adjective	2514	7684	5741
Adposition	4333	14960	13467
Adverb	4173	8476	6116
Conjunction	1576	6737	6684
Determiner	2990	9811	10037
Interjection	225	1427	1109
Noun	11057	30759	23230
Number	1176	2550	1568
Other	1936	1503	3033
Particle	638	352	18
Pronoun	4530	7737	10333
Punctuation	8650	20529	14102
Verb	6506	21793	19460
Continuation	918	4227	3422
Emoticon	449	1076	951
Hashtag	1895	3035	1805
Mention	1984	6519	9070
URL	1923	4494	3019

Table 2: Tagset with occurrence counts in the corpus per language.

Emoticon this tag applies to unicode emoticons and traditional smileys, e.g. ":)";

Hashtag this tag applies to the "#" symbol of twitter hashtags, and to the following token if and only if it is not a proper part-of-speech;

Mention this indicates a twitter "@-mention" such as "@jack" in the example above;

URL indicates URLs e.g. "http://example.com" or "example.com";

A noteworthy guideline is the case of the Hashtag. Twitter hashtags are often just topic words outside of the sentence structure and not really part-of-speech. In this case, the Hashtag PoS tag applies to the word following the "#" symbol. Otherwise, if it is part of the sentence structure, the guideline specifies that it should be labeled as if the "#" symbol was not present.

3.5 Named Entities

Named entities are phrases that contain the names of persons, organizations, and locations. Identifying these in newswire text was the purpose of the CoNLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003). We have adopted the definitions for each named entity class: Person, Location, Organization, and Miscellaneous. In table 3 we show each type of entity in our corpus and the number of tokens annotated with each per language.

Entity Type	German	Italian	Spanish
Location	742	2087	1441
Miscellaneous	995	5802	775
Organization	350	1150	836
Person	757	3701	2321

Table 3: Token counts per named entity type per language in the corpus.

3.6 Sentiment

Each tweet is labeled with its sentiment polarity: positive, neutral/objective, or negative. The choice of this three labels mirrors that of the Semeval Shared Task (Rosenthal et al., 2014). The vast majority of tweets in our corpus was annotated with the Neutral/Objective label as we show in table 4.

Language	Positive	Neutral	Negative	Total
German	334	2924	142	3400
Italian	554	7524	523	8601
Spanish	388	7083	197	7668

Table 4: Message level sentiment polarity annotation counts.

4. Agreement

In order to estimate inter-annotator agreement, for each language, the annotators were given tweets that they annotated in common. We show the number of tweets and tokens in table 5. These were then used to calculate Cohen's Kappa (technically, Fleiss' Kappa for Italian) and we show the results in table 6. The worst agreement between the human annotators occurred when labeling sentiment. Even for humans, it can be challenging to assign sentiment, without context, to a small message.

Language	Tweets	Tokens	Annotators
German	47	791	2
Italian	45	758	3
Spanish	45	721	2

Table 5: Number of tweets and tokens annotated by all annotators for a given language.

Task	German	Italian	Spanish
PoS	0.88 (AP)	0.87 (AP)	0.85 (AP)
NER	0.67 (SUB)	0.42 (MOD)	0.51 (MOD)
Sentiment	-0.07 (Poor)	0.02 (Slight)	0.37 (Fair)

Table 6: Inter Annotator Agreement (Cohen/Fleiss kappa) per task per language. In parenthesis, the human readable interpretation where: AP - Almost Perfect, MOD - Moderate, SUB - Substantial.

5. Format and Availability

The corpus is primarily distributed online[2] as a set of three tab-separated values (TSV) files - one per language. We also distribute the data in language and task specific formats such as a text file containing the German tweets with one word per line followed by a whitespace character and a NER label. These were automatically created using a script described in § 5.2.

5.1 Headers

Each of the TSV files has the same set of headers:

token the token, e.g. "levantan";

tok_id a unique identifier for the token in the current message, composed of the tweet id, followed by the dash character, followed by a token id, e.g. "417649074901250048-47407";

doc_id a unique identifier for the message (tweet id), e.g.: "417649074901250048";

doc_task_sentiment the sentiment label assigned by the annotator;

tok_task_pos the Part-of-Speech tag assigned by the annotator;

tok_task_ner the entity class label assigned by the annotator;

annotator the unique identifier for the annotator.

Note that the combination of the token identifier and the annotator identifier is unique i.e. the combination is present only once in the corpus.

[2] xLiMe Twitter Corpus: `https://github.com/lrei/xlime_twitter_corpus`

5.2 Scripts

In order to facilitate experiments using this corpus as well as to replicate its construction, several python scripts are distributed with the corpus data. We detail the most important scripts here. Namely the tokenizer, the pre-tagger, and the script that converts the sequence tagging tasks (PoS and NER) into a format similar to the CoNLL 2002/2003 format. In this format, there are empty lines which mark the end of a tweet and ”word” lines start with the token followed by a space, followed by a tag.

xlime2conll.py the script used to convert the data into the column format similar to the CoNLL 2003 shared task;

extract_sentiment.py the script used to convert the data into a format that is easy to handle by text classification tools, specifically, a TSV file with the headers: id, text, sentiment;

twokenize.py the tokenizer used to split the tokens in the corpus;

pretag.py the script used to pre-tag the data;

agreement.py the script used to calculate the agreement measures.

6. Acknowledgments

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under xLiMe (FP7-ICT-611346) and Symphony (FP7-ICT- 611875). We would like to thank the annotators that were involved in producing the xLiMe Twitter Corpus. The annotators for German were M, Helbl and I. Škrjanec; for Italian, E. Dervišević, J. Jesenovec, and V. Zelj; and for Spanish, M. Kmet and E. Podobnik.

7. References

De Smedt, T. and Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Derczynski, L., Maynard, D., Aswani, N., and Bontcheva, K. (2013a). Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013b). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP).*, pages 198–206. Association for Computational Linguistics.

Forsyth, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 19–26. IEEE.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Lui, M. and Baldwin, T. (2011). Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*.

O'Connor, B., Krieger, M., and Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM 2010)*, pages 384–385.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1524–1534.

Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Villena Román, J., Lana Serrano, S., Martínez Cámara, E., and González Cristóbal, J. C. (2013). Tass-workshop on sentiment analysis at sepln.

Zhu, X., Kiritchenko, S., and Mohammad, S. M. (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447.

Political Discourse in Polish Internet – Corpus of Highly Emotive Internet Discussions

Antoni Sobkowicz
National Information Processing Institute
E-mail: antoni.sobkowicz@opi.org.pl

Abstract

In this work, we present description and initial statistical analysis on a corpus of comments from most popular polish news-related website, Onet.pl. Presented corpus contains highly informal texts, politically polarized texts, with highly emotive content. We gathered corpus containing 4,829,076 texts and 1,826,906 unique tokens total during 9 month time period which held several important political events in Poland. Presented corpus is freely available, and we intend to update it regularly, with additional texts being currently retrieved.

Keywords: politically related texts, polish language corpus, social media, computer-mediated communications

1. Introduction

Discussion about politics on the internet, especially in such politically polarized country as Poland – with supporters of two dominating parties being very vocal and active - are often very emotive. People tend to not only express their feelings about events but resort to personal insults or insults directed at politicians. This makes these discussions very interesting for analysis.

We have gathered over 4.8 million comments from largest polish news-oriented website, Onet.pl, choosing only comments under political related news, over the 9 months that were very intensive in terms of political events in the country (presidential and parliamentary elections where party ruling for last 8 years lost, Constitutional Tribunal crisis, changes in public media). We have analyzed basic properties of this set and we encourage researchers in text analysis related fields to use it. Collected dataset is freely available, and we intend to update it every three months with new content.

Dataset described in this paper was previously used in several works, although it was not publicly available.

2. Related Work

Corpora regarding political text are widely available, with examples being a multilingual corpus of annotated political programs (Merz et al., 2016), the corpus of political speeches with annotated audience reactions (Guerini et al., 2013) or political speech corpus of Bulgarian (Osenova & Simov, 2012). This corpus however only touches text with a higher degree of formalization.

Corpora build on less formal text sources are also available – based on Tweets (Longhi & Wigham, 2015), blogs (Eisenstein & Xing, 2010) and other sources. These are more similar to corpus described in this paper because of the informality of those sources.

Work on similar kind of dataset – politically related comments in the Polish language, also done on Onet.pl data - was done by Sobkowicz and Sobkowicz (2012), although dataset was highly limited.

3. Corpus Source Description

Corpus was scraped from Onet.pl website, one of the largest and most commented news related websites in Polish internet. Onet.pl is a news website, covering topics from politics to sport and entertainment, with complex comment section under each news piece, news tagging.

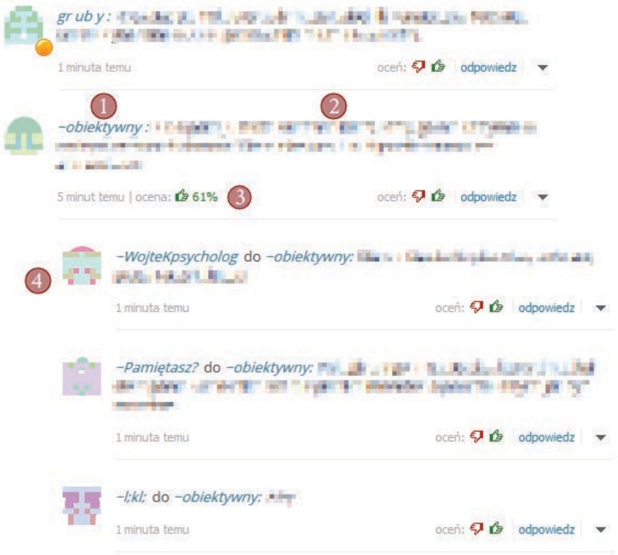


Figure 1: A chunk of typical discussion in the comment section on Onet.pl – source for corpus described in this paper. Elements are as follow: 1 - Comment poster name; 2 - Comment text; 3 - Comment score; 4 - Replies to comment, nested, with information who replied to which post. Texts were blurred out because they may be offensive.

The comment section is tree based, meaning comments that reply to other comments are displayed below with indentation. The user can rate comments, and the average rating is displayed near each comment, along date and time of posting and name of the original poster. An example of such tree and data are is shown in figure 1.

3.1 Discussions on Onet.pl

As Onet.pl does not enforce registrations, the user can post under any nickname. This seems to encourage more heated discussions, with lots of insults – token based analysis of the dataset using only known, heavily emotive negative tokens shown that around 5% of all messages can be considered as directly insulting (insulting other users or other parties connected to the topic of the discussion). Manual sentiment analysis on small randomly selected subset of data shows that purely neutral texts are only 15% of data (146 texts out of 950 assessed). In data analyzed by Sobkowicz and Sobkowicz (2012), neutral texts were 56% of all texts, however, authors used different neutrality and emotiveness measure.

Each posted news piece tends to have several hundred texts, the more dividing in opinion the topic is, the more posts are written by users.

4. Corpus Description

We have gathered comments under articles (along with article text). Data was gathered in three periods, from May – August 2015, September – December 2015 and January – March 2016, with fourth part being currently downloaded.

4.1 Comment Data Description

Comments are scraped from the website while preserving their tree structure, along time of posting and user handle. This information can be used to retrieve back user network if needed. Comments themselves are stored in their raw form, without any alteration to their text. We decided against storing only extracted and processed tokens, as we believe that preserving additional data (such as discussion tree) is very important, and extracting/lemmatizing/stemming can be done when needed.

4.2 Basic Corpus Properties

Corpus contains 4,829,076 texts, with average length of 179 characters and length distribution shown in figure 2. Average length in tokens is 33, with distribution shown in figure 2. Both of distributions seem to follow lognormal distribution as expected from human produced texts (Sobkowicz et al. 2013).

Distribution of unique tokens to a number of texts in the corpus is shown in figure 3 and 4 – non-unique tokens and unique tokens only respectively. Corpus itself contains over 160 million tokens, with 1,826,906 unique tokens (as we do no extract lemmas from the words, this number in inflated by different conjugations).

We do not provide sentiment annotation for the corpus, because given it's size and lack of good sentiment analysis tools for the Polish language, we believe we cannot give accurate or semi-accurate sentiment information for the corpus. This is the case also for lemmatization and POS tagging.

4.3 Anonymization

We believe that given the fact that source website does not require users to register and does not provide any other information about the user beyond their username anonymization is not required for this dataset.

5. Toolset Description

Data was gathered using specialized tools written in Python using scraps library. Scrappers parsed all comment pages, going from first to last, and saving all data to JSON files. These files were then parsed and saved into an SQLite database for easier use.

6. Availability

Corpus is available for free, but taking into consideration the fact that the collected corpus is relatively large in size – around 6GB, we currently do not provide direct download link – instead, we encourage to contact us to prepare data for transfer via selected service.

7. Conclusions and Future Work

We have built new corpus containing politically related comments from under news pieces on largest polish news related site. We gathered over 4.8 million texts spanning 9 months period and calculated basic corpus statistics. In near future, we plan to finish downloading fourth part of data (spanning the time from April to June 2016) and keep corpus up-to-date for foreseeable future.

We encourage researchers to use this corpus and analyze it in greater detail – in the context of linguistics, sentiment analysis, and analysis of human interactions in CMC. We believe that corpus this large, coming from very bi-polar community can be very interesting for researchers.

8. References

Eisenstein, J., & Xing, E. (2010). *The CMU 2008 political blog corpus*. Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Guerini, M., Giampiccolo, D., Moretti, G., Sprugnoli, R., & Strapparava, C. (2013). The new release of corps: A corpus of political speeches annotated with audience reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action* (pp. 86-98). Springer Berlin Heidelberg.

Longhi, J., Wigham, C. R.. (2015) Structuring a CMC corpus of political tweets in TEI: corpus features, ethics, and workflow. *Corpus Linguistics 2015*

Merz N., Regel, S., Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*

Osenova, P., & Simov, K. (2012). The Political Speech Corpus of Bulgarian. In *LREC* (pp. 1744-1747).

Sobkowicz, P., Thelwall, M., Buckley, K., Paltoglou, G., & Sobkowicz, A. (2013). Lognormal distributions of user post lengths in Internet discussions-a consequence of the Weber-Fechner law?. *EPJ Data Science*, *2*(1), 1-20.

Sobkowicz, P., & Sobkowicz, A. (2012). Two-year study of emotion and communication patterns in a highly polarized political discussion forum. *Social Science Computer Review*, 0894439312436512.

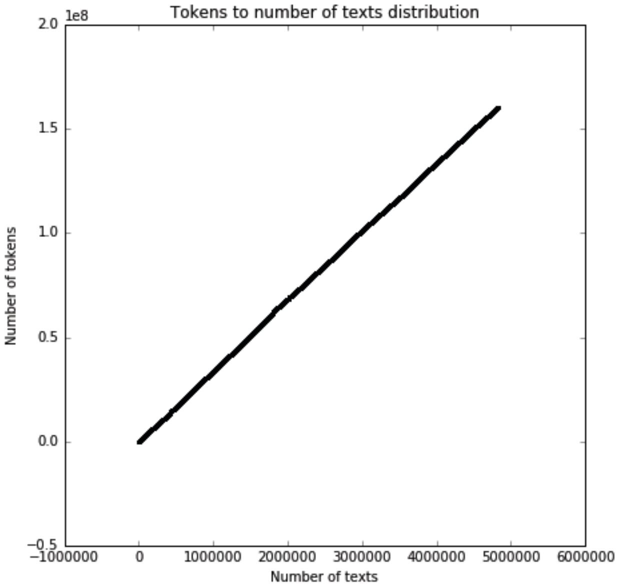


Figure 3: Distribution of number of non-unique tokens to number of texts in corpus.

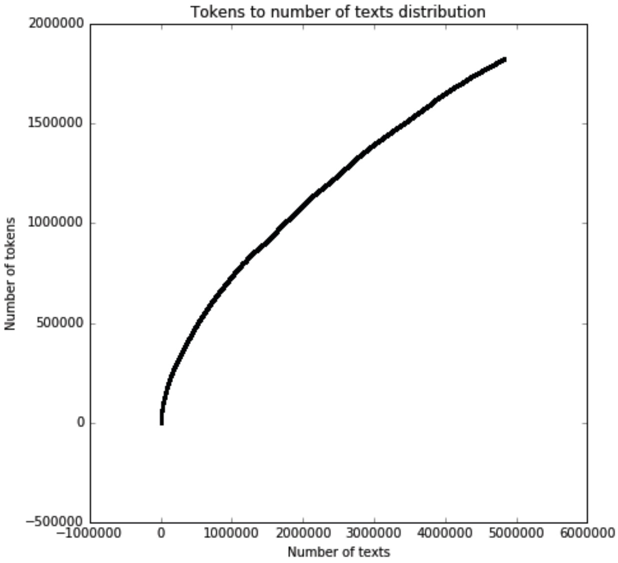


Figure 4: Distribution of number of unique tokens to number of texts in corpus.

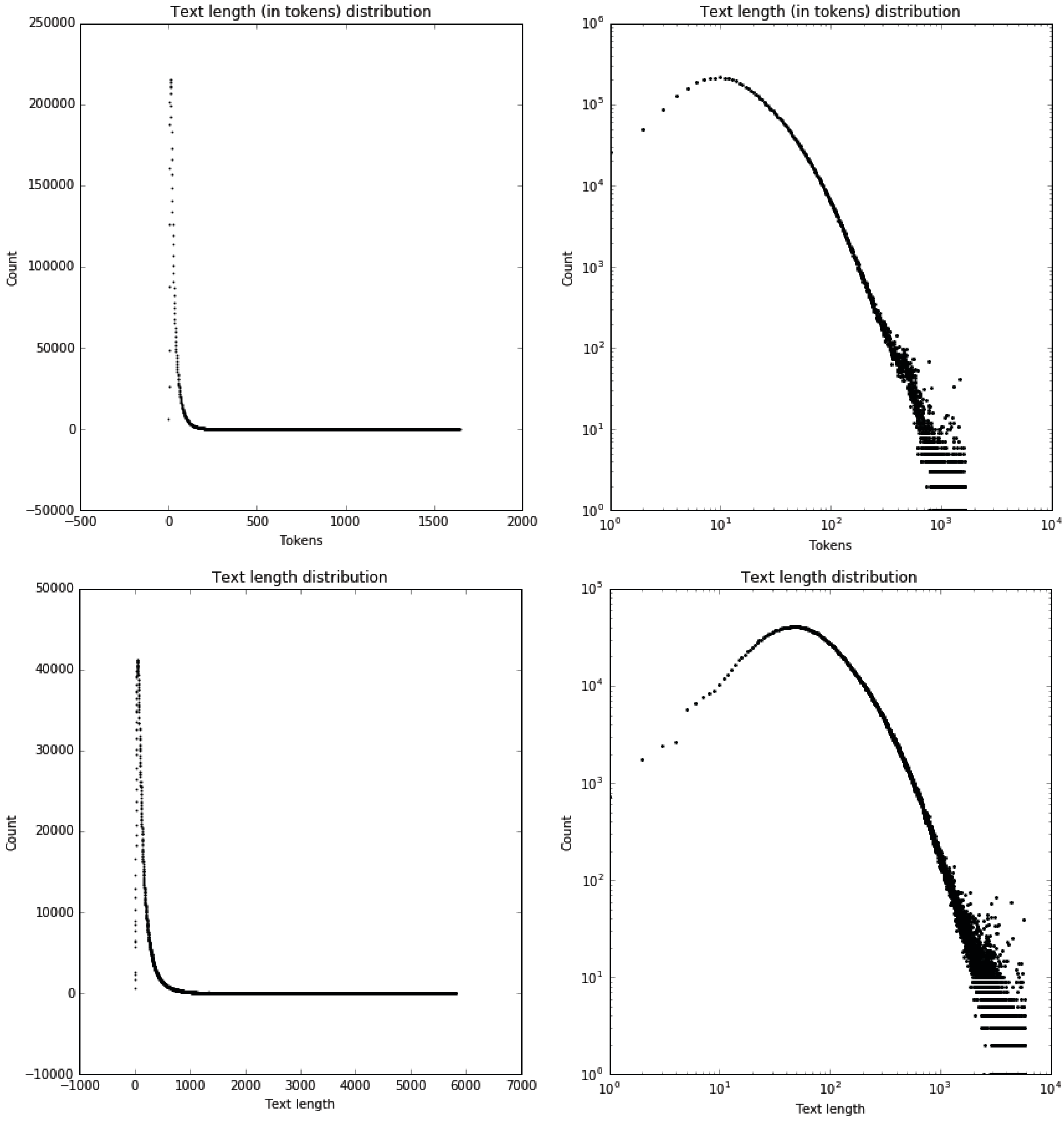


Figure 2: Distributions of text length in the corpus, both in raw character length and in token length. Right figures show the distribution in log-log scale. Both distributions seem to follow log-normal distribution, which seems to be the case for most of human-created according to other research.

Topic Ontologies of the Slovene Blogosphere: A Gender Perspective

Iza Škrjanec[1], Senja Pollak[2]
[1]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[2]Jožef Stefan Institute, Ljubljana, Slovenia
skrjanec.iza@gmail.com, senja.pollak@ijs.si

Abstract

In the past years, blogs have become an increasingly popular genre for publishing content on the Web. Blogs are also one of the five genres of the Janes corpus of Slovene user-generated content. The aim of this paper is to explore the topics of the blog subcorpus of Janes using OntoGen, a semi-automatic and data-driven ontology editor. In addition to the construction of the topic ontology of the blogs from two Slovene blog portals, special focus is placed on the topical variation in entries by male and female bloggers. First, the keywords of selected topics differentiating male and female blog entries are analysed. Next, we present two topic ontologies, one based on blog entries by private female and the other by private male users, and contrast them against each other. The analysis has shown that both groups write about politics, family, romance and sexuality, environment, and nutrition. Men seem to blog more about spectator sports, music and literature, the Roman-Catholic Church, the refugee crisis, and biology; in contrast, female authors discuss religion, emotions and social politics.

Keywords: blogs, topic ontologies, gender, keyword analysis

1. Introduction

In corpus linguistics, corpora serve as the main resource for either testing various hypotheses or developing linguistic theories based on the corpus data. This is why it is important to learn about the properties of the corpus we are working with, e.g. recognizing frequent topics by observing keywords (Kilgarriff, 2012).

For Slovene, the topics of blogs in particular have not yet been studied; however, Logar Berginc and Ljubešić (2013) contrasted two Slovene corpora of various genres against each other: the crawled slWaC[1] corpus and the reference Gigafida[2] corpus. Using the LDA topic modelling method, a number of n topics for each of the corpora was constructed. When comparing the topics, Logar Berginc and Ljubešić found that some topics appeared in both corpora (domestic policy, team sports, finance, war, terrorism, publications and culture, local politics, health and law). The slWaC corpus contains more documents on film and music, travelling and tourism, foreign affairs and classified ads. In contrast, the following topics are more prominent in Gigafida: cities, street traffic, public events, television and radio programs, individual sports, and work. Some differences between the reference corpus and the Janes corpus including blog entries (but also tweets, news comments, forum posts) have been identified through collocation analysis in Pollak (2015).

In this paper, we focus on topical variation between male and female bloggers. For English there have been some studies on how the content in social networks posts or spoken language correlates with the demographic factors of users, such as gender and age. Using data mining techniques, Argamon et al. (2007) found that male bloggers tend to write about religion, politics, business and the Internet more frequently, while female bloggers blog about conversation, domestic environment, fun, romance, and swearing more than men. Schmid (2003) carried out a comparable study on the spoken part of the BNC corpus. He conducted a list of words typical for 14 different topics. Using relative frequencies, he observed which topics are more dominant in the corpus of female and male speakers. An overrepresentation of female speakers was detected in topics dealing with clothing, basic colors, home, food and drink, body and health, and people. In contrast, the domains of work, computing, sports, and public affairs were considered more typical of the male subcorpus. The domains on swearing and car and traffic occurred equally in the speech of both groups.

In comparison to the topic keyword analysis as for example in Logar Berginc and Ljubešić (2013), the approach selected for this study results in hierarchical ontologies which allow the identification of subtopics for each topic, enables the user to be involved in the process of ontology construction, and provides the visualization of the constructed ontologies. In addition to the understanding of the topics of the Slovene blogosphere, the main contribution of our paper is the research of gender and the Slovene language in social media. We thus wish to contribute to existing studies, e.g. on the use of emoticons and expressive punctuation in tweets (Osrajnik et al., 2015) and the discourses about women and men (Škrjanec et al., 2016).

The rest of the paper is structured as follows. In Section 2, the blog subcorpus and the text preparation are presented. The OntoGen tool and the ontology construction process are described in Section 3, and discussed in Section 4. In Section 5, we conclude the paper and suggest further work.

2. Corpus Description and Data Preparation

The corpus of Slovene blogs used in this paper is one of the subcorpora in version 04 of the Janes corpus of user-

[1] http://nlp.ffzg.hr/resources/corpora/slwac/

[2] http://www.gigafida.net/

generated Slovene. The corpus was compiled within the Janes[3] research project and contains various genres of user-generated content: tweets, news comments, forum posts, user and talk pages from Wikipedia, and blog entries and blog comments. In this paper, the focus is placed on the compilation and properties of the blog subcorpus, for which two Slovene blog portals were crawled: *publishwall.si* and *rtvslo.si* (Fišer et al., to appear).

The blog entries of the Janes corpus were contributed by over 800 users, which we annotated for their account type (private or corporate) and gender[4] (female, male and undefined). Corporate accounts belong to different companies or journalists, the rest are private. The gender was manually assigned based on the use of grammatical gender when referring to self; the profile picture and username. If we were not able to identify the user as male or female, the tag "neutral" (meaning "undefined") was used.

For our study, we selected the blog posts of male and female private users. Private users wrote over 29,000 entries altogether (female: 9,056; male: 20,105). For the ontology construction, blog entries in Slovene were taken into consideration.

Disregarding the gender and account type, the average length of blog entries and comments in the entire blog subcorpus is about 70.16 words (85.42 tokens) per entry. Since clustering algorithms perform better on longer texts than on shorter ones, blog entries with minimum of 100 full words (no stop words) were used for ontology construction (9,039 entries by male and 3,771 by female users).

2.1.1 Text Preparation

The original vertical file of the Janes blog subcorpus was parsed into a format supported by OntoGen, in which each blog entry is represented with a single line containing the blog entry ID, category (female or male), and the lemma form of each token. All preprocessing steps were carried out with a simple Python programme. Stop words were removed. OntoGen cannot process diacritics and other special characters, so these were replaced with character sequences that enable the reconstruction of the original form.

3. Topic Ontology Construction

In this section, the OntoGen tool and the construction of three topic ontologies are presented.

3.1 The OntoGen Tool

For the construction of Slovene blog topic ontologies, we used the OntoGen tool[5], which is a semi-automatic data-driven ontology editor that combines text mining techniques with a fairly simple user interface (Fortuna et al., 2007). OntoGen is based on Bag-of-Words (BoW) vector representations of documents, weighted by the Term Frequency-Inverse Document Frequency weights. The tool provides subtopic suggestions based on the k-means clustering algorithm, with the parameter k being set by the user. The user then decides whether to add the clusters to the ontology. The user can also manually move the documents and provide labels for the clusters (topics). Additionally, if the input documents are pre-categorized, a method for grouping the instances according to the labels is also supported.

The user can influence the division into subtopics by employing the Active learning functionality that is based on the SVM (Support Vector Machines) active learning method. The user provides a term or a set of keywords that represent a new subtopic to be added to the ontology. This action is followed by iterative model refinement through user interaction by answering to the question whether a

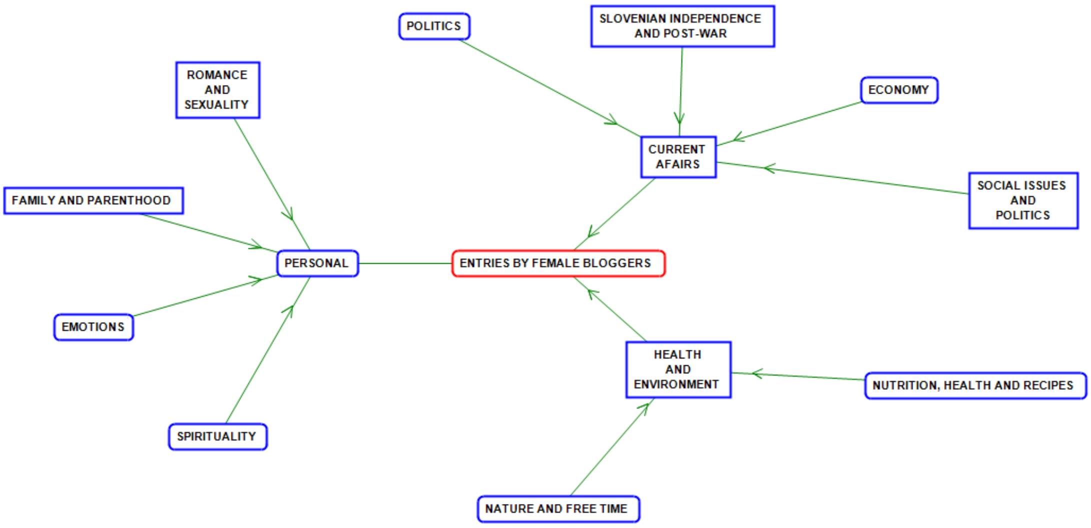


Figure 1: Topic ontology of entries by female bloggers.

[3] The Janes project webpage: http://nl.ijs.si/janes/

[4] The bloggers themselves were not contacted and asked about their self-identification. Thus, the claims on their gender are based solely on the use of grammatical gender.

[5] The OntoGen tool: http://ontogen.ijs.si/

particular document belongs to the topic or not. The user can decide when to stop the active learning process and the new (sub)topic is added to the ontology.

For each topic, OntoGen provides a list of keywords, which are the words that are the most descriptive for the content of cluster, i.e. words with the highest weights in the document centroid vectors (ibid.). Another view is gained by inspecting SVM keywords, which are the words most distinctive for the selected concept with regard to its sibling concepts in the hierarchy (e.g. words contrasting male and female entries categorized in a selected topic).

3.2 The Construction of Three Topic Ontologies

The dataset with entries by private male and female bloggers was imported into OntoGen. The ontology was built using k-means for topic suggestions, and the active learning functionality, as well as by manually arranging the ontology. Because the entries were pre-categorized according to the user gender, we could examine the keywords and SVM keywords of topics, whereby the topics with a more or less comparable number of entries by female and male bloggers were selected for analysis. Two topics (*Romance and sexuality*; *Political system*) and their keywords are presented in Table 1. In addition, we constructed a topic ontology for entries by female (Figure 1) and male (Figure 2) users[6].

	Female		Male	
	Keywords	SVM k.	Keywords	SVM k.
Romance and sexuality	moški, ženska, film, želeti, partner, življenje, ljubezen, prijatelj, ženski, odnos #431	moški, želeti, partner, čutiti, strah, razmišljati, potrebovati, fb, spolnost, telo	ženska, moški, film, sex, prijatelj, ženski, žena, rak, dekle, želeti #329	sex, ženska, žena, film, mati, bivši_žena, obraz, zgodbica, brada, punca
Political system	družba, sistem, obstajati, lasten, ego, narod, zavest, vrednota, človeški, življenje #129	želeti, obstajati, narod, telo, izkušnja, lasten, ego, sposoben, različen, zavest	družba, sistem, kapitalizem, družben, problem, človeški, življenje, planet, znanost, vrednota #503	družba, sistem, bitje, sodoben, demokracija, ideja, planet, materialen, svoboda, stoletje

Table 1: Keywords for the topics *Romance and sexuality*, and *Political system*.

4. Discussion

A keyword list can tell us something more about the main ideas and concepts users blog about concerning a particular topic. After constructing a common topic ontology of entries by both groups of users, we observed the entries on romance and sexuality to compare the keywords and SVM keywords of both groups. From keywords in Table 1, we can learn that male and female bloggers use similar keywords (“woman”, “man”, “want”) with some variation. Observing the SVM keywords, which point out the differences, it is evident that female bloggers use more verbs (“feel”, “think”, “need”), while male bloggers focus more on the participants (“ex-wife”, “girlfriend”, “mother”) and appearance (“face”, “beard”). The keyword “crab”/”cancer” suggests that the topic is still somewhat noisy. The keywords for the topic *Political system* also reveal similarity between entries by men and women (“society”, “system”, “life”, “human”), whereas in entries by female bloggers the topic of nation comes forward. In the entries by male bloggers, terms like “capitalism”,

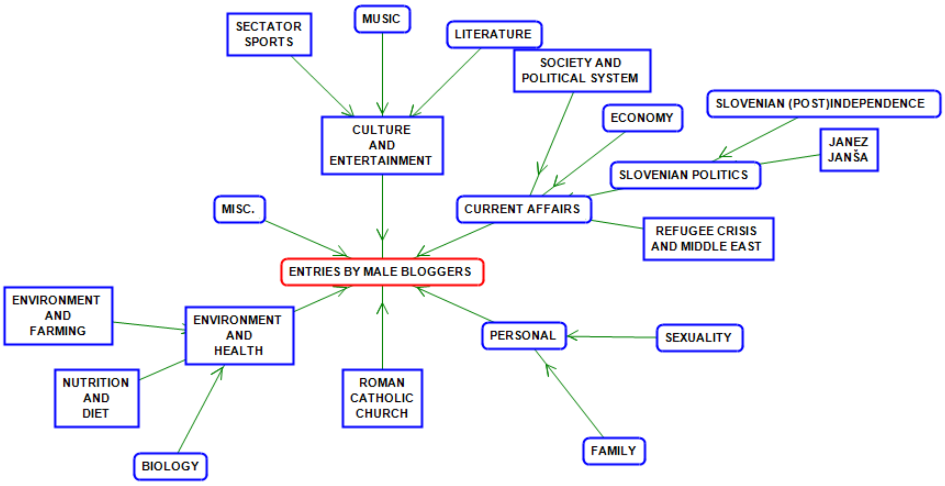


Figure 2: Topic ontology of entries by male bloggers.

[6] For the common ontology, lemmas of uni- and bigrams with the minimum frequency of 20, and for the gender specific ontology, the minimum frequency was set to 10.

"democracy" and "freedom" indicate a specific political issue.

A topical comparison of blog entries by female and male bloggers (Figures 1 and 2) shows some interesting similarities and differences. Both groups seem to write about the environment, nutrition, family and parenthood, sexuality, and politics, in particular the subtopic on Slovenian politics and the (post)independence era (the Independence War, post-war killings and the role of former communists today). Another common topic is economy (mostly Slovene and EU). One of the more prominent topics of male bloggers is that on the Slovene politician Janez Janša, mostly concerning his 2013–2015 trial for corruption. An evident topic on current affairs is also that of the refugee crisis in the male ontology. In contrast to female bloggers, male authors contributed a significant number of entries on biology, spectator sports, music and literature. They also discuss the role of the Roman Catholic Church. In turn, female bloggers write more about spirituality in connection to various religious beliefs, and nature. Emotions are also a prevalent blog topic of female users; additionally, they pay special attention to social politics and issues, such as handicapped people and their social rights.

5. Conclusion

In the paper, we described the process of topic ontology construction and keyword analysis of blog entries from two Slovene blog portals. The goal was to contrast the topics covered by female and male bloggers.

To avoid over-generalization on gendered topics, it is important to take into account the distribution of blog entries among bloggers. Some topics are heavily dominated by a very small number of bloggers (*Biology*, *Social issues and politics*), but this is not visible in the ontology. When using quantitative methods to explore gender and language use, it seems the tendency is to favour differences, while backgrounding similarities, what Baker (2014) calls the "difference mindset". The findings of studies such as this one may suggest and show mostly the differences. However, the language and topics of a single gendered group is not homogenous, which is what Baker (ibid.) discovered when he contrasted "same-sex" parts of the BNC spoken among each other using Manhattan Distance for a list of keywords. He found that some pairs of "same-sex" parts vary more than pairs of "mixed-sex" combinations.

In spite of considering this issue, the analysis has shown that some topics (*Refugee crisis*, *Janez Janša, Biology*, *Spectator sports*, *Music and literature*) seem more prominent in entries by male bloggers, while female bloggers typically contribute to topics like *Religion*, *Nature*, *Emotions*, *Social politics*. When writing about mutual topics (*Romance and sexuality*, *Political system*), female and male bloggers discuss them from different perspectives. Our methodology can be applied to explore the topics of the entire blog subcorpus, including corporate users and those undefined in terms of gender. The information on the predominant topic of the entry could enrich the existing blog metadata: user gender, account type, the linguistic and technical standardness and sentiment of the text. In the future, the automatization of the topic labelling could be performed by combining clustering and terminology extraction as shown in Fortuna et al. (2008). Adding the topic to the metadata enables a more fine-grained analysis of discursive strategies for the same topic with regard to the gender of the user, which is something we plan to carry out in the future.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014-2017).

7. References

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker and Johnatan Schler (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Baker, P. (2014). *Using Corpora to Analyze Gender*. London: Bloomsbury.

Fišer, D., Erjavec, T., Ljubešić, N. (to appear): Janes v0.4: korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0 – Special Issue*.

Fortuna, B., Grobelnik, M., Mladenić, D. (2007). OntoGen: Semi-automatic Ontology Editor. *HCI International 2007*, July 2007, Beijing. 309–318.

Fortuna, B., Lavrač, N., Velardi, P. (2008). Advancing Topic Ontology Learning through Term Extraction. In Ho, T., Zhou, Z. (eds), *Proceedings of PRICAI 2008: Trends in Artificial Intelligence*. Hanoi, Vietnam, December 15-19, 2008. 626–635.

Kilgarriff, A. (2012). Getting to know your corpus. In Sojka, P., Horak, A., Kopecek, I. (eds), *Proceedings of the 15th International Conference on Text, Speech and Dialogue* (TSD2012), pages 3-15. Brno, Czech Republic: Springer.

Logar Berginc, N., Ljubešić, N. (2013). Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, 1 (1): 78–110.

Osrajnik, E., Fišer, D., Popič, D. (2015). Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. Fišer, D. (ed), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 50–74.

Pollak, S. (2015). Identifikacija spletno specifičnih kolokacij pogostega besedišča. Fišer, D. (ed), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba FF UL, 57–62.

Schmid, H. J. (2003). Do men and women really live in different cultures? Evidence from the BNC. In: Wilson, A., Rayson, R. and McEnery, T. (eds), *Corpus Linguistics by the Lune. Lódź Studies in Language 8*. Frankfurt: Peter Lang. 185-221.

Škrjanec, I., Sobočan, A. M., Pollak, S. (2016). The lexical environments of woman and man in the corpus of Internet Slovene. In: Granić, J., Kecskes, I. (eds), *Proceeding of the 7th INPRA Conference.* 10-12 June 2016, Split, Croatia. 161.

Linguistic Characteristics of Dutch Computer-Mediated Communication: CMC and School Writing Compared

Lieke Verheijen
Radboud University (Nijmegen, the Netherlands)
E-mail: lieke.verheijen@let.ru.nl

Abstract

Computer-mediated communication has become essential in many youths' lives. Because language in CMC frequently deviates from standard language norms, it is feared to harm youngsters' traditional literacy skills. To determine if and, if so, how social media affect their writing skills, we first need to establish how CMC actually differs from the standard language. This paper presents findings of a study comparing CMC texts and school essays by youths from the Netherlands. Linguistic analyses were done with T-Scan, software specifically designed for Dutch texts. A range of lexical measures (lexical diversity, 'special' words, lexical density, ellipses) and syntactic measures (dependency lengths, subordinate clauses, sentence length, D-level) were studied. Results reveal that in comparison to their school writings, Dutch youths' computer-mediated communication is syntactically less complex, contains more omissions, and is lexically more diverse, different, and dense. These youths thus employ different registers in the writing contexts of CMC and school.

Keywords: computer-mediated communication, social media, writing, register, literacy

1. Introduction

Most youths' daily lives are nowadays filled with computer-mediated communication. Instant messaging, texting, and other social media are essential for them to keep in touch with friends and family. In computer-mediated messages, it is key to communicate effectively, expressively, and informally. As a result, CMC writings frequently differ from standard language conventions (e.g. Thurlow & Brown, 2003; Crystal, 2008; Frehner, 2008; Cougnon & Fairon, 2014). Notable differences are nonstandard orthography and syntax, as in '*fyi i'll B @home l8er 2night, u OK with that? car broke down ☹*'. This sentence contains abbreviations, omissions, an emoticon, and lacks capitalisation and punctuation at the appropriate places. Such deviations in CMC from the 'official' language norms are a source of worry for many parents and language teachers: they fear it damages youths' traditional literacy skills.

2. Research Goals

This paper presents a study that is part of my PhD project into the impact of CMC on literacy. In order to determine whether and, if so, how youths' social media use affects their writings at school, it is imperative to first investigate what youths' CMC actually looks like and how it differs from the standard language. The main goal of this study is to explore in what ways the informal language used by Dutch youths in CMC differs from their more formal school writings. These questions were analysed by means of a manual analysis, as well as an automatic analysis; the present paper focuses on the latter.

3. Methodology

3.1 Materials

For my study into Dutch written CMC, I used a corpus of CMC texts by youths between 12 and 23 years old, with MSN chats, SMS, tweets, and WhatsApp chats. These social media represent four CMC genres: instant messaging with an internet application, text messaging, microblogging, and instant messaging with a mobile phone app. The first three genres were selected from SoNaR ('STEVIN Nederlandstalig Referentiecorpus'), a reference corpus of written Dutch (Treurniet & Sanders, 2012; Oostdijk et al., 2013). WhatsApp chats were gathered especially for the purposes of my project, via a website where youths could voluntarily donate their messages, http://cls.ru.nl/whatsapptaal/. Table 1 shows specifics of the CMC corpus. For comparison, I also collected school writings. These were written by youths of similar ages as the CMC texts, of different educational levels. Table 2 shows more details on the school essays.

Genre	Years of collection	Age group	# words	# chats or contributors
MSN	2009-2010	12-17 18-23	45,051 4,056	106 21
SMS	2011	12-17 18-23	1,009 23,790	7 42
Twitter	2011	12-17 18-23	22,968 99,296	25 83
WhatsApp	2015	12-17 18-23	55,865 140,134	11 / 84 23 / 132
total	2009-2015	12-23	392,169	

# chats: MSN, WhatsApp; # contributors: SMS, Twitter, WhatsApp

Table 1: CMC texts.

Educational level	Years of production	Age group	# words	# texts
lower secondary (*vmbo*)	2013-2014	± 14-15, 3rd grade	50,143	128
higher secondary (*vwo*)	2013-2014	± 14-15, 3rd grade	50,070	153
lower tertiary (*mbo*)	2012-2014	± 17-18, 2nd grade	39,793	137
higher tertiary (*uni*)	2012-2014	± 18-19, 1st grade	50,175	169
total	2012-2014	± 14-19	190,181	587

Table 2: School essays.

3.2 Method

A quantitative corpus study was conducted. For the first part of the analysis, frequencies of several linguistic features were counted manually in the CMC texts. Yet this paper focuses on the second/automatic part of the analysis, comparing the CMC texts to school writings with T-Scan – software specifically designed for Dutch texts (Pander Maat et al., 2014). On the basis of theoretical considerations, a range of relevant lexical and syntactic measures were selected. It was hypothesized that CMC texts, compared to school essays, are lexically more diverse, different, and dense; contain more omissions; and are syntactically less complex. Independent *t*-tests were conducted to compute whether differences were significant; one-tailed probability values are reported here.

4. Results and Discussion

4.1 Lexical Analysis

The measure of textual lexical diversity (MTLD) is the average length of sequential word strings in a text that maintain a type-token ratio (TTR) above a specified threshold (McCarthy & Jarvis, 2010). The MTLD depends on the TTR, which is calculated by dividing the number of types (different words) by the number of tokens (total number of words). Although the TTR is a classic measure, the MTLD is more reliable, because it is insensitive to text length. A higher MTLD value indicates more lexical diversity: more different words or *differently spelled* words. On average, the CMC writings had a higher lexical diversity (M = 119.62, SE = 14.39) than the school writings (M = 76.10, SE = 2.23), $t(10) = -2.08$, $p < 0.05$. Figure 1 shows that the MTLD was higher in the CMC texts, with the exception of WhatsApp chats by 12-17-year-olds.[1] The higher lexical diversity depends on the orthographic variation in written CMC, due to textisms (unconventional spellings, deviating from the standard language norms), misspellings ('errors', as judged by linguistic prescriptivists), and typos (incorrect key presses or false predictions by predictive software). This confirms the hypothesis that CMC is lexically more diverse.

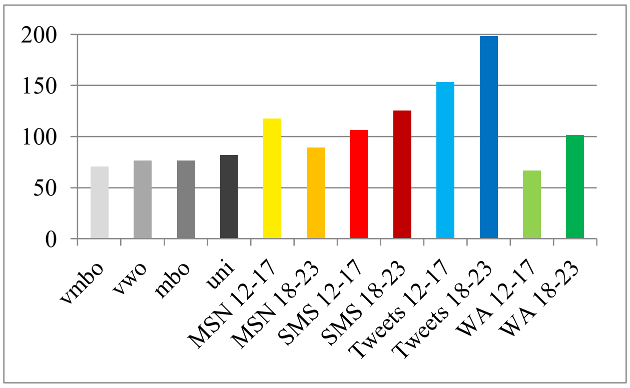


Figure 1: Measure of textual lexical diversity (MTLD).

T-Scan computes the density of 'special words', measured per one thousand words. This includes names, loanwords, numbers, Roman numerals, and times. On average, the CMC writings had a higher density of 'special words' (M = 140.77, SE = 33.20) than the school writings (M = 28.58, SE = 4.02), $t(10) = -3.35$, $p < .01$. Figure 2 illustrates this and shows that there is much variation between CMC genres. The greater frequency of 'special words' is because of textisms, misspellings, typos, and URLs in CMC – character strings that T-Scan cannot recognize as words, since they deviate orthographically from Standard Dutch and are not listed in any standard dictionaries. Tweets in particular include many URLs and 'words' of the format *@username*, within messages in response to another user's tweet (replies) or messages directed at another user (mentions). This higher density endorses the hypothesis that CMC is lexically more different from the standard language.

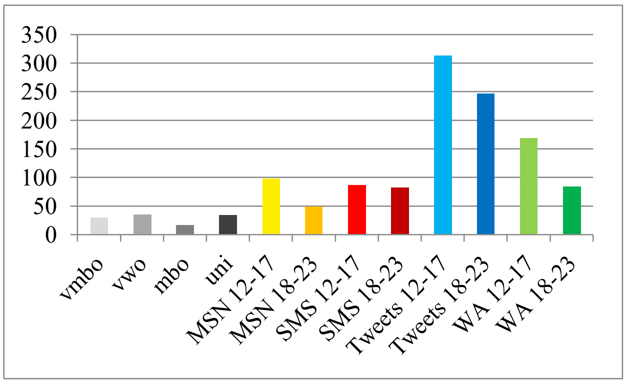


Figure 2: Density of 'special words'.

The third lexical measure that was selected is lexical density. This is the number of content words (nouns, verbs, adjectives, and adverbs) per one thousand words (e.g. Johansson, 2008). When a text has a high lexical density, it contains many content words and few function words. On average, the CMC writings had a higher lexical density (M = 531.70, SE = 9.28) than the school writings (M = 481.31, SE = 2.68), $t(10) = -3.71$, $p < .01$, as shown in Figure 3. This is due to the frequent omission of function words in CMC, which is known for its concise writing style, somewhat similar to that of telegrams or newspaper headlines. The findings from T-Scan thus support the hypothesis that CMC is lexically denser.

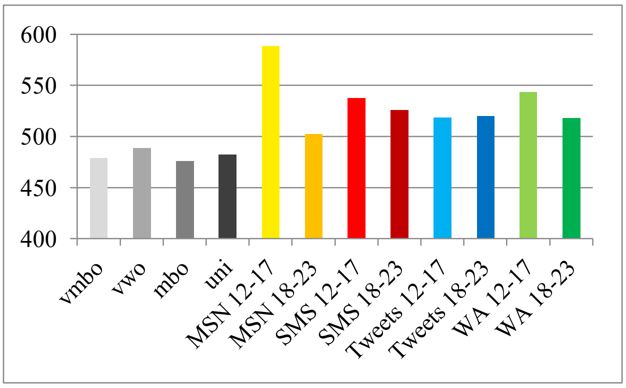


Figure 3: Lexical density.

[1] This apparent exception can be attributed to the frequent repetition of chain messages and certain words in a spam-like manner by one contributor; excluding this outlier, the MTLD would be 92.70 – higher than the school essays, as hypothesized.

Another interesting measure is the density of elliptical constructions, quantified as the number of finite verbs without a subject per one thousand words. On average, the CMC writings had a higher density of ellipses (*M* = 25.86, *SE* = 3.17) than the school writings (*M* = 8.60, *SE* = 1.18), *t*(10) = -5.10, *p* < .001. Figure 4 shows that the CMC writings of all genres contained more elided subjects (though just barely for MSN chats by 18-23 year olds). This backs up the abovementioned results on lexical density: informal written CMC contains fewer function words than formal school essays, at least partly due to the frequent omission of grammatical subjects.

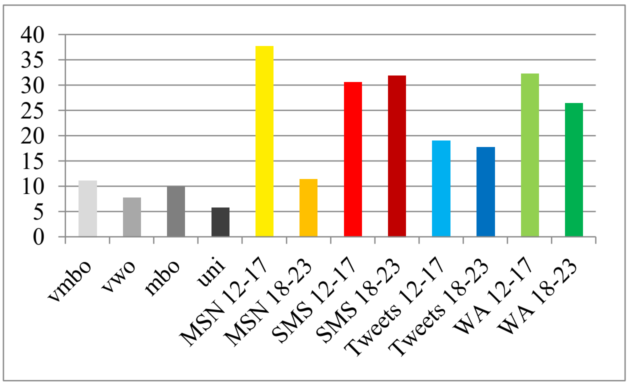


Figure 4: Density of ellipses.

4.2 Syntactic Analysis

One measure of syntactic complexity is the average of all dependency lengths per sentence. The dependency length is the distance between a head (of a sentence or phrase) and its dependent, such as a finite verb and the subject or an article and the corresponding noun. T-Scan expresses the distance in number of words that need to be skipped from head to dependent. Texts with a higher average dependency length contain more discontinuous structures, making them syntactically more complex and more difficult to process for readers (Gibson, 2000). On average, the CMC writings had a lower average of all dependency lengths per sentence (*M* = 0.63, *SE* = 0.06) than the school writings (*M* = 1.59, *SE* = 0.10), *t*(10) = 9.04, *p* < .001. It is clear from Figure 5 that the CMC texts of all genres had lower average dependency lengths, no matter what the writer's age or educational level. This supports the idea that CMC is syntactically less complex.

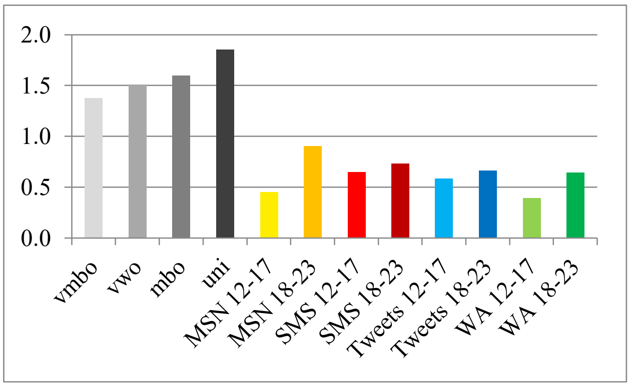


Figure 5: Average of all dependency lengths per sentence.

T-Scan also measures the average number of subordinate clauses per sentence. It includes both finite (relative, adverbial, and complement clauses) and infinitival subclauses. A higher density of subclauses is indicative of greater syntactic complexity. On average, the CMC writings had a lower average no. of subordinate clauses per sentence (*M* = 0.14, *SE* = 0.02) than the school writings (*M* = 0.80, *SE* = 0.06), *t*(10) = 10.21, *p* < .001. Figure 6 clearly shows that the CMC texts overall contained fewer subordinate clauses. Again, the lower syntactic complexity of CMC is confirmed by T-Scan.

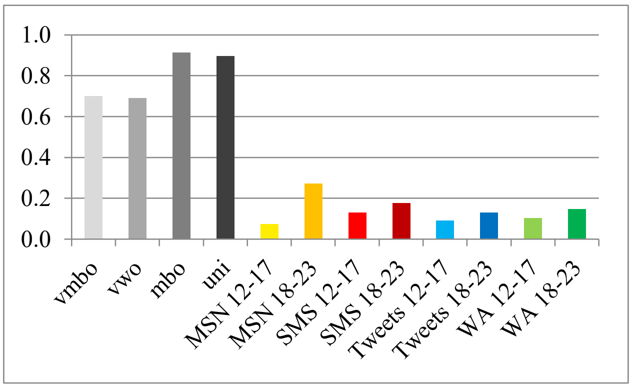


Figure 6: Average no. of subordinate clauses per sentence.

Another complexity measure provided by T-Scan is the average sentence length, which is measured in number of words. A higher average sentence length indicates more syntactic complexity. On average, the CMC writings had a lower average sentence length (*M* = 6.55, *SE* = 0.28) than the school writings (*M* = 16.33, *SE* = 0.79), *t*(10) = 14.76, *p* < .001. Figure 7 shows that the texts of all four CMC genres contained much shorter sentences than the school essays, irrespective of the writer's educational level or age. Once more, the hypothesis is confirmed.

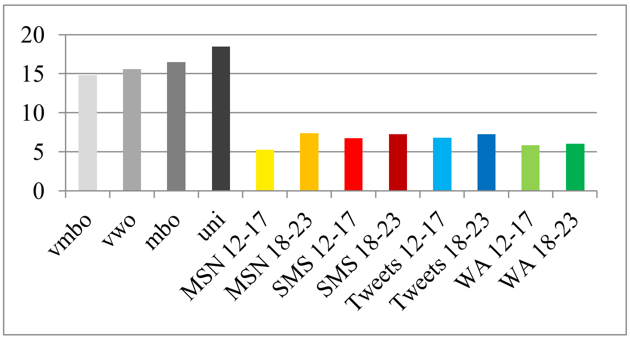


Figure 7: Average sentence length.

A final relevant syntactic measure is the so-called D-level. The D-level of a text is determined on the basis of a classification and rank order of sentence types in eight increasingly complex developmental levels, in the order in which children learn these constructions (Rosenberg & Abbeduto, 1987; Covington, 2006). The assumption is that a higher D-level value suggests more syntactic complexity. On average, the CMC writings had a lower D-level (*M* = 0.88, *SE* = 0.08) than the school writings (*M* = 2.87, *SE* = 0.10), *t*(10) = 15.51, *p* < .001. The CMC texts of all four genres had lower D-levels, as can be seen in Figure 8. This result is in line with the proposed hypothesis on syntactic complexity.

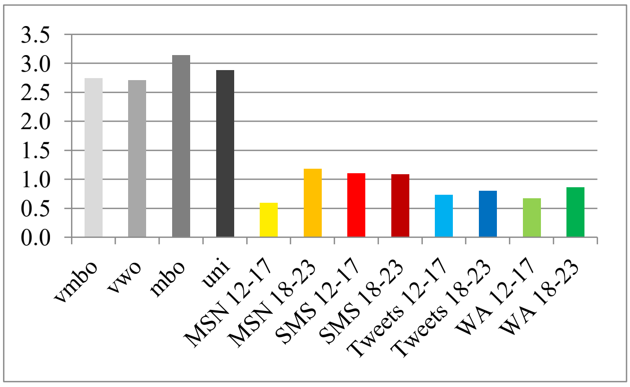


Figure 8: D-level.

5. Conclusion

To conclude, the lexical and syntactic analysis of CMC texts of four social media support my hypothesis: in comparison to school writing, CMC is lexically more diverse, different, and dense, while syntactically it contains more omissions and is less complex. This proves that Dutch youths in secondary and tertiary education employ a different register in informal computer-mediated communication than in texts written in more formal settings. These results are hopeful: perhaps deviations from the standard language in youngsters' CMC do not cause great interference with their traditional writing skills after all – they might be quite capable of keeping the registers separate, as societal norms expect them to do.

6. Future Work

A limitation of the present study is that the materials compared here, i.e. CMC discourse and texts written at school, were not produced by the same writers. In addition, they have been collected over a relatively long time span, of six years. For a more precise answer to the question if and, if so, how CMC use affects school writing, I plan to conduct research in which (a) social media data and school texts of the same students are collected and analysed and (b) additional information about writers' use of CMC and social media (in terms of frequency/intensity) are gathered through surveys. Future work will include one more genre, namely posts from the social networking site Facebook. Furthermore, it unfortunately exceeded the scope of this paper to closely examine variation between texts of different genres, educational levels, ages; this may also be explored further. Still, this study can serve as a fruitful basis for analyses on the impact of written computer-mediated communication on young people's literacy skills.

7. Acknowledgements

This study is part of a research project funded by the Dutch Organisation for Scientific Research (NWO), project number 322-70-006. I would like to thank Wilbert Spooren and the anonymous reviewers for their useful comments on previous versions of this paper.

8. References

Cougnon, L.-A., & Fairon, C., Eds. (2014). *SMS Communication: A Linguistic Approach*. Amsterdam: John Benjamins.

Covington, M.A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). *How Complex is That Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-Level Scale*. CASPR Research Report 2006-01. University of Georgia: Artificial Intelligence Center.

Crystal, D. (2008). *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

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

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. In Y. Miyashita, A.P. Marantz & W. O'Neil (Eds.), *Image, Language, Brain*. Cambridge: MIT Press, pp. 95--126.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers in Linguistics*, 53, 61--79.

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381--392.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme*. Heidelberg: Springer, pp. 219--247.

Pander Maat, H., Kraf, R., van den Bosch, A., Dekker, N., van Gompel, M., Kleijn, S., Sanders, T., & van der Sloot, K. (2014). T-Scan: A new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal*, 4, 53--74.

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8(1), 19--32.

Thurlow. C., & Brown, A. (2003). Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1.

Treurniet, M., & Sanders, E. (2012). Chats, tweets and SMS in the SoNaR corpus: Social media collection. In D. Newman (Ed.), *Proceedings of the First Annual International Conference on Language, Literature & Linguistics*. Singapore: Global Science and Technology Forum, pp. 268--271.

A Multimodal Analysis of Task Instructions for Webconferencing-supported L2 Interactions: A Pilot Study of the ISMAEL Corpus

Ciara R. Wigham,† H. Müge Satar*

†Clermont Université, Laboratoire de Recherche sur le Langage
(LRL), MSH, 4 rue Ledru, 63000 Clermont-Ferrand.

* Boğaziçi Üniversitesi
School of Foreign Languages, Bebek, 34342, Istanbul
Email: ciara.wigham@univ-bpclermont.fr, muge.satar@boun.edu.tr

Abstract

This pilot study examines how trainee language teachers use the different semiotic resources available to them during webconferencing-supported interactions to give task instructions. The sub-corpus examined is taken from the ISMAEL corpus (Guichon *et al*., 2014) that structured interaction data from a six-week telecollaborative exchange between trainee teachers of French and learners of French, who majored in Business. The study explores, firstly, how the corpus of synchronous CMC interactions was structured in order to be used by researchers who were not involved in the pedagogical project. Secondly, we will describe how the interactions were transcribed with reference to a multimodal interactional analysis approach. Thirdly, a sequential analysis of two trainee teachers' instruction-giving practices for a role-play task will be presented. The aim of the pilot study is to determine whether research and pedagogical leads emerge that warrant a larger investigation of the corpus with relation to multimodal instruction-giving practices.

Keywords: instruction-giving, LEarning and TEaching Corpora (LETEC), multimodality, teacher-training, webconferencing

1. Introduction and Research Aims

Tasks in the second language classroom allow for authentic communication with a focus on meaning (Ellis, 2003; Nunan, 2004). Alongside recent pedagogical moves towards task-based language teaching (TBLT) approaches, telecollaboration is also gaining increasing interest and research has started to explore how, by bringing together different student populations from different cultures and languages, telecollaboration can support language learning and help prepare students for physical mobility programmes, or, if involving teacher-trainee populations, prepare trainees for online mediated teaching contexts (Guth & Helm, 2010). Many telecollaboration programmes based on TBLT use synchronous means of communication to bring together the student populations that are in geographically distant locations. However, as Guichon & Cohen underline whilst "synchronicity is generally seen as bringing real value to online pedagogical interactions […], research investigating the potential of a broad array of channels has been much less frequent" (2014:332).

In any foreign-language classroom, instruction-giving is a significant part of teacher-talk time. Indeed, in TBLT, specific teacher roles include guiding and facilitating learning during task completion and explaining the purpose, expected results and task completion steps in understandable ways for learners (Raith & Hegelheimer, 2010). Although a limited number of studies have explored teachers' instruction-giving practices (see Section 2), research on instruction-giving practices in synchronous online contexts is currently non-existent.

This pilot study attempts to bridge the research gaps mentioned above by focusing on how trainee teachers of French as a foreign language give task instructions during webconferencing-supported interactions and, more specifically, how they use the multimodal semiotic resources available to them during these practices. The data examined in this qualitative study is taken from the ISMAEL corpus (Guichon *et al*., 2014) that structured the interaction data from a six-week telecollaborative exchange between undergraduate Business students learning French at an Irish higher education institution and trainee teachers on a Master's programme in Teaching French as a Foreign language at a French University. In our paper presentation, we will, firstly, examine how the corpus was structured. Then, drawing on multimodal interactional analysis and conversation analysis approaches, we will examine a sub-corpus of two trainee teachers' instruction giving practices for a role-play rehearsal task (Nunan, 2004). In particular, we examine how the trainee teachers contextualise instruction-giving sequences. The aim of the pilot study is to discern whether a larger investigation of the corpus would be pertinent and more specific research questions such a study could address.

2. Instruction-giving

Instructions are defined as directives, explanations or questions, etc. used by the teacher in order "to get the students to do something" (Watson Todd, 1997:32). Instructions could constitute such a crucial aspect of the classroom activities that successful task outcomes may depend on effective instructions (Watson Todd *et al.*, 2008). Seedhouse (2008) investigated instruction-giving practices from a conversational analysis approach, focusing on how teachers create, manage and maintain a shift in focus through the use of discourse markers, changes in the spatial configuration of participants and metadiscoursal comments. He describes how semiotic means, through the proxemics distance placed between the teacher and resources allowed a shift in focus.

Markee (2015a) examined instructions from an ethnomethodological perspective, concluding that "non-verbal aspects of communication are a vital part of instructions" (p. 126). These non-verbal aspects included gaze, cultural artifacts, gestures and embodied actions. His observation of overlaps between teacher instructions and learner responses indicated that instructions are not monologues, but they have an interactional nature. According to Markee (2015a), teachers' instructions in the classroom comprise six fragments: "(1) how [students] will be working (in dyads or small groups); (2) what resources they will need; (3) what tasks they have to accomplish; (4) how they will accomplish the task; (5) how much time they have to accomplish these tasks; (6) and why they should do something" (pp. 120-121). Markee (2015b) concluded that further research is needed on teachers' instruction-giving practices particularly in second language teaching.

Whilst Markee appears to be referring to face-to-face teaching contexts, his statement appears all the more true for computer-assisted language learning contexts as we failed to identify any studies specifically that detailed instruction-giving sequences in synchronous online pedagogical interactions. This observation was the starting point for the analysis presented in this paper.

3. Methodology

This section presents our research methodology. The corpus design will be the focus of the first part of our paper presentation.

3.1 ISMAEL Corpus and the Pedagogical Context

This study draws on the ISMAEL corpus (Guichon *et al.,* 2014) that structured data from a telecollaboration project between Business undergraduates at Dublin City University (DCU) and trainee teachers (henceforth, trainees) at Université Lyon 2 (Lyon2) on a French as a foreign language Master's programme. For the Lyon2 students, the exchange formed part of an optional module in online teaching that aims to help the trainees develop professional skills to teach French online and to analyse their online teaching practice and develop reflective analysis around this. For the undergraduate DCU students, the exchange composed part of a 12-week blended French for Business module that had CEFR level B1.2 as its minimum exit level (Council of Europe, 2001).

Participants completed six 40-minute weekly online sessions via webconferencing in autumn 2013. Two of the trainees planned each session (except the introductory session) around a theme of Business French according to the needs of DCU students as they prepare for an internship in France. Therefore, the topics for the sessions were preparing for an internship, project management, pitching a project, interviews, and labour law. The online webconferencing sessions took place on *Visu* (Guichon, Bétrancourt, & Prié, 2012) as part of a larger circular learning design (detailed in Guichon & Wigham, 2016). In this presentation, we will only draw on the data from the synchronous sessions.

Twelve of the 18 students (eight females, four males) and all of the trainees (ten females, two males) gave permission for their data to be included in the ISMAEL corpus. Thus, the corpus includes data of 7 groups. Because of differently sized groups, five groups comprised a trainee working with two learners whilst the other two groups were learner-trainee pairs. Currently, 24 of the 35 synchronous interactions included have been transcribed, totalling 13h04m30s of data. Pseudonyms are used for all personal information.

During the structuration phase of the ISMAEL corpus, the different participants' webcam videos had been extracted from the *Visu* software and imported into the transcription software *ELAN* (Sloetjes & Wittenburg 2008). The spoken interaction of all the online sessions had been transcribed and, using the timestamps created in *Visu*, the parallel text chat logs had been synchronized with these transcriptions. With regards to LEarning and TEaching Corpora (LETEC, Reffray *et al.,*2012), the learning design for the telecollaboration project, as well as documents related to the research protocol, was also available within the corpus.

3.2 Sub-corpus Examined

This preliminary study examines data from the fourth session of the telecollaboration project. During this session, participants engaged in a role-playing task that concerned project management. This task was planned in three stages. First, the trainees would introduce the roles for the learners (co-workers at McDonalds) and for themselves (manager). At this stage, learners needed to collaboratively find a new formula for children's birthday parties organized at the fast-food restaurant. During the second stage, the learners were asked to list the actions required to execute their new idea in text chat. In the final stage, the trainee (in the role of the manager) would guide a reflection session on the ideas of the learners (i.e. the employees) using questions such as: What action would you need to put into place first: which is the most important for you? Why?

A sub-corpus of the instruction-giving interaction data from two of the seven teacher trainees (Samia, Etienne) was chosen for analysis. Samia is a 23 year old female who has completed several teaching observation placements and who has experience of one-to-one tuition and some French language teaching at first school in Germany. One of her learners spoke English as his first language (Sean) whilst the other's (Angela) mother tongue was German.

Etienne is a 24 year old male who has no formal teaching experience. He had been involved in running conversation workshops in French as a foreign language at an American University over a five-month period. Etienne's leaners were Conor, who was of Irish origin and Sophie who was a Spanish speaker (L1). Neither of the trainees was involved in preparing the lesson plan for this session which had been prepared by their classmates. Samia's session lasted 35m21s whilst Etienne's session lasted 20m46s. Figures 1 and 2 give an overview of the verbal interaction data for these sessions.

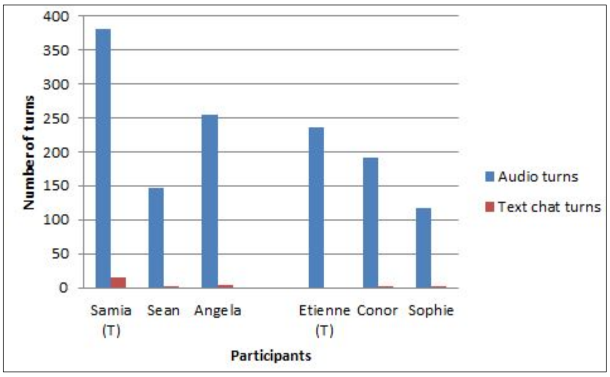


Figure 1: Overview of verbal interaction data.

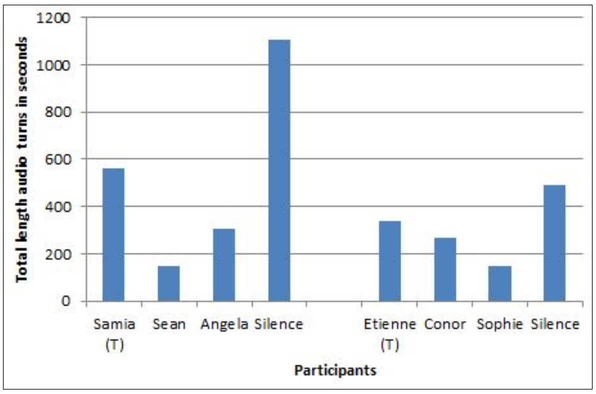


Figure 2: Total length audio turns.

3.3 Analysis approach and procedures

Data for this presentation was analysed using multimodal interactional analysis (Norris, 2004) which aims to explore people's meaning-making practices in the moment-by-moment construction of interaction with an emphasis on "how people employ gesture, gaze, posture, movement, space and objects to mediate interaction in a given context" (Jewitt, 2011: 34). For the verbal data, we also make use of conversation analysis techniques. The initial step in the analysis was to identify instruction-giving sequences for the role-playing task by isolating trainees' transition into the task and the several fragments that were introduced to cover all aspects of the instructions. The second analysis step was the annotation of the co-verbal acts that accompanied task instructions. The co-verbal actions included gaze, facial expressions, head movements, gestures and distance between the webcam and the participant.

It is worth noting that the approach to the analysis of the sub-corpus involved a researcher who was closely in the data collection, data transcription and the structuration of the corpus and an 'outsider' who did not know the participants and the context (cf. Guichon, in print). Both researchers worked on the sub-corpus together, constantly comparing their interpretations of the data and how the instruction-giving sequences were organised. We will briefly touch on the advantages and disadvantages of data analysis that involves 'insider' and 'outsider' researchers.

4. Preliminary Findings

In the second part of our paper presentation, we will look closely at the interaction data and will present a sequential analysis of each of the two instruction-giving sequences. Due to space constraints, it is not possible here to go into depth concerning the micro-analysis conducted. Rather, we summarise the analysis of each case.

The analysis of the instruction-giving sequence in the session conducted by Samia shows a clear step-by-step approach to instruction giving. Gaze plays an important role in punctuating these steps.

Samia combines the audio and text chat modalities to elicit key vocabulary for the task and concept check these items.

Gaze shifts, accompanied by vocatives play an important part in assigning learner roles. Samia then makes use of the visual mode to communicate, through a change in proximity, that she is giving greater control of the floor to learners as they begin the task and, thus, that she wishes to step out of her interaction management role. A shift in pronoun use to the inclusive 'we' also allows her to show verbally that she has moved into the fictitious role of manager rather than the managerial role of task instruction-giver.

In contrast, in the analysis of Etienne's instruction-giving sequence, the trainee first of all sets the context for the task by checking the concept of children's birthday parties and then proceeds by indicating his role and providing examples of possible themes. This helped learners identify what constitutes the trainee's expectations concerning successful task completion. However, as they had not yet been given their roles some confusion ensues. Learner role allocation was achieved through a side-sequence during a long task-preparation phase rather, as was the case with Samia, as a main step in the instruction-giving process. The trainee's multimodal interaction during this phase is of particular interest. In the visual mode he attempts to remove his presence from the interactional order through a change in posture and proximity, underlining that this is an individual-work phase. Gaze change during this preparation phase allows the trainee to monitor whether he has covered all of the information points that are necessary for the task and prompt Etienne to introduce a side sequence in which he allocates learner roles. The laughter and posture change that follow help to signal the learners' better understanding of the task instructions.

5. Discussion

Our initial analysis suggests that in order to draw pedagogical conclusions, it would be of particular interest to further examine instruction-giving sequences with reference to how the beginning and different stages of the task instructions are marked; how the trainees allocate roles required by the task during these sequences and trainees deal with key lexical items.

With reference to these points, the data examined in this pilot investigation suggests, firstly, that changes in proximity to the webcam may be a successful technique to highlight changes in role and show learners that the trainee is moving into his fictional role required by the task. Secondly, the multimodal analysis sheds light on different strategies employed by the trainees to introduce vocabulary for the task. Whilst Samia used elicitation to concept-check key vocabulary that she often then put into the text chat modality, Etienne preferred to use pre-emptive vocabulary explanation to establish the context for the task and used reduced proximity to signal when he was willing to leave the floor/interactional order.

Thirdly, combining vocatives in the audio modality and gaze in the visual mode appeared effective in role allocation whilst the other session demonstrates what happens when task instructions, especially role allocation are not complete and how the resulting confusion and uneasiness can be resolved.

The presentation will conclude with pedagogical recommendations highlighting the need to raise teacher trainees' awareness of the multimodal features of webconferencing that can be employed to facilitate instruction-giving.

6. Acknowledgments

The authors are grateful to the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon for its support within the program 'Investissements d'Avenir' (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR).

7. References

Ellis, R. (2003). *Task-based language learning and teaching.* New York: Oxford University Press.

Guichon, N. (in print). Sharing a multimodal corpus to study webcam-mediated language teaching. *Language Learning & Technology.*

Guichon, N., Bétrancourt, M., Prié, Y. (2012). Managing written and oral negative feedback in a synchronous online teaching situation. *Computer assisted language learning*, 25(2), 181–197.

Guichon, N., Blin., F., Wigham, C.R., & Thouësny, S. (2014) *ISMAEL LEarning and TEaching Corpus.* Dublin, Ireland: Centre for Translation and Textual Studies & Lyon, France: Laboratoire ICAR.

Guichon, N. & Cohen, C. (2014). The Impact Of The Webcam On An Online L2 Interaction. *Canadian Modern Language Review.* 70(3), 331–354.

Guichon, N. & Wigham, C. R. (2016). A semiotic perspective on webconferencing-supported language teaching, *ReCALL*, 28(1), 62-82.

Guth, S. & Helm, F. (2010). *Telecollaboration 2.0.* New York:Peter Lang.

Jewitt, C. (2011). Different approaches to multimodality. In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis*, (pp. 28-39). London: Routledge

Markee, N. (2015a). Giving and following pedagogical instructions in task-based instruction: An ethnomethodological perspective. In P. Seedhouse and C. Jenks (Eds.) *International Perspectives on the ELT Classroom,* (pp.110-128). Basingstoke: Palgrave MacMillan.

Markee, N. (2015b). Teachers' instructions: Toward a collections-based, comparative research agenda in classroom conversation analysis. Paper presented at *HUMAN Social Interaction and Applied Linguistics Postgraduate Conference*, 08 September 2015, Hacettepe University, Ankara. [https://sial2015hu.files.wordpress.com/2015/09/1-ankara-paper-final.pdf]

Norris, S. (2004). *Analyzing multimodal interaction: a methodological framework*. London: Routledge.

Nunan, D. (2004). *Task-Based Language Teaching.* Cambridge: Cambridge University Press.

Raith, T. & Hegelheimer, V. (2010). Teacher Development, TBLT and Technology. In M. Thomas & H. Reinders (Eds.), *Task-Based Language Learning and Teaching with Technology*, (pp.154-175). London: Continuum.

Reffay, C., Betbeder, M-L. & Chanier, T. (2012). Multimodal learning and teaching corpora exchange: lessons learned in five years by the Mulce project', *Int. J. Technology Enhanced Learning*, 4(1/2), 11–30.

Seedhouse, P. (2008) Learning to Talk the Talk: Conversation Analysis as a Tool for Induction of Trainee Teachers. In Garton, S. & Richards, K. (eds). *Professional encounters in TESOL: discourses of teachers in training* (pp.42-57). Basingstoke: Palgrave Macmillan.

Sloetjes, H. & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008).

Watson Todd, R. (1997). *Classroom Teaching Strategies*. London: Prentice Hall.

Watson-Todd R, Chaiyasuk I, and Tantisawatrat N (2008) A functional analysis of teachers' instructions. *RELC Journal*, 39, 25-50.

Linguistic Analysis of Emotions in Online News Comments - an Example of the Eurovision Song Contest

Ana Zwitter Vitez,[+*] **Darja Fišer**[*†]

[+] Faculty of Humanities, University of Primorska, Titov trg 5, 6000 Koper, Slovenia
[*] Department of Translation, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia
[†] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: ana.zwitter@guest.arnes.si, darja.fiser@ff.uni-lj.si

Abstract

The aim of the study is to identify linguistic differences of positive and negative comments on the example of online comments on the news about the Eurovision song contest. The results show that positive comments have a typical exclamation form and simple sentence structure, and include more informal vocabulary and orthographic variation. Negative comments, on the other hand, are more likely to be formulated as statements with a complex syntax structure and with a neutral vocabulary and standard orthography. The detected differences can be explained by the communicative function of the negative comments that act as reviews and therefore call for thorough argumentation and build an individual's reflective identity.

Keywords: online news comments, linguistic analysis, sentiment analysis, Eurovision song contest

1. Introduction

Identifying emotions in language is a relevant field of research because of the strong connection between the physiological arousal of an emotion and its social display (Mygovych, 2013). If we understand how people feel, we can analyse or even predict how they will react in certain situations. This is why sentiment analysis can be used for predicting societal changes, election results, and customer satisfaction (Liu, 2015).

In online comments, analysis of emotions is particularly interesting because comments enable users to formulate their own opinion and to find their own identity independently of the official media content. Wright (2009)[1] even claims that "for many businesses, online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace".

Online news comments of the Eurovision song contest represent a specific dataset because they usually evoke polarised emotions (either users strongly support or hate Eurovision song contestants and/or their songs). The comments often even exceed the scope of the song contest itself and refer to wider political and societal issues (e.g. *Azerbejdžan podeli Rusiji 12 točk in si s tem zagotovi dostavo plina za še eno leto #Eurovision. / Azerbaijan gives Russia 12 points, thus guaranteeing their gas supply for another year #Eurovision*).

The aim of this paper is to provide a linguistic analysis of online comments in order to detect syntactic, lexical and orthographic differences between positive and negative comments.

2. Emotion Analysis in CMC

Different approaches have been developed to analyse emotions in computer-mediated communication. In data mining three basic categories are most often used: positive, negative, neutral (Smailović, 2013). These models are very useful on big datasets to study overall trends but are not always reliable for linguistic analyses of individual texts.

In discourse analysis, much more fine-grained sentiments are typically examined: happiness (Stefanowitch 2004), shame (Retzinger, 1991), and even irony (Haverkate, 1990). These approaches are very interesting for qualitative analyses but cannot be scaled for emotion identification on bigger datasets.

In this paper we focus on a qualitative analysis of a small dataset of news comments on the Eurovision song contest in Slovenian in order to examine linguistic characteristics of opinionated texts. Once comments were manually attributed a sentiment category, they were analysed on the syntactic, lexical and orthographic level.

3. Methodology

3.1 Sample Creation

The analysis was performed on a sample extracted from the Janes corpus v0.4 (Fišer et al., 2016). The sample contains 70 comments referring to an article announcing that the Slovenian representative was selected to compete in the finals of the Eurovision song contest[2] published on the national television and radio online news portal RTV Slovenija. Only opinionated comments were taken into account for the study. Neutral, factual and objective comments (e.g. *Bjørn Einar Romøren*) were discarded as were off-topic comments or direct replies to a previous comment that were part of an internal debate that had nothing to do with the article they appeared under (e.g. *Kje je kolega XX? Upam, da ni zaspal! / Where is our camerad XX? I hope he hasn't fallen asleep!*).

3.2 Sentiment Annotation

First, comments were manually attributed a sentiment category (positive, negative) by two annotators. Disagreements were detected in 6 cases (8%), which were discussed in order to reach a systematic final decision. 4

[1] URL: http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html

[2] URL: http://www.rtvslo.si/zabava/glasba/evrovizija-2014/slovenija-s-tinkaro-v-evrovizijski-finale-vidimo-se-v-soboto/336356

of the cases involved comments which consisted of two parts, expressing a different sentiment each (e.g. *Tinkari pa zaželim srečo,četudi je ta Evrovizija zadnjih 10 let z uvedbo polfinalov čisti cirkus in šov,ki ga prav tako dolgo ne jemljem več tako resno. / I wish Tinkara all the best, even though for the past 10 years the Eurovision and its semi-finals have been nothing but a circus and a show that I am not taking seriously anymore*.). In such cases, the annotators agreed to determine the prevalent sentiment in the comment. In 2 of the cases, it was not clear out of context whether the comments were meant literally, as a joke or cynical (e.g. *Pričakujem 12 točk iz Makedonije. / I am expecting 12 points from Macedonia.*). In such cases, the entire discussion thread was examined for a wider context and annotated accordingly.

3.3 Linguistic Analysis

Each sentence in the sample was analysed for sentence type (statement, exclamation question, order), sentence structure (simple, complex), vocabulary characteristics and orthography (formal, informal). Examples of the analysis are presented in Tables 1 and 2.

SLO: ***Držim pesti!*** **ENG:** ***Fingers crossed!***	
Sentiment	positive
Sentence form	exclamation
Sentence structure	simple
Vocabulary	/
Orthography	informal

Table 1: Linguistic analysis of a positive comment.

SLO: ***Kolikor slišim, se je včeraj slabo odrezala.*** ***ENG: As far as I heard, she did not fare well last night.***	
Sentiment	negative
Sentence form	statement
Sentence structure	complex
Vocabulary	slabo
Orthography	standard

Table 2: Linguistic analysis of a negative comment.

4. Results and discussion

Comments with the same sentiment label were compared in order to detect the shared linguistic properties on the syntactic, lexical and orthographic level. The sample contains slightly more negative (53%) than positive (47%) comments.

4.1 Syntax

As can be seen from Figure 1, a large majority (86%) of the negative comments are statements (e.g. *Ne, ne bo. / No, it won't.*) with only a few examples of questions (8%) and exclamations (5%). Among the positive comments, on the other hand, there is a similar share of statements (48%) and exclamations (45%) (e.g. *Srečno! / Good luck!*). While positive comments contain no questions, there are a couple of commands (6%) (e.g. *Uživajmo in ne nergajmo kot stare babe. / Let's enjoy the show and not whine like old ladies.*).

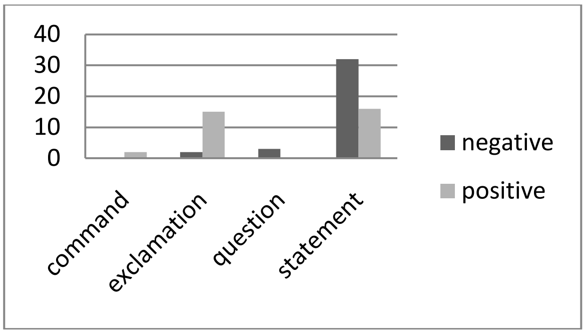


Figure 1: The distribution of different types of sentences in positive and negative comments.

The majority of the sentences in the negative comments are complex (62%) while nearly half of the sentences in the positive comments (49%) are simple. For illustration, Figure 2 contains examples of a complex negative sentence and a simple positive one.

Negative, complex:
SLO: Če zaupaš našim medijem, so še skoraj vsako leto bile kritike glede naših pesmi pozitivne, ampak rezultata pa nobenega in isto bo letos.
ENG: According to our media, despite positive reviews our songs were unsuccessful almost every year and this year won't be any different.
Positive, simple:
SLO: Imam dober občutek.
ENG: I have a good feeling about this.

Figure 2: Examples of simple and complex sentences in positive and negative comments.

4.2 Vocabulary

The vocabulary level was manually annotated following the criterion of whether a comment is characterised by a specific lexical unit carrying an opinion or not.

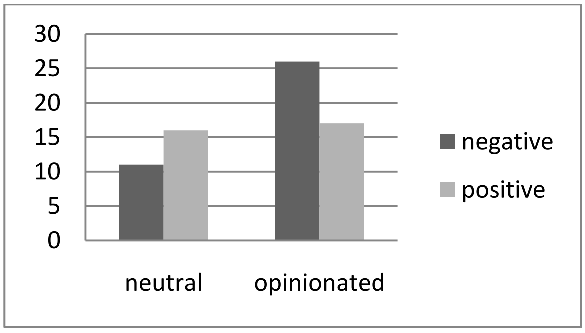


Figure 3: The distribution of neutral and opinionated vocabulary in positive and negative comments.

As Figure 3 shows, vocabulary in the negative comments is heavily opinionated (70%), e.g. *kuhna (inside deal), davkoplačevalci (taxpayers), lajna (broken record)*. About half of the positive comments are characterised by opinionated vocabulary (51%), but this vocabulary is not topic-specific and usually expresses general support, e. g. *Srečno! (Good luck!), upam (I hope), podpiramo (we support*).

4.3 Orthography

At the orthographic level the following phenomena that are typical of CMC language were observed: informal orthography (e.g. *dejmo* instead of *dajmo*), use of all-caps (e.g. *BO*), non-standard use of punctuation (e.g. *Dajmo klobasica!!!:)* / *Do it*), and emoticons *(;))*. As can be seen in Figure 4, while distinctly standard orthography (78%) is used in the negative comments, nearly half of the positive comments (42%) contain non-standard orthographic features.

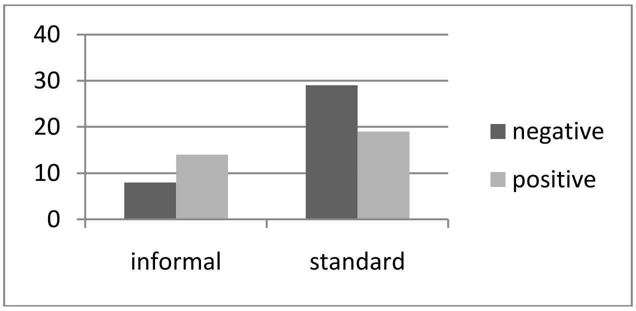


Figure 4: The distribution of non-standard and standard orthography in positive and negative comments.

For illustration, Figure 5 contains an example of a positive comment with standard orthography and an example of a positive comment with non-standard spelling.

Negative, non-standard:
SLO: *dejmo naši!!!*
ENG: *c'mon, team!!!*
Positive, standard:
SLO: *Danes imajo naši turisti še zadnji dan turistovanja na davkoplačevalske stroške*.
ENG: *Today is the last vacation day attaxpayers' expense for our tourists.*
Figure 4.

Figure 5: Examples of standard and non-standard orthography in positive and negative comments.

5. Conclusions

The aim of the study was to identify linguistic characteristics of positive and negative comments on the example of comments on articles about the Eurovision song contest. The results show that positive comments are fewer, typically have an exclamation form and simple sentence structure, and contain more informal vocabulary and orthography. Negative comments are more numerous, are typically represented as statements with a more complex syntax structure and with distinctly general vocabulary as well as standard orthography.

While it is true that the analysed sample is small and limited to a single topic, the results are very homogenous and consistent throughout the analysis. A plausible explanation for such a discrepancy is the critical function of the negative comments which calls for thorough argumentation, not affect, and from the position of a reflective individual who acts in his own capacity, not as a member of regional or social groups adherence to which usually shows through the use of typical vocabulary and orthography.

Our future work plan is to extend the analysis on a wider range of highly opinionated topics (sports, politics, religion, product and service reviews) and text types (blogs and blog comments, tweets, forum posts, Wikipedia talk pages). In addition, the set of sentiments will be more fine-grained in order to distinguish between different types of negative or positive sentiment such as support and cynicism that deserve special treatment.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Non-standard Internet Slovene" (J6-6842, 2014–2017).

7. References

Fišer, D., Erjavec, T., Ljubešić, N. (2016): JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2), 67–100.

Haverkate, H. (1990). A speech act analysis of irony. *Journal of Pragmatics,* 14(1), 77–109.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press.

Mygovitch, I. (2013). Secondary nomination in the modern English language: affective lexical units. *Visnik LNU imeni Tarasa Ševčenka,* 1(1), 206–214.

Ritchie, G. (2004). The Linguistic Analysis of Jokes. *Journal of Literary Semantics,* 33(2), 196–197.

Smailović, J., Grčar, M. Lavrač, N., Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences* 285: 181–203.

Stefanowitsch, A. (2004). Happiness in English and German: A metaphorical-pattern analysis. In M. Achard and S. Kemmer (eds.) *Language, Culture, and Mind*, 137–149. CSLI Publications.

Retzinger, M. S. (1998). *Violent Emotions: Shame and Rage in Marital Quarrels.* Newbury Park, CA: Sage Publications.

Alternative Endings of Slovene Verbs in Third Person Plural: A Corpus Approach

Gašper Pesek, Iza Škrjanec, Dafne Marko
Ljubljana
E-mail: gasper.pesek@gmail.com, skrjanec.iza@gmail.com, dafne.marko@gmail.com

Abstract

This paper is concerned with the alternative endings of some Slovene verbs in third person plural. Certain Slovene verbs in this form can take the endings *-jo* or *-do* (*jejo* or *jedo*), whereby the two possible choices are normatively evaluated in various ways in Slovene language manuals. The paper introduces a corpus-driven approach to the problem by first extracting potential verbs with this phenomenon, after which the use of both endings is compared in the subcorpora of Janes (a corpus of Slovene user-generated content) and in the Kres corpus.

Keywords: alternative endings, Slovene verbs, user-generated content, standard language

1. Introduction

Slovene is a morphologically rich language with some alternative forms in the inflection paradigm. Related studies using a corpus approach (Može, 2013; Arhar Holdt, 2013) show that Slovene language manuals either insufficiently describe the phenomena in question, or prescriptively evaluate one variant as more prestigious than the other. This paper is concerned with Slovene verbs that can take the endings *-do* or *-jo* in third person plural. We observe their behavior in the Janes corpus of user-generated Slovene (Fišer et al., 2016) and in the Kres corpus (Logar Berginc et al., 2012), which mostly contains standard written Slovene.

The rest of the paper is structured as follows: in Section 2, the alternative verb endings *-jo* and *-do* are presented together with their evaluation in different Slovene language manuals. In Section 3, we briefly present the two corpora and the data extraction process. The verb forms and their frequencies in the Janes subcorpora and in the Kres corpus are analyzed in Section 4, while Section 5 provides a genre comparison. Section 6 concludes the paper and suggests further work.

2. Problem Description and the Aim of the Paper

Certain Slovene verbs can take two different endings in third person plural: *-jo* or *-do*. In Slovene grammar (Toporišič, 2004), this characteristic is observed in five athematic verbs (*jejo* – *jedo*; *grejo* – *gredo*; *bojo* – *bodo; vejo* – *vedo; dajo* – *dado*; 'they eat/go/will be/know/give', respectively). According to paragraph 891 in the Slovene normative language manual *Slovenski pravopis 2001*, the ending *-jo* is frequently used instead of *-do*, which is especially true for "literary conversational language" (*knjižni pogovorni jezik*) – considered less appropriate for written texts – and derivatives (by means of prefixation) of the previously mentioned athematic verbs, e.g. *prepovejo* – *povedo,* where the derivative with the prefix *pre-* is written with the ending *-jo*, whereas the original verb is written with the ending *-do*. The choice between two alternative forms seems to be puzzling for the users of Slovene, as can be seen from several questions posted in specialized[1] as well as general[2] online forums, such as Med.Over.Net. In the replies, two main factors for the use of appropriate endings are emphasized: the medium (spoken or written) and register (literary or conversational). The ending *-do* is considered more common in written language and formal communication.

In this regard, an analysis of the phenomenon in computer-mediated communication or user-generated content would be interesting, as the language on the Internet is heavily influenced by spoken language (Crystal, 2006); however, a spectrum of genres are considered as CMC, differing in synchronicity, message size, privacy settings, communication norms, etc. (for a list of factors, see Herring, 2007).

The aim of this paper is to use a corpus-based approach to determine the general tendencies of using the endings *-do* and *-jo*. We expect to find a preference for the ending *-jo* in the Janes corpus and a preference for the ending *-do* in the Kres corpus. We intend to observe the usage patterns with regard to different genres in Janes, and to contrast the tendencies in the Janes corpus with those in the Kres corpus, which is a corpus of standard written language.

3. Methodology

3.1 Corpus Description

For the analysis, two corpora were queried. The Janes v0.4 corpus is a corpus of user-generated Slovene. It contains over 175 million words or 9 million documents, published between 2002 and 2016 in five different genres: tweets, forum posts, blog entries and their comments, online news and their comments, and user and page talk from the Slovene Wikipedia.

[1] E.g.: http://www2.arnes.si/~lmarus/suss/arhiv/suss-arhiv-000103.html (accessed: 10 August, 2016)

[2] E.g.: http://med.over.net/forum5/viewtopic.php?t=2547265, http://med.over.net/forum5/viewtopic.php?t=10424263 (accessed: 10 August, 2016)

The Kres corpus, which contains nearly 100 million words, is a collection of standard written Slovene with a balanced genre structure: it consists of periodicals (newspapers and magazines), fiction and non-fiction, documents from the Web, and other genres, published between 1990 and 2011.

3.2 Data Extraction

The Sketch Engine concordance (see Kilgarriff, 2014) was used for corpus scanning. Employing a CQL expression, we used the Kres corpus to first extract verbs that end in *-do*, after which we noted the frequency of the forms with *-do* and *-jo*[3] in the five Janes subcorpora[4], as well as the entire Janes and Kres corpora. Thus, we collected 17 verbs[5]: *biti* ('be'), *iti* ('go'), *dati* ('give'), *vedeti* ('know'), *izvedeti* ('find out'), *zvedeti* ('find out'), *poizvedeti* ('inquire'), *zavedeti* ('realize'), *jesti* ('eat'), *pojesti*[6] ('eat up'), *najesti* ('sate'), *povedati* ('tell'), *izpovedati* ('confess'), *dopovedati* ('get across'), *napovedati* ('predict'), *odpovedati* ('cancel'), and *prepovedati* ('forbid').

4. Analysis

4.1 Distributions of Variants in the Janes Subcorpora

Because Slovene's 5 athematic verbs are a linguistically unique group, they call for a separate initial analysis. With regard to the future tense form of the verb *biti*, all subcorpora indicate a virtually exclusive preference (nearly 100% of concordances) for the ending *-do*. The opposite is true of *dati*, where *-jo* has almost completely replaced its older alternative[7]. In the case of *vedeti*, *-do* is preferred in all subcorpora – least prominently in the forum subcorpus, and most prominently in the blog (nearly 90% of concordances) and Wiki talk subcorpora (over 90%)[8]. *Iti* displays a relatively even distribution between the two alternatives, with the exception of the blog subcorpus, where a preference for *-do* is observed (roughly 80%). *Jesti* has revealed a general preference for the *-do* ending, with a relatively equal distribution in the tweet subcorpus, a slight preference for *-do* in the forum subcorpus, and a strong preference for *-do* in the comment (roughly 70%), blog (roughly 80%), and Wiki talk (roughly 85%) subcorpora. The following paragraphs summarize the specifics of each subcorpus.

In the tweet subcorpus (Figure 1[9]), the *-do* ending is preferred (60% or more) for *najesti* and *povedati*; *-jo* is preferred for *izpovedati*, *napovedati*, *odpovedati*, and *prepovedati*. All other verbs display a relatively equal distribution of both endings.

The forum subcorpus (Figure 2) reveals a considerable preference (roughly 70%) for the *-do* ending with the verb *zvedeti*. A relatively equal distribution of both endings is observed with *izvedeti*, *zavedeti*, *pojesti*, and *povedati*. All other verbs show a preference (60% or more) for *-jo*, especially in the case of *izpovedati* and *dopovedati*, which have only produced concordances with *-jo*.

In the blog subcorpus (Figure 3), there is an overall preference for *-do*. A relatively even distribution of both endings can be seen for *poizvedeti*, *zavedeti*, *najesti*, *izpovedati*, *odpovedati*, and *prepovedati*. *Napovedati* shows a notable preference (over 60%) for *-jo*.

The news comment subcorpus (Figure 4) reveals a notable preference (60% or more) for *-do* in the cases of *izvedeti* and *povedati*, while *zvedeti* and *poizvedeti* have only provided concordances with *-do*. *Zavedeti*, *pojesti*, and *izpovedati* have shown a relatively equal distribution of the two endings, whereas the others (*najesti*, *napovedati*, *odpovedati*, and *prepovedati*) have shown a notable preference for *-jo*. *Dopovedati* did not produce any concordances.

In Wiki talk (Figure 5), the verbs that have produced concordances show a strong preference for the *-do* ending, with the exception of *povedati* and *izvedeti*, which have displayed a relatively even distribution of the two ending variants.

4.2 The Janes Subcorpora in Relation to Kres

This subsection examines each of the Janes subcorpora in relation to Kres. The 5 athematic verbs will once again be described separately.

In all of the Janes subcorpora, the ratios of *-jo* to *-do* for the verb *biti* are virtually the same as in Kres. For *iti*, on the other hand, all Janes subcorpora display a much stronger preference for *-jo*, except for the blog subcorpus, as its ratio of *-jo* to *-do* is practically the same as in Kres. *Dati* almost exclusively employs the *-jo* ending in both Kres as well as all of the Janes subcorpora. For *vedeti*, the tweet, forum, and news comment subcorpora prefer *-jo* compared to Kres, whereas the blog and Wiki talk subcorpora show ratios similar to the ones in Kres (which shows a slightly stronger preference for *-do*). Finally, *jesti* displays a notable preference for *-jo* in the tweet, forum, and news comments subcorpora. There is an overlap with the Kres ratio in the Wiki talk subcorpus, and a slight preference for *-jo* with the ratio in the blog subcorpus.

The distributions of alternate endings for the verbs *poizvedeti*, *zavedeti*, and *dopovedati* in the tweet subcorpus are very similar to those in Kres. With *najesti*, however,

[3] CQL expression for verb extraction:
[word=".+do" & tag="G.*"]
CQL expression for form frequency, e.g., *bojo*:
[word="(b|B)ojo" & tag="G.*"]

[4] In the news subcorpus, only the comment section was taken into consideration, excluding the news because they are not user-generated content.

[5] The verb *zajesti* was also extracted from the Kres corpus, but was not included in the analysis due to low or zero frequencies in the Janes subcorpora.

[6] Because the third person plural forms of the verbs *pojesti* ('eat up') in *peti* ('sing') overlap (*pojejo* – 'they eat/sing'), all concordances were analyzed manually.

[7] Since the word form *dado* is used mostly as a proper noun (but has been erroneously annotated as a verb form), all concordances were analyzed manually.

[8] There is an overlap between the word forms *vedo* (third person plural) and *vedo* (a participle in masculine singular form). The latter is a regional, non-standard variation of the participle *vedel*, used mainly in northeastern Slovenia. Since all derivatives of the verbs *vedeti*, *jesti* and *povedati* follow the same pattern, there might be a slight deviation from the actual frequency of verbs ending in *-do* for third person plural.

[9] For charts, see Appendix 1.

there is a higher preference for *-do* in the tweets. All other verbs show a notable to strong preference for *-jo* in comparison with Kres.

In the forum subcorpus, all remaining verbs show a higher preference for *-jo*, albeit in varying degrees.

The verbs in the blog subcorpus display a slightly higher preference for *-jo* with the verb *zavedeti*, and a stronger preference for *-jo* with *zvedeti*, *pojesti*, *najesti*, and *napovedati*. The ratios for *poizvedeti*, *povedati*, and *odpovedati* are similar to the ones in Kres, while the remaining verbs (*izvedeti*, *izpovedati*, *dopovedati*, and *prepovedati*), seem to prefer *-do* more than Kres.

In the news comment subcorpus, *pojesti*, *najesti*, *povedati*, *napovedati*, and *odpovedati* show a higher preference for *-jo*. The ending ratios for *izvedeti*, *zavedeti*, *izpovedati*, and *prepovedati* are almost the same as the ones in Kres. Interestingly, *zvedeti* and *poizvedeti* show a higher preference for *-do*.

In the Wiki talk subcorpus, there is a slightly higher preference for *-jo* with *izvedeti* and a notably higher preference for *-jo* with *povedati*. *Pojesti*, however, shows a slightly higher preference for *-do*, while *prepovedati* shows a notably higher preference for *-do*.

5. Genre Comparison

To be able to draw more general conclusions, this section considers only the verbs with a minimum frequency of 10 in all corpora, excluding the Wiki talk subcorpus due to its small size[10]. The following verbs meet this criterion: *biti*, *iti*, *dati*, *vedeti*, *izvedeti*, *zvedeti*, *jesti*, *pojesti*, *povedati*, *odpovedati*, and *prepovedati*.

Having compared all of the genres with one another, we have identified three recurring patterns. In all (sub)corpora, the verb *biti* is predominantly used with the *-do* ending, while the verb *dati* is practically never used with the *-do* ending anymore. Taking into account all the remaining verbs, the blog subcorpus and the Kres corpus generally display comparable tendencies with eight of them (*iti*, *vedeti*, *izvedeti*, *jesti*, *pojesti*, *povedati*, *odpovedati*, and *prepovedati*), whereby the verbs in question show a similar preference for *-do* in some cases, and equal shares of both endings in others.

The second evident pattern is the similarity of the forum, tweet, and news comment subcorpora, as they contain a similar distribution of the endings in the verbs *iti*, *jesti*, *pojesti*, *povedati*, *odpovedati*, and *prepovedati*. In the case of two verbs (*vedeti* and *izvedeti*), the tweet and forum subcorpora show a similar tendency of an equal distribution for both endings.

6. Conclusion

The paper describes the use of endings *-jo* and *-do* in certain Slovene verbs in third person plural. The corpus analysis shows a strong preference for *-do* both in the Kres corpus (12 out of 17 verbs) and in the Janes corpus (10 out of 17 verbs). However, different verbs display completely different tendencies in the analyzed subcorpora, meaning that a general conclusion concerning the use of endings *-jo* and *-do* cannot be made. The verbs *biti* and *dati* proved to be the only ones with the same pattern – *biti* is almost exclusively realized as *bodo*, and *dati* as *dajo*. Regarding the genres in the Janes corpus (tweets, forum posts, blog entries, news comments and on Wikipedia discussions and user pages), a conclusion can be made that there are evident similarities between the blog subcorpus and the Kres corpus, while the tweet, forum, and news comment subcorpora display fairly comparable tendencies as well. Thus, it is important to emphasize that different genres in CMC may not always have the same linguistic characteristics and should therefore not be understood as a homogenous language variety. For a more precise description of the phenomenon, other derivatives (e.g., *spovedati*, *zapovedati*), which were not extracted in the automatic process, should also be included into discussion. The use of the endings *-jo* and *-do* should also be analyzed from a normative perspective, taking into account the language manuals and dictionaries with their descriptions of the analyzed verbs.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014–2017). We would also like to thank the reviewers for their useful comments and suggestions.

8. References

Arhar Holdt, Š. (2013). Študentje, škratje in nadškofje: končnica *-je* v imenovalniku množine pri samostalnikih prve moške sklanjatve. *Slovenščina 2.0*, 1(1), pp. 134–154.

Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

Fišer, D., Erjavec, T., Ljubešić, N. (2016). Janes v0.4: korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* (to appear).

Herring, S.C. (2007). A faceted classification scheme for computer-mediated discourse, *Language@Internet*: http://www.languageatinternet.org/articles/2007/761 (accessed: 13 August 2016).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt Š. and Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; FDV.

Može, S. (2013). Raba kratkega nedoločnika: korpusni pristop. *Slovenščina 2.0*, 1 (1), pp. 155–175.

SP 2001 – Slovenski pravopis. Ljubljana: SAZU – ZRC SAZU – Založba ZRC.

Toporišič, J. (2004). *Slovenska slovnica*. Maribor: Obzorja.

[10] For absolute and relative frequencies of verb forms, see https://www.dropbox.com/s/ducs6y7ei75vn9p/Abs_rel.xlsx?dl=0.

Appendix 1[11]

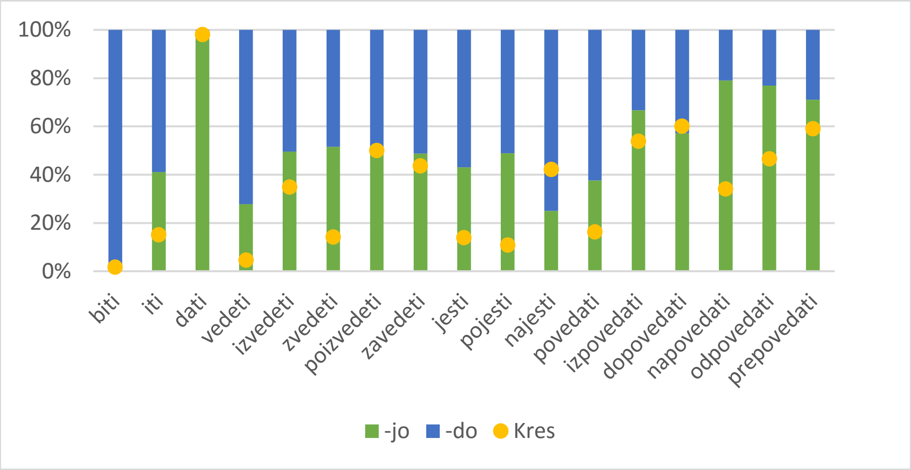


Figure 1: Percentage of *-jo*/*-do* in the tweet subcorpus (green/blue) and the Kres corpus

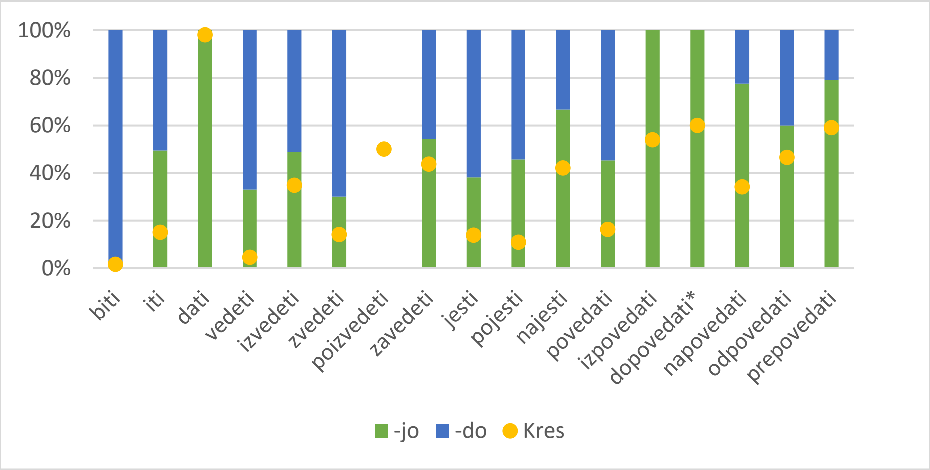


Figure 2: Percentage of *-jo*/*-do* in the forum subcorpus (green/blue) and the Kres corpus (yellow).

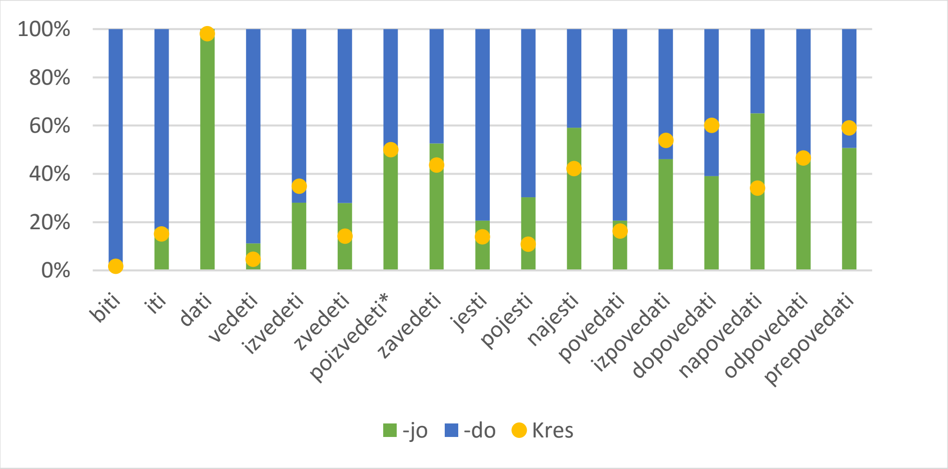


Figure 3: Percentage of *-jo*/*-do* in the blog subcorpus (green/blue) and the Kres corpus (yellow).

[11] The asterisk (*) indicates that one or both verb forms appeared only once. Empty columns indicate that the forms for a particular verb were not found in the Janes corpus. The yellow dot defines the limit between percentage for *-jo* (below it) and *-do* (above it).

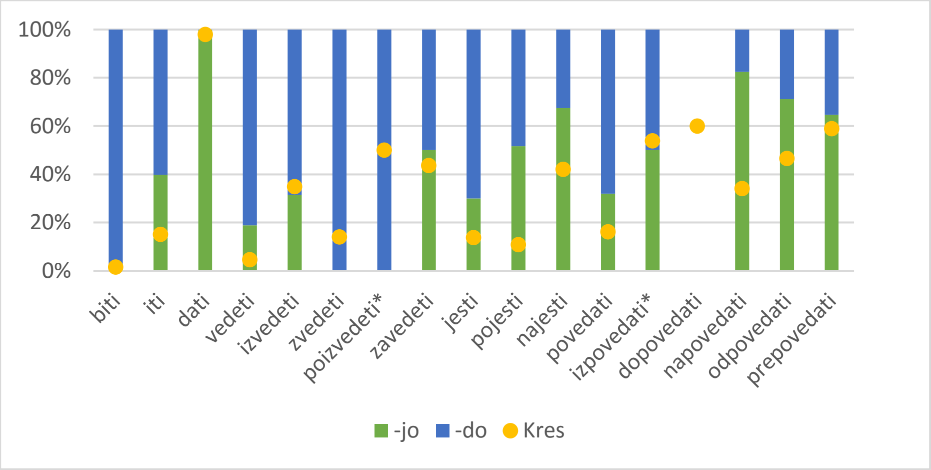


Figure 4: Percentage of *-jo*/*-do* in the news comment subcorpus (green/blue) and the Kres corpus (yellow).

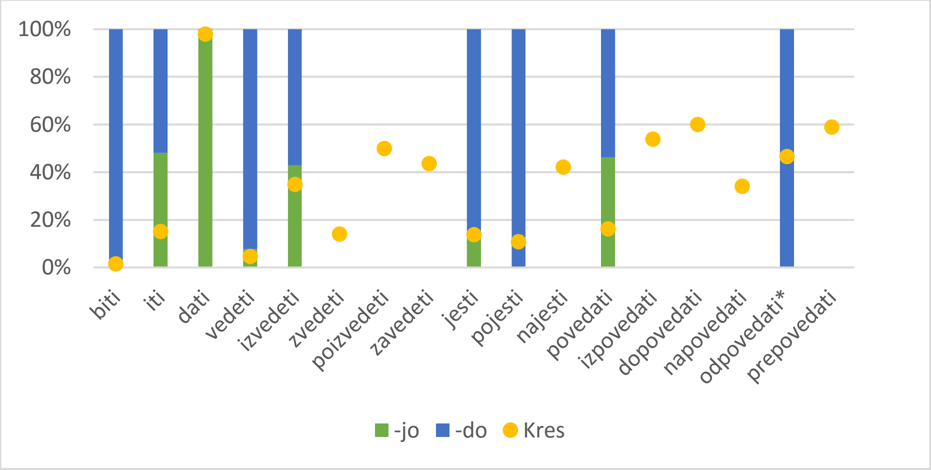


Figure 5: Percentage of *-jo*/*-do* in the Wiki talk subcorpus (green/blue) and the Kres corpus (yellow).

Geolocating German on Twitter
Hitches and Glitches of Building and Exploring a Twitter Corpus

Bettina Larl, Eva Zangerle
University of Innsbruck
E-mail: bettina.larl@uibk.ac.at, eva.zangerle@uibk.ac.at

Abstract

Languages, and thus Linguistics, have always been influenced by technological developments and new media forms and every development brought new methods and approaches of how language can or should be studied and explored. About 16% of the EU residents speak German as a native language and this makes it the widest spread language within the European Union. German is a pluricentric language with three standard varieties: German Standard German, Swiss Standard German and Austrian Standard German. The official borders between Germany, Austria and Switzerland also form the boundary between the three standards.

Because of easy access and informal communication methods, more and more oral markers find their way into written language. This is often showcased on social media platforms such as Twitter. Every tweet includes language output in the form of short messages that can contain different regional characteristics. Tweets can be geolocated, which means these language outputs can be assigned to the geographic location they were tweeted from.

To explore research questions like "Is there a connection between the language output and the geographic location tweets were sent from?" and "Could, for example, lexical varieties be allocated to a specific region by geolocation information provided in tweets?" We are building a Twitter Corpus. The Corpus contains tweets collected via the Twitter streaming API, using a binding box around the rough approximation of the Deutscher Sprachraum and re-filtering the results for Tweets sent within Germany, Austria, Switzerland and South Tyrol/Italy. This paper shows preliminary findings of hand sampling a random sample of 1,000,000 Tweets.

Keywords: Twitter, geolocation, German

The #Intermittent Corpus: Corpus Features, Ethics and Workflow for a CMC Corpus of Tweets in TEI

Julien Longhi
Cergy-Pontoise University, AGORA
E-mail: julien.longhi@u-cergy.fr

Abstract

This poster aims to describe issues encountered whilst structuring a corpus of tweets compiled from the key word intermittent (arts worker) in order to analyse a discursive topic related to the controversy surrounding the status of French arts workers. This corpus is part of the CoMeRe project (CoMeRe, 2014): it aims to build a kernel corpus of computer-mediated communication (CMC) genres with interactions in the French language. Three key words characterize the project: variety, standards and openness. A variety of interactions was sought: public or private interactions as well as interactions from informal, learning and professional situations. The CoMeRe project structured the corpora in a uniform way using the Text Encoding Initiative format (TEI, Burnard & Bauman, 2013) and described each corpus using Dublin Core and OLAC standards for metadata (DCMI, 2014; OLAC, 2008). The TEI model was extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse (Chanier et al., 2014). The term 'openness' also characterizes the project: The corpora have been released as open data on the French national platform of linguistic resources (ORTOLANG, 2013) in order to pave the way for scientific examination by partners not involved in the project as well as replicative and cumulative research.

This poster presentation aims to give an overview of the corpus building process using, as a case study, a corpus of tweets cmr-intermittent (Longhi et al., 2016). The following steps led to the choice of tweets:

1) In 2015, with the creation of a threshold of at least 10 tweets with the #intermittent (s), we identified 215 accounts, each of which had produced at least 10 tweets explicitly referenced as contributing to this theme (in order to have representative accounts).

2) By gathering all of the tweets sent by those 215 people, we collected 586, 239 tweets.

3) 10,876 of the 586, 239 tweets contained the #: #intermittent(s): the #intermittent corpus corresponds to these 10, 876 tweets.

The poster will focus, firstly, on how features that are specific to Twitter were included and structured in the interaction space TEI model. We will exemplify how certain features are accounted for in TEI. These include hashtags that label tweets in order that other users can see tweets on the same topic and at signs that allow users to mention or reply to other users. Secondly, the poster will evoke some of the ethical and rights issues that had to be considered before publishing this corpus of tweets. Finally, the workflow and multi-stage quality control procedure adopted during the corpus building process will be illustrated.

Keywords: tweets, corpus, TEI, CMC corpora

The Construction of a Teletandem Multimodal Data Bank

Queila Barbosa Lopes
São Paulo State University "Júlio de Mesquita Filho" - UNESP
E-mail: queilalopes@gmail.com

Abstract

The discussion presented here represents the initial reflection of my doctoral research. The main purpose of this research is to propose an organization of a multimodal data bank in semi-integrated and integrated Teletandem (Aranha & Cavalari, 2014) modalities. "Teletandem is a virtual, autonomous, and collaborative context that uses online teleconferencing tools (text, voice, and webcam images of VoIP technology, such as Skype) to promote intercontinental and intercultural interactions between students who are learning a foreign language" (Telles, 2015: 2). During these interactions, the interactants produce some genres to communicate. All production is saved in the computers and then saved in external hard disks. As result, we have a considerable amount of research data. The data bank organization will be based on the bazermanian conception of genres system according to which genres occur in an activity system (Bazerman, 1994; 2005). According to this conception, every social activity is done through genre sets, which are interrelated within a genre system, occurring in an activity system. The argument is grounded on the socio-rhetorical genre approach, which comprehend genres as a typified and socially situated action. Based on this assumption, I believe it will be possible to propose an organization of the data bank which will optimize researcher's time and it will make possible future diachronic studies. It will also help to understand how teletandem learning of a foreign language works. The question that guides my research is "Considering the genres characteristics of teletandem practice, how is it possible to organize a multimodal data bank in integrated and semi-integrated Teletandem? I will try to use the methodology proposed by Chanier and Wigham (2016) "to transform […] data from online learning situations". It will be also relevant to consider the concept of learning scenario (Foucher, 2010) as the space where there is the occurrence of one genre instead of another. Data were collected from 2012 to 2015, when around 655 hours of video interaction were recorded, 477 chats, 849 reflexives diaries, 180 questionnaires (initial and final) and 1444 texts were produced. Texts were also revised and rewritten by the participants. The objective of this presentation is to share a) the status of this work to the international community of researchers on Computer Mediated Communication, so that the work can be improved by the comments of its members, and b) especially the questions I have faced during this stage of the research.

Keywords: Teletandem, genre system, activity system, learning scenario

Graphic Euphemisms in Slovenian CMC

Mija Michelizza, Urška Vranjek Ošlak
Fran Ramovš Institute of the Slovenian Language ZRC SAZU
Novi trg 4, SI-1000 Ljubljana
E-mail: mmija@zrc-sazu.si, uvranjek@zrc-sazu.si

Abstract

Taboo words have been a part of human communication for as long as language has existed. Scientists are not in agreement on why taboo words emerged but it seems as if certain words have always been out-of-bounds for language users or have always caused certain negative feelings or reactions. Everyday communication (written and spoken) is filled with taboo words either expressed openly or disguised and concealed as more or less harmless. The evolution of the Internet and its many communication possibilities have led to a new (and somewhat less hidden) growth of taboo words, especially swear words and words designed to insult the recipient of the communication.

The poster presents the analysis of graphic euphemisation of chosen swear words in Slovenian CMC (on Twitter and in online news comments) and identifies different ways in which CMC users disguise taboo words, mostly in order to avoid automatic detection and deletion of their tweets or comments. The search was performed with search queries kur* and piz*. Most common graphic euphemism types are the substitution of a letter with a non-letter symbol and the insertion of a non-letter symbol (eg. kur**, piz.ijo). Repeated letters are also common (eg. pizzzzzda). Substitutions of letters with visually similar symbols (eg. kur@...), other letters or letter combinations with similar pronunciation (eg. kurz, kurchiti, pyzda, pisda) are less frequent. The analysis also shows that CMC users are very innovative; juxtaposition, puns and various word formation procedures (eg. pizdapaponedeljek, pizdarna (< pisarna), kurbenizon (< kombinezon)) are very common even though their primary role is language play rather than taboo word encryption.

Keywords: graphic euphemisms, CMC, taboo words, Slovenian language

Author Index

Author Index