HAL
open science

Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)

Egon W. Stemle, Ciara R. Wigham

To cite this version:

Egon W. Stemle, Ciara R. Wigham. Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17). 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17), Oct 2017, Bolzano, Italy. cmc-corpora conference series, 2017, cmc-corpora pre-conference proceedings, 10.5281/zenodo.1040714 . hal-01614310v1

HAL Id: hal-01614310

https://hal.science/hal-01614310v1

Submitted on 10 Oct 2017 (v1), last revised 7 Nov 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Conference Proceedings

Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)

3-4 October 2017, Eurac Research, Italy

Editors
Egon W. Stemle
Ciara R. Wigham

Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)

Editors: Egon W. Stemle, Ciara R. Wigham

Bolzano, 2017
First edition

Conference web site: https://cmc-corpora2017.eurac.edu
This publication is available at: https://cmc-corpora2017.eurac.edu/proceedings/

This publication was supported by

eurac research

cmc-corpora Conference Series

CLARIN
Common Language Resources and Technology Infrastructure

This publication was compiled with LuaLaTeX. The LaTeX template is based on KOMA script and can be found online: https://github.com/cmc-corpora/.

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

Preface

This volume presents the proceedings of the 5th edition of the annual conference series on CMC and Social Media Corpora for the Humanities (cmc-corpora2017). This conference series is dedicated to the collection, annotation, processing, and exploitation of corpora of computer-mediated communication (CMC) and social media for research in the humanities. The annual event brings together language-centered research on CMC and social media in linguistics, philologies, communication sciences, media and social sciences with research questions from the fields of corpus and computational linguistics, language technology, text technology, and machine learning.

The 5th Conference on CMC and Social Media Corpora for the Humanities was held at Eurac Research on October, 4th and 5th, in Bolzano, Italy. This volume contains extended abstracts of the invited talks, papers, and extended abstracts of posters presented at the event. The conference attracted 26 valid submissions. Each submission was reviewed by at least two members of the scientific committee. This committee decided to accept 16 papers and 8 posters of which 14 papers and 3 posters were presented at the conference. The programme also includes three invited talks: two keynote talks by *Aivars Glaznieks* (Eurac Research, Italy) and *A. Seza Doğruöz* (Independent researcher) and an invited talk on the Common Language Resources and Technology Infrastructure (CLARIN) given by *Darja Fišer*, the CLARIN ERIC Director of User Involvement.

We wish to thank all colleagues who have contributed to the conference and to this volume with their papers, posters, and invited talks. Thanks also to all members of the scientific committee and to the local coordinating committee without whom the conference would not have taken place. Whilst previous events in the conference cycle were held in Dortmund, Germany (2013 and 2014), Rennes, France (2015) and Ljubljana, Slovenia (2016), we hope that the Bolzano 2017 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.

We look forward to welcoming colleagues at the 2018 conference to be held in Antwerp, Belgium to continue the scientific exchange.

September 30, 2017
Bolzano/Bozen

Ciara R. Wigham, Université Clermont Auvergne (France)
Egon W. Stemle, Eurac Research (Italy)

Chair of the Scientific Committee and chair of the Organizing Committee.

Table of Contents

Preface iii
Committees vi

Invited Talks 1
Think Global, Write Local – Patterns of Writing Dialect on SNS 2
Aivars Glaznieks
Small vs. Big Data in Language Research: Challenges and Opportunities 3
A. Seza Doğruöz
CLARIN Survey of CMC Resources and Tools 4
Darja Fišer

Papers 5
The Impact of WhatsApp on Dutch Youths' School Writing 6
Lieke Verheijen and Wilbert Spooren
Modeling Non-Standard Language Use in Adolescents' CMC: The Impact and Interaction of Age, Gender and Education 11
Lisa Hilte, Reinhild Vandekerckhove and Walter Daelemans
Investigating Interaction Signs across Genres, Modes and Languages: The Example of OKAY 16
Laura Herzberg and Angelika Storrer
Anonymisation of the Dortmund Chat Corpus 2.1 21
Harald Lüngen, Michael Beißwenger, Laura Herzberg and Cathrin Pichler
Emoticons as multifunctional and pragmatic Resources: a corpus-based Study on Twitter . . 25
Stefania Spina
Corpus-Based Analysis of Demonyms in Slovene Twitter 30
Taja Kuzman and Darja Fišer
European Language Ecology and Bilingualism with English on Twitter 35
Steven Coats
Reliable Part-of-Speech Tagging of Low-frequency Phenomena in the Social Media Domain 39
Tobias Horsmann, Michael Beißwenger and Torsten Zesch
Developing a protocol for collecting data in Higher Education: assessing natural language metadata for a Databank of Oral Teletandem Interactions. 44
Paola Leone
The #Idéo2017 Platform 46
Julien Longhi, Claudia Marinica, Nader Hassine, Abdulhafiz Alkhouli and Boris Borzic
Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? 52
Michael Beißwenger, Ciara Wigham, Carole Etienne, Holger Grumt Suárez, Laura Herzberg, Darja Fišer, Erhard Hinrichs, Tobias Horsmann, Natali Karlova-Bourbonus, Lothar Lemnitzer, Julien Longhi, Harald Lüngen, Lydia-Mai Ho-Dac, Christophe Parisse, Céline Poudat, Thomas Schmidt, Egon Stemle, Angelika Storrer and Torsten Zesch
"You're trolling because..." – A Corpus-based Study of Perceived Trolling and Motive Attribution in the Comment Threads of Three British Political Blogs 56
Márton Petykó
Fear and Loathing on Twitter: Attitudes towards Language 61
Damjan Popič and Darja Fišer
A Comparative Study of Computer-mediated and Spoken Conversations from Pakistani and U.S. English using Multidimensional Analysis 65
Muhammad Shakir and Dagmar Deuber

Posters 70

MoCoDa 2: Creating a Database and Web Frontend for the Repeated Collection of Mobile Communication (WhatsApp, SMS & Co) **71**
Michael Beißwenger, Marcel Fladrich, Wolfgang Imo and Evelyn Ziegler

Public Service News on Facebook: Exploring Journalistic Usage Patterns and Reaction Data **72**
Daniel Pfurtscheller

The graphic realization of /l/-vocalization in Swiss German WhatsApp messages **73**
Simone Ueberwasser

Appendix 74

Author Index **75**

Keyword Index **76**

Committees

Scientific Committee

Chair	
Ciara Wigham	Université Clermont Auvergne
Co-Chairs	
Michael Beißwenger	Universität Duisburg-Essen
Darja Fišer	Faculty of Arts, University of Ljubljana
Members	
Andrea Abel	Eurac Research
Steven Coats	University of Oulu
Daria Dayter	University of Basel
Tomaž Erjavec	Dept. of Knowledge Technologies, Jožef Stefan Institute
Jennifer Frey	Eurac Research
Aivars Glaznieks	Eurac Research
Axel Herold	Berlin-Brandenburgische Akademie der Wissenschaften
Julien Longhi	Université de Cergy-Pontoise
Harald Lüngen	Institut für Deutsche Sprache
Maja Miličević	University of Belgrade
María-Teresa Ortego-Antón	Departamento de Lengua Española (Área de Traducción e Interpretación) - Universidad de Valladolid
Céline Poudat	BCL, Université Nice-Sophia Antipolis, CNRS
H. Müge Satar	Newcastle University
Stefania Spina	Università per Stranieri di Perugia
Egon W. Stemle	Eurac Research
Angelika Storrer	Universität Dortmund

Coordinating Committee

Michael Beißwenger	Universität Duisburg-Essen
Darja Fišer	Faculty of Arts, University of Ljubljana
Ciara Wigham	Université Clermont Auvergne

Organizing Committee

Chair	
Egon W. Stemle	Eurac Research
Members	
Irina Iokhno	Eurac Research
Daniela Gasser	Eurac Research

Invited Talks

Contents

Think Global, Write Local – Patterns of Writing Dialect on SNS **2**
Aivars Glaznieks

Small vs. Big Data in Language Research: Challenges and Opportunities **3**
A. Seza Doğruöz

CLARIN Survey of CMC Resources and Tools **4**
Darja Fišer

Think Global, Write Local – Patterns of Writing Dialect on SNS

Aivars Glaznieks
Institute for Applied Linguistics, Eurac Research
Viale Druso 1, 39100 Bolzano, Italy
aivars.glaznieks@eurac.edu

Abstract

Social Network Sites (SNS) claim that they are on a mission to connect the world. They facilitate communication among people wherever they are located. Consequently, many users of SNS communicate with a broad and heterogenic group of friends on different occasions and thereby express various aspects of their identities (such as gender, age, ethnic background etc.). One aspect may also be a local identity.

Users of SNS can show their local identity linguistically by using a regional variety. Sometimes, the use of single regionally marked words or sporadic regiolectal spellings are sufficient to identify the regional background of the writer (Androutsopoulos and Ziegler, 2003); in other cases entire text messages and conversations appear in dialectal spellings meaning that the dialect appears as the main variety of the conversation (Siebenhaar, 2008). The extent of dialect use in computer-mediated communication (CMC) may depend on various factors such as the individual dialect skills, the vividness and prestige of the respective dialect in the community, emotional involvement in the given topic, age, gender, the intended recipient, and other factors probably interacting with each other (Peersman et al., 2016).

The use of regional dialects in written CMC is one reason (amongst others) why language in CMC often differs from the respective standard languages. Since no orthographic rules are usually available for writing in dialect, it is up to the users to represent their dialect in a proper but readable and comprehensible way. Users have to construct their regiolectal language variety on the basis of the orthography of the respective standard language, which usually allows also for variation. One reason for this may be various adequate possibilities to represent a dialect word within a given writing system (e.g. German, cf. Dürscheid and Stark (2013)). Another reason may be the (sometimes very slight) phonetic differences between regionally close dialects that writers want (or do not want) to turn up in the dialect respelling (Sebba, 2007). Therefore, dialect respellings are not always coherent (neither with respect to a group of dialect speakers nor with respect to individual writers) but usually appear in various forms (Müller, 2011). However, unifications of respellings in CMC are described for pidgin languages (Heyd, 2016) and also occur in dialectal CMC (Tophinke, 2008).

Over the last decade, researchers started to compile corpora containing different genres of CMC. Such CMC corpora enable a systematic analysis of the way dialect features are reflected in written communication. In my talk, I will focus on patterns of the regional dialect(s) in the DiDi Corpus, a collection of Facebook messages from around 100 South Tyrolean writers (`http://www.eurac.edu/didi`). I will provide examples of regional features, analyse the distribution of such features, and discuss challenges of identifying local writings on SNS.

Keywords: orthography, dialect writing, facebook messages

References

Androutsopoulos, J. and Ziegler, E. (2003). Sprachvariation und Internet: Regionalismen in einer Chat-Gemeinschaft. In Jannis Androutsopoulos et al., editors, *Standardfragen: soziolinguistische Perspektiven auf Sprachgeschichte, Sprachkontakt und Sprachvariation*, page 251279. Peter Lang, Frankfurt a. M.

Dürscheid, C. and Stark, E. (2013). Anything goes? SMS, phonographisches Schreiben und Morphemkonstanz. In Martin Neef et al., editors, *Die Schnittstelle von Morphologie und geschriebener Sprache*, page 189209. de Gruyter, Berlin.

Heyd, T. (2016). Global varieties of English gone digital: Orthographic and semantic variation in digital Nigerian Pidgin. In Lauren Squires, editor, *English in Computer-Mediated Communication*, pages 101–122. de Gruyter, Berlin.

Müller, C. M. (2011). Dialektverschriftung im Spannungsfeld zwischen standardnah und lautnah. Eine korpuslinguistische Untersuchung der Rubrik 'Dein SMS' in der Aargauer Zeitung. In Helen Christen, et al., editors, *Struktur, Gebrauch und Wahrnehmung von Dialekt. Beiträge zum 3. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD), Zürich, 7.-9. September 2009*, pages 155–178, Wien. Praesens Verlag.

Peersman, C., Daelemans, W., Vandekerckhove, R., Vandekerckhove, B., and Van Vaerenbergh, L. (2016). The Effects of Age, Gender and Region on Non-standard Linguistic Variation in Online Social Networks. *eprint arXiv:1601.02431*, pages 1–24, 1.

Sebba, M. (2007). *Spelling in Society*. Cambridge University Press, Cambridge.

Siebenhaar, B. (2008). Quantitative Approaches to Linguistic Variation in IRC: Implications for Qualitative Research. *Language@Internet*, 5(4):1–14.

Tophinke, D. (2008). Regional schreiben: Weblogs zwischen Orthographie und Phonographie. In Helen Christen et al., editors, *Sprechen, Schreiben, Hören. Zur Produktion und Perzeption von Dialekt und Standardsprache zu Beginn des 21. Jahrhunderts. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20.-23. September 2006*, pages 153–179, Wien. Praesens Verlag.

Small vs. Big Data in Language Research: Challenges and Opportunities

A. Seza Doruz
Independent Researcher
a.s.dogruoz@gmail.com

Abstract

Mobile communication tools and platforms provide various opportunities for users to interact over social media. With the recent developments in computational research and machine learning, it has become possible to analyze large chunks of language related data automatically and fast. However, these tools are not readily available to handle data in all languages and there are also challenges handling social media data. Even when these issues are resolved, asking the right research question to the right set and amount of data becomes crucially important.

Both qualitative and quantitative methods have attracted respectable researchers in language related areas of research. When tackling similar research problems, there is need for both top-down and bottom-up data-based approaches to reach a solution. Sometimes, this solution is hidden under an in-depth analysis of a small data set and sometimes it is revealed only through analyzing and experimenting with large amounts of data. However, in most cases, there is need for linking the findings of small data sets to understand the bigger picture revealed through patterns in large sets.

Having worked with both small and large language related data in various forms, I will compare pros and cons of working with both types of data across media and contexts and share my own experiences with highlights and lowlights.

Keywords: social media data, machine learning, small vs. large data sets, multilingualism

References

Nguyen, D. and Doğruöz, A. (2014). Word level language identification in online multilingual communication. In *EMNLP*.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational linguistics*.

Papalexakis, E. and Doğruöz, A. S. (2015). Understanding multilingual social networks in online immigrant communities. In *Proceedings of the 24th International Conference on World Wide Web*, pages 865–870. ACM.

CLARIN Survey of CMC Resources and Tools

Darja Fišer
University of Ljubljana
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Abstract

With the growing volume and importance of computer-mediated communication, the need to understand its linguistic and social dimensions, along with CMC-robust language technologies is on the rise as well. This is reflected in the increasing number of conferences, projects and positions involving analysis of CMC in a wide range of disciplines in Digital Humanities, Social Sciences and Computer Science. As a result, a number of valuable CMC corpora, datasets and tools are being developed (Beißwenger et al., 2017) but unfortunately, due to non-negligible technical, legal and ethical obstacles, not many are being shared and reused.

Since it is the mission of CLARIN to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for researchers in Digital Humanities and Social Sciences (Krauwer and Hinrichs, 2014), it is our goal to have a good overview of the available resources and tools, to offer support to their developers to overcome the technical, legal and ethical obstacles and deposit them to the CLARIN infrastructure, as well as to the researchers with diverse backgrounds, such as linguistics, media studies, psychology etc., but also to interested parties from the educational, commercial, political, medical and legal sectors of the society who are interested in using them.

The first step in this direction was an interdisciplinary workshop[1] on the creation and use of social media which was organized within the Horizon 2020 CLARIN-PLUS project on 18 and 19 May 2017 in Kaunas, Lithuania. The aims of the workshop were to demonstrate the possibilities of social media resources and natural language processing tools for researchers with a diverse research background and an interest in empirical research of language and social practices in computer-mediated communication, to promote interdisciplinary cooperation possibilities, and to initiate a discussion on the various approaches to social media data collection and processing.

The workshop also served as a platform to conduct a survey[2] of corpora, datasets and tools of computer-mediated communication in the languages spoken in countries that are members and observers of CLARIN ERIC. Apart from identifying the existing resources and tools, our motivation was to establish to which extent they are accessible through the CLARIN infrastructure and how the information and accessibility of them could be further optimized from a user perspective.

In this talk, I will give an overview of the identified corpora, the smaller, more focused datasets and tools that are tailored to processing computer-mediated communication. The focus of the talk will be on the comprehensiveness of the provided metadata, level of availability and accessibility of the identified resources and tools and the degree of their actual or potential inclusion in the CLARIN infrastructure. I will also discuss the simple and long-term possibilities of enriching the current state of the infrastructure and provide guidelines for creating and depositing CMC resources with a CLARIN center.

Keywords: CLARIN ERIC, research infrastructure, language resources, NLP tools, computer-mediated communication

References

Beißwenger, M., Chanier, T., Erjavec, T., Fišer, D., Herold, A., Lubešić, N., Lüngen, H., Poudat, C., Stemle, E., Storrer, A., and Wigham, C. (2017). Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In *Selected Papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 2628 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, pages 1–18. Linköping University Electronic Press, Linköpings universitet.

Krauwer, S. and Hinrichs, E. (2014). The clarin research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531. European Language Resources Association (ELRA).

[1] https://www.clarin.eu/event/2017/clarin-plus-workshop-creation-and-use-social-media-resources

[2] https://office.clarin.eu/v/CE-2017-1064-Resources-for-computer-mediated-communication.docx

Papers

Contents

The Impact of WhatsApp on Dutch Youths' School Writing 6
Lieke Verheijen and Wilbert Spooren

Modeling Non-Standard Language Use in Adolescents' CMC: The Impact and Interaction of Age, Gender and Education 11
Lisa Hilte, Reinhild Vandekerckhove and Walter Daelemans

Investigating Interaction Signs across Genres, Modes and Languages: The Example of OKAY 16
Laura Herzberg and Angelika Storrer

Anonymisation of the Dortmund Chat Corpus 2.1 21
Harald Lüngen, Michael Beißwenger, Laura Herzberg and Cathrin Pichler

Emoticons as multifunctional and pragmatic Resources: a corpus-based Study on Twitter .. 25
Stefania Spina

Corpus-Based Analysis of Demonyms in Slovene Twitter 30
Taja Kuzman and Darja Fišer

European Language Ecology and Bilingualism with English on Twitter 35
Steven Coats

Reliable Part-of-Speech Tagging of Low-frequency Phenomena in the Social Media Domain 39
Tobias Horsmann, Michael Beißwenger and Torsten Zesch

Developing a protocol for collecting data in Higher Education: assessing natural language metadata for a Databank of Oral Teletandem Interactions. 44
Paola Leone

The #Idéo2017 Platform 46
Julien Longhi, Claudia Marinica, Nader Hassine, Abdulhafiz Alkhouli and Boris Borzic

Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? 52
Michael Beißwenger, Ciara Wigham, Carole Etienne, Holger Grumt Suárez, Laura Herzberg, Darja Fišer, Erhard Hinrichs, Tobias Horsmann, Natali Karlova-Bourbonus, Lothar Lemnitzer, Julien Longhi, Harald Lüngen, Lydia-Mai Ho-Dac, Christophe Parisse, Céline Poudat, Thomas Schmidt, Egon Stemle, Angelika Storrer and Torsten Zesch

"You're trolling because..." – A Corpus-based Study of Perceived Trolling and Motive Attribution in the Comment Threads of Three British Political Blogs 56
Márton Petykó

Fear and Loathing on Twitter: Attitudes towards Language 61
Damjan Popič and Darja Fišer

A Comparative Study of Computer-mediated and Spoken Conversations from Pakistani and U.S. English using Multidimensional Analysis 65
Muhammad Shakir and Dagmar Deuber

The Impact of WhatsApp on Dutch Youths' School Writing

Lieke Verheijen, Wilbert Spooren
Radboud University (Nijmegen, the Netherlands)
Email: lieke.verheijen@let.ru.nl; w.spooren@let.ru.nl

Abstract

Today's youths are continuously engaged with social media. The informal language they use in computer-mediated communication (CMC) often deviates from spelling and grammar rules of the standard language. Therefore, parents and teachers fear that social media have a negative impact on youths' literacy skills. This paper examines whether such worries are justifiable. An experimental study was conducted with 500 Dutch youths of different educational levels and age groups, to find out if social media affect their productive or perceptive writing skills. We measured whether chatting via WhatsApp directly impacts the writing quality of Dutch youths' narratives or their ability to detect 'spelling errors' (deviations from Standard Dutch) in grammaticality judgement tasks. The use of WhatsApp turned out to have no short-term effects on participants' performances on either of the writing tasks. Thus, the present study gives no cause for great concern about any impact of WhatsApp on youths' school writing.

Keywords: computer-mediated communication, social media, WhatsApp, writing, literacy

1. Introduction

Youths are nowadays constantly using computer-mediated communication such as WhatsApp, Facebook chat, Snapchat, and Twitter. Examples (1)–(3) present chat messages by Dutch youths:

(1) **OMG!** Had je mijn mijn verhaal gezien
Hahahahhaahhaaha kwam ik pas vanochtend achter
k kan me **nie** eens herinneren **da** ik die gemaakt heb
Miss in **mn** slaap **ofzo hagahagagaa**
('OMG! Did you see my my story
Hahahahhaahhaaha only found out this morning
i cant even remember making dat
Mayb in me sleep or somethin hagahagagaa')

(2) Beetje te vroeg ik val echt in slaap maar alvast **happy birthdayyyyyy toooooooo youuuuuuuuuuu**! 🖤🖤🐱🐱🏃🏃💙💙🎉🎉 **loveyouuuuuu xxxxxxxxxx**
('Bit too early I'm really falling asleep but anyway…')

(3) **Liefie**💙 gaat ie weer met jou? Wat het je😩😟 bel me **weneer** je online bent😎✓ **ly**💙💙💙💙 zie je morgen☺ **BEL ME** ☺ **chatt**🤝😮
('Luv💙 you doin okay again? What hare you😩😟 call me whn you are online😎✓ ly💙💙💙💙 see you tomorrow☺ CALL ME ☺ honeyy🤝😮')

All the words in bold deviate from the Dutch standard language norms. They contain non-standard abbreviations, letter repetitions, phonetic respellings, overuse of capitalisation, and emoji. Such deviations from Standard Dutch lead to fears that informal written CMC may negatively interfere with writing in more formal settings (Spooren, 2009). These fears have existed for decades now;[1] it is just the medium under critique that changes every few years. Yet there are also scholars who point out that youths' literacy skills may benefit from social media use, via creativity with language, greater exposure to written texts, more engagement in writing, and greater metalinguistic awareness (Wood, Kemp, & Plester, 2013).

[1] Or, for that matter, for centuries. See Deutscher (2005) for a historical note on the concerns about language deterioration.

2. Research Goals and Hypotheses

Since prior research does not provide a conclusive answer about the impact of CMC on literacy (Verheijen, 2013), our research aims to contribute to this debate. The goal of this study is to determine whether Dutch youngsters' constant use of social media affects the way they write at school. We conducted an experiment to find out whether engaging in synchronous written CMC directly impacts their productive or perceptive school writing skills. Our focus was on the chat program WhatsApp, because this is currently a very popular medium among Dutch youths.
Furthermore, we aim to establish whether the demographic variables of age, education, and gender have an impact on youths' writing skills – specifically, if they have a mediating effect on the possible impact of WhatsApp on those writing skills. Therefore, the analysis will include four independent variables: not only condition (exposure vs. non-exposure to WhatsApp), but also educational level (lower, intermediate, higher), age group (adolescents vs. young adults), and gender (boys vs. girls). We hypothesize that a greater impact of CMC use on school writing skills may be displayed by youths of a younger age group or lower educational level, as it may be more difficult for them to keep these registers separate.

3. Methodology

3.1 Participants

Participants were 500 youths from secondary and tertiary educational institutions, from different educational levels and age groups, in Nijmegen and surroundings. The data collection period lasted from October to December 2016. Testing took place in an educational setting. Most participants were tested in class; only a small number in higher tertiary education voluntarily took part outside of class. The latter were reimbursed for their participation with gift certificates of € 5. Students from higher tertiary education belonged to different faculties and studies, including communication and information sciences,

biology, and literary and cultural studies. The adolescents (N = 300) were around 14 years old ($\bar{x}$ age = 14.2 yrs, range 13-16; 151 male, 149 female), all in the third grade. The young adults (N = 200) were around 20 years old ($\bar{x}$ age = 20.4 yrs, range 18-27; 72 male, 128 female). Table 1 shows an overview of the participants. Afterwards, underage participants were given a document with more information about the study and the researchers' contact details, to take home to their parents or caretakers.

		Educational level		
		lower	intermediate	higher
Age group	adolescents: secondary education	101	92	107
	young adults: tertiary education	102	-	98

Table 1: Overview of participants.

3.2 Data Collection

3.2.1. Priming: WhatsApp vs. Colouring

All classes that were tested were divided into two groups. The experimental groups were primed with CMC via social media: they were instructed to chat via WhatsApp on their own smartphones for fifteen minutes, in small groups of three or four students. They could chat about whatever they preferred; no specific conversation topics were provided, in order to generate as natural chat conversations as possible. During that time, the control groups performed a non-CMC-related control task, namely colouring mandalas. These tasks were chosen because, in a pilot study, they proved to be effective in revealing differences with respect to orthography and language correctness (Riemens, 2016).

3.2.2. Measuring Productive Writing Skills: Stories

To test their productive writing skills, all participants wrote a story in class, starting with the following sentence: "I was alone in a dark room. My hand groped for the light switch, but suddenly…" [translated from Dutch]. The formal writing genre that was tested was that of narrative storytelling. Since not all classes had easy access to computers and laptops, all stories were hand-written for consistency's sake.

3.2.3. Measuring Perceptive Writing Skills: GJTs

Participants also completed grammaticality judgement tasks (GJTs), to test their receptive grammar and spelling skills. These consisted of twenty sentences in which they had to spot and correct 'language errors'. These were orthographic deviations typical of CMC: various types of textisms (phonetic respelling, reduplication of letter, shortening, single letter homophone, initialism); missing capitalisation, diacritics, and punctuation; spelling 'errors' that are heavily frowned upon by Dutch language prescriptivists (*is/eens*, *d/t*, *jou/jouw*); emoticons; omissions; English borrowings; and extra spacing. Five sentences contained no orthographic deviations, so participants could spot and correct fifteen 'errors'.

3.3 Data Analysis

3.3.1. GJT Scores

For the grammaticality judgement tasks, two scores were computed for each participant. First, the choice score: whether they correctly identified the sentence as containing an 'error' or not (max. 20 points). Second, the correction score: whether they correctly managed to correct that 'error' (max. 15).

3.3.2. T-Scan Analysis

The stories were automatically analysed with T-Scan, software for conducting complexity analyses of Dutch texts (Pander Maat et al., 2014). T-Scan provided us with a staggering 411 variables for each text, out of which a theory-based selection of 27 relevant variables was made:

1) Zin_per_doc: number of sentences per essay
2) Word_per_doc: number of words per essay
3) Let_per_wrd: number of letters per word
4) Wrd_per_zin: number of words per sentence
5) Bijzin_per_zin: number of subordinate clauses per sentence
6) Pv_Frog_d: density of finite verbs
7) D_level: D-level
8) Nom_d: density of nominalisations
9) Lijdv_d: density of passive forms
10) AL_gem: average of all dependency lengths per sentence
11) AL_max: maximal dependency length per sentence
12) Bijw_bep_d: density of adverbials
13) TTR_wrd: type-token ratio (for words)
14) MTLD_wrd: measure of textual lexical diversity (for words)
15) Inhwrd_d: density of content words
16) Pers_vnw_d: density of personal and possessive pronouns
17) Ww_mod_d: density of modal verbs
18) Huww_tijd_d: density of auxiliary verbs of time
19) Koppelww_d: density of copula verbs
20) Imp_ellips_d: density of imperatives and elliptical constructions
21) Vg_d: density of conjunctions
22) Lidw_d: density of articles
23) Nw_d: density of nouns
24) Tuss_d: density of interjections
25) Spec_d: density of names and special words
26) Interp_d: density of punctuation
27) Afk_d: density of abbreviations

3.3.3. Exploratory Factor Analysis

Because the twenty-seven variables selected from T-Scan were still too many to put into a regression analysis, we used an exploratory factor analysis (with the extraction method of principal component analysis, PCA), to further reduce these to a set of writing components indicative of the writing quality of stories.

An orthogonal rotation method was chosen, namely varimax with Kaiser normalization: this method, which does not allow correlations between factors, facilitated the interpretation of results, since it maximizes the spread of loadings for a variable across all factors. There was no multicollinearity, because none of the correlation coefficients were $r \geq .84$. Missing values were replaced with the mean, because listwise deletion would result in a

loss of participants in the analysis, and pairwise deletion would lead to a non-positive definite matrix. The Keyser-Meyer-Olkin measure was well above .5 (*KMO* = .644), which verified the sampling adequacy for the analysis. Bartlett's test of sphericity showed that correlations between items were sufficiently large for PCA: χ^2 (351) = 6267.569, p < .001. The proportion of residuals with an absolute value greater than 0.05 was 50%. An initial analysis yielded eigenvalues for each component in the data. The large sample size of this study (500 participants) allowed us to use a scree plot with eigenvalues over 1 for deciding how many components to extract. The inflexion of the scree plot justified retaining three components.

Table 2 shows the results of the PCA after rotation. The items that cluster on the same components suggest that component 1 represents syntactic complexity, 2 lexical richness, and 3 writing productivity. The total variance explained by the three factors is 38.08%. The resulting factor scores were saved as Anderson-Rubin variables, so they did not correlate.

Rotated Component Matrix			
Writing variable	Rotated factor loadings		
	1	2	3
AL_max	**.868**	.141	.106
D_level	**.818**	-.193	.016
Bijzin_per_zin	**.792**	-.089	.077
AL_gem	**.764**	.221	.232
Wrd_per_zin	**.720**	.004	-.087
Interp_d	**-.718**	-.077	.065
Vg_d	**.556**	-.287	.072
Tuss_d	-.240	-.117	.186
Bijw_bep_d	.221	-.004	.107
Spec_d	-.147	.016	-.057
Pv_Frog_d	-.167	**-.762**	-.037
Nw_d	-.059	**.698**	-.166
Pers_vnw_d	-.102	**-.680**	.066
Let_per_wrd	-.002	**.624**	-.153
Inhwrd_d	-.045	**.554**	.054
Lidw_d	-.047	**.520**	-.232
MTLD_wrd	-.060	**.450**	.004
Nom_d	-.034	**.423**	.028
Koppelww_d	-.127	-.189	.020
Ww_mod_d	.057	-.145	.136
Imp_ellips_d	.039	-.099	.066
Word_per_doc	.023	.004	**.917**
TTR_wrd	-.141	.281	**-.800**
Zin_per_doc	**-.529**	-.038	**.782**
Huww_tijd_d	-.078	-.043	-.378
Lijdv_d	-.038	.028	-.139
Afk_d	-.095	.053	-.095
Eigenvalues	4.496	3.246	2.539
% of variance	16.652	12.021	9.405

Note: loadings > .40 appear in bold and colour.

Table 2: PCA rotated factor loadings for the story analysis.

3.3.4. Linear Multiple Regression

The next step of the statistical analysis was linear multiple regression. The outcome variables were the three A-R factor scores resulting from the exploratory factor analysis of the stories and the two GJT scores. The predictor variables were condition (colouring versus WhatsApp), the three demographic variables educational level, age group, and gender, plus all interactions between condition and the demographic variables. As we had no preconceived ideas about which variables would be significant predictors, they were all entered with the forced entry method. The first block of the regression only contained the main effects. The interactions were entered in subsequent blocks.[2]

4. Results and Discussion

Table 3 shows the means and standard deviations of participants' performances on the writing tasks:

	Dependent variables				
	Stories			GJTs	
Independent variables	syntactic complexity: $\bar{x}$ (*SD*)	lexical richness: $\bar{x}$ (*SD*)	writing productivity: $\bar{x}$ (*SD*)	choice score: $\bar{x}$ (*SD*)	correction score: $\bar{x}$ (*SD*)
Condition:					
Colouring, *N* = 207	-0.02 (1.00)	0.06 (1.04)	0.00 (1.09)	14.44 (3.16)	13.82 (1.03)
WhatsApp, *N* = 201	0.07 (1.05)	0.09 (0.96)	0.03 (0.99)	14.71 (3.04)	13.74 (0.99)
Educational level:					
Lower, *N* = 203	0.19 (1.06)	-0.03 (0.93)	-0.17 (1.04)	12.68 (2.87)	13.53 (0.99)
Higher, *N* = 205	-0.13 (0.96)	0.18 (1.06)	0.20 (1.00)	16.44 (2.00)	14.03 (0.96)
Age group:					
Adolescents, *N* = 208	0.01 (1.14)	-0.24 (0.92)	0.00 (1.08)	14.10 (3.09)	13.75 (1.01)
Young adults, *N* = 200	0.05 (0.88)	0.40 (0.98)	0.03 (1.00)	15.06 (3.04)	13.81 (1.01)
Gender:					
Male, *N* = 179	0.16 (1.16)	0.09 (1.00)	-0.14 (1.11)	14.25 (3.14)	13.60 (1.02)
Female, *N* = 229	-0.08 (0.89)	0.06 (1.0)	0.14 (0.97)	14.82 (3.06)	13.92 (0.98)
TOTAL, *N* = 408	0.03 (1.02)	0.07 (1.00)	0.01 (1.04)	14.57 (3.10)	13.78 (1.01)

Table 3: Descriptive statistics.

4.1 Syntactic Complexity

One writing component was syntactic complexity, presented in Table 2 in column '1'. Educational level was a significant negative predictor: higher educated youths wrote syntactically less complex stories. At a first glance, this may seem surprising. However, this rather fits the genre of narrative storytelling, which does not require complex, long sentences – as opposed to, for example, expository discussion as in essays. So the higher educated youths showed more mastery of the genre of stories. Gender was a significant negative predictor: male participants wrote syntactically more complex stories.

[2] Participants of the intermediate secondary educational level were eventually omitted, as they were not part of the original research plan and would decrease the reliability of the analyses because of an empty cell in the research design: no youths of intermediate tertiary education were tested (see Table 1).

	Dependent variable: syntactic complexity		
Independent variables	*B*	*SE B*	*β*
Condition	0.09	0.10	0.04
Educational level	**-0.32**	**0.10**	**-0.16****
Age group	0.06	0.10	0.03
Gender	**-0.24**	**0.10**	**-0.12***
R^2 / Adjusted R^2	.04 / .03		
ANOVA	F (4, 403) = 4.24 (p < .01)		

Table 4: Regression results for syntactic complexity.[3]

4.2 Lexical Richness

Another writing component was lexical richness. Table 2 shows the variables that loaded onto this component, in the column labelled '2'. Lexical richness was positively predicted by educational level and age group: the stories of higher educated and of older participants were lexically richer. In addition, there was a significant interaction between gender and condition. For boys, WhatsApp had a small significant positive effect on their stories' lexis; for girls, the effect was negative but non-significant.

	Dependent variable: lexical richness		
Independent variables	*B*	*SE B*	*β*
Condition	0.32	0.19	0.16
Educational level	**0.31**	**0.13**	**0.16***
Age group	**0.64**	**0.13**	**0.32*****
Gender	0.11	0.13	0.06
Educational level × condition	-0.14	0.19	-0.06
Age group × condition	0.08	0.19	0.03
Gender × condition	**-0.50**	**0.19**	**-0.22****
R^2 / Adjusted R^2	.13 / .12		
ANOVA	F (7, 400) = 8.90 (p < .001)		

Table 5: Regression results for lexical richness.

4.3 Writing Productivity

The third component of the stories, writing productivity, is presented in column '3' in Table 2. It was positively predicted by educational level: youths with a higher educational level produced significantly longer stories. Gender was a significant positive predictor of writing productivity too: female participants wrote longer stories.

	Dependent variable: writing productivity		
Independent variables	*B*	*SE B*	*β*
Condition	0.03	0.10	0.01
Educational level	**0.37**	**0.10**	**0.18*****
Age group	-0.01	0.10	0.00
Gender	**0.27**	**0.10**	**0.13****
R^2 / Adjusted R^2	.05 / .04		
ANOVA	F (4, 403) = 5.19 (p < .001)		

Table 6: Regression results for writing productivity.

4.4 GJT Choice Score

For the grammaticality judgement tasks, educational level and age group were significant positive predictors of the choice score, so higher educated youths and older youths were more successful in spotting 'language errors'.

	Dependent variable: GJT choice score		
Independent variables	*B*	*SE B*	*β*
Condition	0.21	0.24	0.03
Educational level	**3.77**	**0.24**	**0.61*****
Age group	**0.99**	**0.24**	**0.16*****
Gender	0.35	0.24	0.06
R^2 / Adjusted R^2	.40 / .39		
ANOVA	F (4, 403) = 67.21 (p < .001)		

Table 7: Regression results GJT choice score.

4.5 GJT Correction Score

The correction score was significantly positively predicted by educational level and gender: both higher educated and female participants were more successful in correcting 'language errors'. The interaction between gender and condition was also significant. For girls, WhatsApp had a small significant negative effect on their correction score; for boys, the effect was positive but non-significant.

	Dependent variable: GJT correction score		
Independent variables	*B*	*SE B*	*β*
Condition	-0.09	0.19	-0.05
Educational level	**0.39**	**0.13**	**0.20****
Age group	-0.11	0.14	-0.06
Gender	**0.51**	**0.14**	**0.25*****
Educational level × condition	0.22	0.19	0.10
Age group × condition	0.31	0.19	0.13
Gender × condition	**-0.44**	**0.20**	**-0.20***
R^2 / Adjusted R^2	.11 / .09		
ANOVA	F (7, 400) = 6.71 (p < .001)		

Table 8: Regression results for GJT correction score.

An overview of the results of all the linear multiple regressions is presented in Table 9 below:

	Dependent variables				
	Stories			GJTs	
Independent variables	syntactic complexity	lexical richness	writing productivity	choice score	correction score
Main variables:					
Condition					
Educational level	−	+	+	+	+
Age group		+		+	
Gender	−		+		+
Interactions:					
EL × C					
AG × C					
G × C		−			−
EL × AG × C					
EL × G × C					
AG × G × C					
EL × AG × G × C					

Note: + = positive predictor, − = negative predictor, C = condition, EL = educational level, AG = age group, G = gender.

Table 9: Overview of regression results.

[3] For all tables with results, *p < .05, **p < .01, ***p < .001.

5. Conclusion

This paper reports on an experimental study measuring whether the use of WhatsApp has a direct impact on the writing quality of Dutch youths' stories or on their ability to detect 'spelling errors' in grammaticality judgement tasks. Educational level was a significant positive predictor for four writing variables, and age group for two. Gender predicted three writing variables. Condition did not affect the writing variables. We can thus conclude that WhatsApp does not appear to impact Dutch youths' productive or perceptive writing skills. Only two minor interactions between condition and gender were found, which suggests that perhaps there might be a slight impact of WhatsApp, moderated by gender, in which boys' lexical richness might benefit from CMC and girls' ability to correct language errors might be affected by it.

Two objections to this conclusion may be raised. One might doubt whether our measuring instrument was sensitive enough to detect differences in writing quality. However, the effectiveness of our testing method is confirmed by finding main effects for three demographic variables: these show that analysing the stories with T-Scan, as well as the GJT scores, are successful ways to detect differences in youths' writing skills. One might also argue that our experimental manipulation, the use of WhatsApp for fifteen minutes, was not strong enough to generate any effects. That cannot be the case either, because we found some interactions with gender; moreover, the prime already proved to yield significant results in a pilot study conducted in advance. All in all, the present study gives no cause for concern about the impact of WhatsApp on school writing.

6. Future Work

We hypothesized that particularly writers of a younger age group and lower educational level could experience possible interference of social media on their school writings. Prior research also suggests that youths of a lower educational track have more trouble distinguishing informal online writing (CMC) from more formal offline writing repertoires (Vandekerckhove & Sandra, 2016). Further research could explore other ways to test for such interference. Perhaps effects of social media crop up in minor orthographic details of their school writings, such as non-standard punctuation, capitalisation, spacing, or diacritics, because in the pilot study, these were the items on which WhatsApp use had the greatest impact. The frequent omission of punctuation and capitalization (sentence-initial or with proper names) in school writings was also noted by Vandekerckhove and Sandra (2016). The stories written for the present experiment could thus be analysed for the occurrence of such non-standard orthographic details.

In addition, the WhatsApp chats produced by roughly half of the participants during the priming phase were nearly all collected afterwards, of course with their consent (sent to the first author via email), but were not analysed. If properly formatted and annotated, the CMC data thus compiled could be a valuable corpus for further analysis. We could study the nature of these WhatsApp interactions, e.g. for the use of textisms, to find out to which extent these chats actually differ from Standard Dutch in terms of orthography and grammar and whether the amount of deviations affected the direct impact of CMC use on the writing tasks.

7. Acknowledgements

This study is part of a research project funded by the Dutch Organisation for Scientific Research (NWO), project number 322-70-006.

8. References

Deutscher, G. (2005). *The Unfolding of Language*. London: Random House.

Pander Maat, H., et al. (2014). T-Scan: A new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal*, 4, pp. 53--74.

Riemens, N. (2016). De directe invloed van WhatsApp op schrijfvaardigheid. BA thesis, Radboud University.

Spooren, W. (2009). Bezorgde ouders? De relatie tussen chat en schrijfkwaliteit. In W. Spooren, M. Onrust, & J. Sanders (Eds.), *Studies in Taalbeheersing* 3. Assen: Van Gorcum, pp. 331--342.

Vandekerckhove, R., and Sandra, D. (2016). De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext. *Tijdschrift voor Taalbeheersing*, 38(3), pp. 201--234.

Verheijen, L. (2013). The effects of text messaging and instant messaging on literacy. *English Studies*, 94(5), pp. 582--602.

Wood, C., Kemp, N. and Plester, B. (2013). *Text Messaging and Literacy: The Evidence*. New York, NY: Routledge.

Modeling Non-Standard Language Use in Adolescents' CMC: The Impact and Interaction of Age, Gender and Education

Lisa Hilte, Reinhild Vandekerckhove, Walter Daelemans
CLiPS Research Center, University of Antwerp
Postal address: Prinsstraat 13, B-2000 Antwerp, Belgium
E-mail: {lisa.hilte, reinhild.vandekerckhove, walter.daelemans}@uantwerpen.be

Abstract

The present paper deals with Flemish adolescents' informal computer-mediated communication (CMC) in a large corpus (2.9 million tokens) of chat conversations. We analyze deviations from written standard Dutch and possible correlations with the teenagers' gender, age and educational track. The concept of non-standardness is operationalized by means of a wide range of features that serve different purposes, related to the chatspeak maxims of orality, brevity and expressiveness. It will be demonstrated how the different social variables impact on non-standard writing, and, more importantly, how they interact with each other. While the findings for age and education correspond to our expectations (more non-standard markers are used by younger adolescents and students in practice-oriented educational tracks), the results for gender (no significant difference between girls and boys) do not: they call for a more fine-grained analysis of non-standard writing, in which features relating to different chat principles are examined separately.

Keywords: computer-mediated communication, non-standardness, teenage talk, language modeling

1. Introduction

Adolescents' informal CMC tends to deviate from formal standard writing in many ways: alternative spelling, non-standard capitalization, emoticons, … These deviations can be related to the three main principles behind chatspeak, i.e. the principles of orality, economy and expressive compensation (Androutsopoulos, 2011, 149; see section 3.2 for definitions and examples). While many CMC-studies report on just one type of features or present a small selection, the present study examines a wide array of 11 non-standard features and relates their frequency to three independent variables.

In the following sections, we will describe the goal of this study (section 2), as well as the dependent and independent variables (section 3). Next, we present the corpus and methodology (section 4), and finally, we will discuss and evaluate the results (section 5).

2. Goal of the Paper

We try to capture the impact of three aspects of the adolescent authors' profile on their CMC writing practices: their gender, age, and educational track. The latter variable has been largely neglected in CMC research. The same accounts for potential interactions between these variables: as boys and girls age, do their online writing practices evolve in a similar way? And do the same age and gender patterns emerge in different education types? In the end we want to demonstrate that the inclusion of a wide range of both independent and dependent variables is a prerequisite for a correct assessment of variation patterns in adolescents' CMC.

3. Dependent and Independent Variables

3.1. Independent Variables

All participants are high school students living in Flanders, the Dutch-speaking part of Belgium. We examine three social variables: the adolescents' gender, their age and their type of education (i.e. educational track).

Both gender and age are treated as binary variables: boys are compared to girls, and younger teenagers (13-16) to older ones (17-20). For educational track, we distinguish the three main types of secondary education in Belgium: ASO, TSO and BSO. ASO or General Secondary Education is theory-oriented and prepares students for higher education, whereas BSO or Vocational Secondary Education is practice-oriented, preparing students for a manual profession. TSO or Technical Secondary Education constitutes a more hybrid in-between level.

3.2. Dependent Variables

We selected 11 different linguistic features which are all deviations from the formal writing standard.

The largest set of features consists of 7 expressive markers which convey emotional or social involvement (see Hilte, Vandekerckhove & Daelemans, 2016 for a detailed analysis of these expressive markers):

1. non-standard capitalization
   e.g. *IK ZWEER HET* 'I swear'
2. emoticons and emoji
   e.g. *dammn we look so hot* 😍🔥💋❤
3. combinations question and exclamation marks
   e.g. *Echt**?!*** 'Really?!'
4. deliberate repetition ('flooding') of letters
   e.g. ***yeeeesss*** 'yes'
5. deliberate repetition ('flooding') of punctuation marks
   e.g. *Wat**???*** 'What???'

6. onomatopoeic rendering of laughter
   e.g. *hahaha*
7. rendering of kisses and hugs
   e.g. *Dankje* ***xxx*** 'Thank you xxx'

For orality (i.e. the underlying chatspeak principle to write 'as you speak'), we take one feature into account:

8. non-standard Dutch lexemes (informal Dutch, ranging from colloquial speech, regiolect/dialect words and slang to youth language ...)
   e.g. *vertel het* ***sebiet*** (std. Dutch: *vertel het* ***straks***, 'tell it later')

The economy principle ('make your message as concise as possible') was operationalized with:

9. chatspeak abbreviations and acronyms (i.e. non-standard shortened words or phrases)
   e.g. ***Omg*** *yes* (full version: '**Oh my god** yes')

The final category contains two features that do not really fit into one of the three chat principle categories but are characteristic of (Dutch) CMC and atypical of formal standard writing:

10. English words[1] (used in a Dutch conversation)
    e.g. *echt* ***nice*** 'really nice'
11. Discourse markers # (*hashtag*, to indicate a topic or express a feeling about it) and @ (*at*, to directly address one person in a group conversation)
    e.g. *#bestfriends*
    e.g. *@sarah*

4. Corpus and Methodology

4.1 Corpus

The corpus consists of Flemish teenagers' informal chat conversations and contains 2 885 084 tokens or 488 014 posts. The number of chatters in the corpus is 1384. The distributions for the social variables age, gender and education can be found in Table 1. We note that (dialect) region is a quasi-constant: almost all tokens (over 96%) are collected from participants living in the central Antwerp-Brabant region. The same holds for medium and year: almost all tokens (over 99%) are extracted from instant (i.e. synchronous) messages on Facebook/Messenger, WhatsApp or iMessage, and the vast majority of the tokens (87%) were produced in 2015-2016. Students consented to donate their conversations, and for minors, parents' consent was asked too. All chat material was anonymized before analysis – the participants' names were replaced by serial numbers, which are linked to the features of their social profiles (e.g. gender).

Variable	Subgroups	Tokens
Gender	Boys	985 928 (34%)
	Girls	1 899 156 (66%)
Age	Younger (13-16)	1 584 373 (55%)
	Older (17-20)	1 300 711 (45%)
Education	General (ASO)	920 114 (34%)
	Technical (TSO)	1 213 483 (42%)
	Vocational (BSO)	751 487 (26%)
Total		2 885 084

Table 1: Distributions for gender, age and education.

4.2. Methodology

4.2.1. Feature Extraction

All feature occurrences were extracted automatically using Python scripts. For a random test set, the software's output was compared to human annotation, which rendered a satisfying f-score of 0.90 (average for all 11 features).

4.2.2. Statistical Language Modeling

We statistically analyze the use of non-standard features by constructing a generalized linear mixed model (GLMM) on a token-level, using the R package 'lme4' (Bates et al., 2017). The GLMM tries to model and predict the response variable, which is the probability of a token containing at least one non-standard feature. As a random effect, we add the chatters' ID to account for individual variation between the participants as well as for their unbalanced contributions.

5. Results

5.1 Modeling Non-Standardness

Our model of non-standard language use (conceptualized in a binary way, i.e. the probability that a token contains at least one non-standard feature) lets the three social factors age, gender and education interact with each other, while adding a random effect for individual variation among the chatters. Table 2 shows the raw output of the model, i.e. the estimates and significance scores for the different levels of the factors, always in comparison to the reference category (older teenage boys in the General Education System/ASO). To evaluate a factor's significance as a whole (and not just in comparison to the reference group, but to all other levels as well), we performed extra Anova analyses. These results are shown in Table 3. Furthermore, we added extra effect tests using the 'Effects' package in R (Fox et al., 2016; Fox, 2003).

[1] Although the use of English words in Dutch conversations could also be related to the orality principle, we argue that it should be dealt with separately, since it is indicative of the extent to which youngsters connect with international chat culture.

	estimate	std. error	z	p	sig.
(Intercept)	-1.18023	0.03466	-34.06	< 2e-16	***
ageYoung	0.07991	0.02159	3.70	0.000215	***
genderGirls	-0.15825	0.04613	-3.43	0.000603	***
eduBSO	0.10890	0.05611	1.94	0.052263	.
eduTSO	0.03414	0.05153	0.66	0.507682	
ageYoung: genderGirls	0.24210	0.02486	9.74	< 2e-16	***
ageYoung: eduBSO	0.03532	0.04537	0.78	0.436242	
ageYoung: eduTSO	0.06508	0.02485	2.62	0.008826	**
genderGirls: eduBSO	0.17528	0.07487	2.34	0.019227	*
genderGirls: eduTSO	0.06882	0.07001	0.98	0.325614	
ageYoung: genderGirls: eduBSO	-0.09562	0.04986	-1.92	0.055141	.
ageYoung: genderGirls: eduTSO	-0.06868	0.02905	-2.36	0.018077	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 2: Output of the GLMM in comparison to the reference level (older boys ASO).

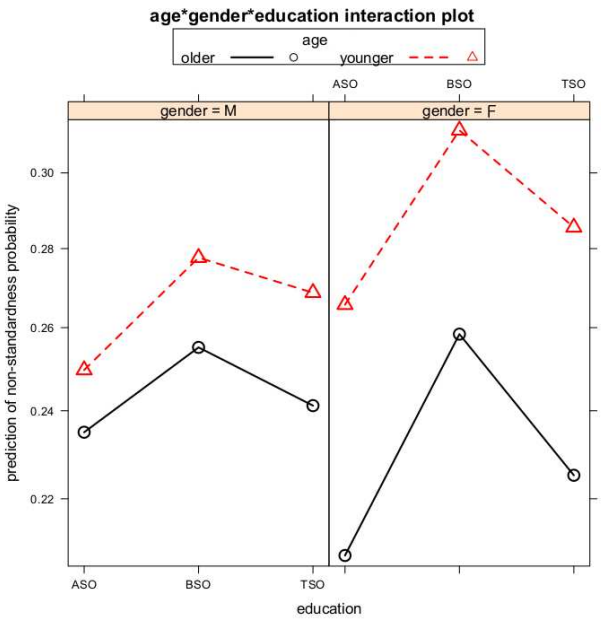


Figure 1: Interaction age*gender*education for non-standardness.

	Chisq	Df	Pr(>Chisq)	Sig.
age	2207.4925	1	< 2.2e-16	***
gender	0.4481	1	0.503223	
education	30.1873	2	2.786e-07	***
age:gender	232.6871	1	< 2.2e-16	***
age:education	10.6975	2	0.004754	**
gender: education	3.0855	2	0.213793	
age:gender:education	6.5998	2	0.036887	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 3: Output of the GLMM's Anova.

5.2. Effects and Interpretation

The different effects captured by the model are visualized in Figure 1 (i.e. the plot of the three-way interaction), on which the predicted probabilities for non-standardness are plotted for the different social variables.

The red dotted lines, representing the younger teenagers, are consistently higher than the black solid ones, representing the older teenagers, across all gender and education groups. This is a very consistent main effect for age, which is also significant (see Table 3, also confirmed by additional effect tests in which the other variables are kept constant): the younger teenagers (13-16 years old) use significantly more non-standard features than the older ones (17-20 years old). These results correspond to our expectations: non-standard language use is said to peak during adolescence, around the age of 16 ('the adolescent peak' – which is also the boundary between our two age categories) and thus decreasing as the teenagers age (Holmes, 1992, 184).

The two panels in Figure 1 represent the two genders, with the males on the left and females on the right. Clearly, at a younger age (compare the red dotted lines), girls outperform boys in non-standardness in each education type. However, this is no longer true at an older age, where girls only very slightly outperform boys in the Vocational System but use fewer non-standard features in the General and Technical Education Systems. The Anova (Table 3) and additional effect tests reveal that there is no significant *main* effect for gender, i.e. the model does not predict significantly different probabilities for non-standard features for girls compared to boys. However, the interactions between gender and age and between gender, age and education are significant. Consequently, gender is still an important factor in the model, as it is part of higher-order (interaction) terms which significantly impact on the response variable: in other words, in order to truly capture the gender effect, age and education have to be included in the analyses. As for the interaction between age and gender, Figure 1 shows that the decrease in non-standardness as the adolescents age is much stronger for the girls than for the boys. Again, these results correspond to our expectations, as in previous research, girls were found to converge more towards the adult standard as they grew older than boys (see Eisikovits, 2006, 43-44). Eisikovits ascribes this different age pattern to a difference between (working class) boys' and girls' attitude towards society when they graduate from high school; while accepting the responsibilities of adulthood, girls converge towards mainstream societal norms, whereas boys more strongly insist on their autonomy (2006, 48-49). We note that these preference patterns are confirmed for middle class participants by Vandekerckhove (2000, 302).

Finally, the Anova (Table 3) and additional effect tests reveal a significant main effect for type of education. The separate data points in Figure 1 reveal a consistent pattern across gender and age groups: the lowest probability of non-standardness is predicted for the teenagers in the General (theoretical) System (ASO), followed by the ones in the Technical (hybrid) System (TSO), and then by the ones in the Vocational (practical) System (BSO). Furthermore, additional effect tests showed that all three types are significantly different from each other. A possible explanation for these results concerns the level of proficiency in and familiarity with written standard Dutch in the different education types, which might increase as the school type becomes more theoretical. Apart from linguistic *skills*, *attitudinal* differences might be a factor too, as the prototypical chatspeak features may simply be more popular and considered to be *cooler* among students in the Vocational System. (For a more thorough analysis, see Hilte, Vandekerckhove & Daelemans, *fc*) Finally, the differences between education types are larger for the girls than for the boys. This could indicate a higher sensitivity for girls for this social factor.

Below, we present an alternative way to visualize the effects captured by the model. Figure 2 facilitates grasping the different 'age*gender' interactions in the three school systems. Clearly, in the more theoretical education types (General and Technical Education / ASO resp. TSO), the gender effect is opposite in the two age groups. At a younger age, the girls outperform the boys in non-standardness, but at a later age, they use fewer non-standard markers. In the Vocational System (BSO), however, the girls outperform the boys in non-standardness at a younger age and use more or less the same number of non-standard markers at an older age. Although there is still an interaction (girls' use of non-standard features decreasing more strongly than boys'), it is much less outspoken than the 'classical' pattern in the other two education types, and results in a convergence of the two genders rather than in an (opposite) divergence.

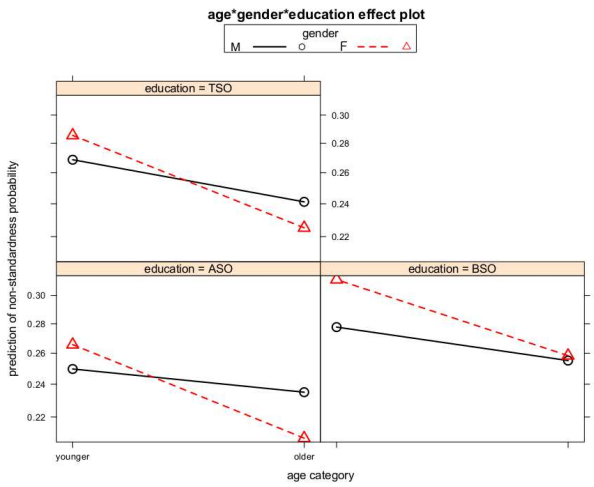


Figure 2: Interaction age*gender*education for non-standardness, alternative visualization.

6. Conclusion

We modeled Flemish adolescents' non-standard language use in their informal computer-mediated communication. We found that age, gender and education interact and influence the occurrence of non-standard features. Whereas the impact of age (lower frequencies in older teenagers' CMC) and education (lower frequencies for students in more theoretical educational tracks) might confirm expectations based on related research, the gender findings are quite surprising. The observation for the main effect of gender (i.e. no significant difference) does not correspond to previous research, as female language use is generally found to be more 'standard-oriented'.

However, this might be related to the operationalization of the notion of non-standardness in our research design: clearly expressive markers, which appear to be highly favored by women (see Hilte, Vandekerckhove & Daelemans, 2016, 31-32), might behave completely different in terms of indexing non-standardness from markers of regional non-standard speech. Consequently, a priority for future research will be the declustering of the set of 'non-standard' features and the consequent construction of different models for each subset, so that potential different preference patterns for these subsets can emerge. Still then, as we have shown in this preliminary study, gender cannot be studied in isolation, since the interactions with age and education are a prerequisite for a correct and nuanced evaluation of its impact.

7. Acknowledgements

We thank Giovanni Cassani, Dominiek Sandra and Koen Plevoets for their help and advise with the statistical modeling.

8. References

Androutsopoulos, J. (2011). Language Change and Digital Media: A Review of Conceptions and Evidence. In T. Kristiansen & N. Coupland (Eds.), *Standard Languages and Language Standards in a Changing Europe*. Oslo: Novus Press, pp. 145--161.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2017). Package 'lme4'. Url: https://cran.r-project.org/web/packages/lme4/lme4.pdf

Eisikovits, E. (2006). Girl-Talk/Boy-Talk: Sex Differences in Adolescent Speech. In J. Coates (Ed.), *Language and Gender: A Reader*. Oxford: Blackwell, pp. 42--54.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15).

Fox, J., Weisberg, S., Friendly, M., Hong, J., Andersen, R., Firth, D., & Taylor, S. (2016). Package 'effects'. Url: http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2016). Expressiveness in Flemish Online Teenage Talk: A

Corpus-Based Analysis of Social and Medium-Related Linguistic Variation. In D. Fišer, & M. Beisswenger (Eds.), *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia, 27-28 September 2016,* pp. 30--33.

Hilte, L., Vandekerckhove, R., & Daelemans, W. *(forthcoming).* Adolescents' Social Background and Non-Standard Writing in Online Communication.

Holmes, J. (1992). *An Introduction to Sociolinguistics.* London / New York: Longman.

Parkins, R. (2012). Gender and Emotional Expressiveness: An Analysis of Prosodic Features in Emotional Expression. *Griffith Working Papers in Pragmatics and Intercultural Communication,* 5(1), pp. 46--54.

Vandekerckhove, R. (2000). Structurele en sociale aspecten van dialectverandering. De dynamiek van het Deerlijkse dialect. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Wolf, A. (2000). Emotional Expression Online: Gender Differences in Emoticon Use. *Cyberpsychology & Behavior,* 3(5), pp. 827--833.

Investigating Interaction Signs across Genres, Modes and Languages: The Example of OKAY

Laura Herzberg, Angelika Storrer
Department of German Linguistics, University of Mannheim
Schloss, Ehrenhof West, D-68131 Mannheim
E-mail: lherzber@mail.uni-mannheim.de, astorrer@mail.uni-mannheim.de

Abstract

This paper presents results of a case study that compared the usage of OKAY across genre types (Wikipedia articles vs. talk pages), across modes (spoken vs. written language), and across languages (German vs. French CMC data from Wikipedia talk pages). The cross-genre study builds on the results of Herzberg (2016), who compared the usage of OKAY in German Wikipedia articles with its usage in Wikipedia talk pages. These results also form the basis for comparing the CMC genre of Wikipedia talk pages with occurrences of OKAY in the German spoken language corpus FOLK. Finally, we compared the results on the usage of OKAY in German Wikipedia talk pages with the usage of OKAY in French Wikipedia talk pages. With our case study, we want to demonstrate that it is worthwhile to investigate interaction signs across genres and languages, and to compare the usage in written CMC with the usage in spoken interaction.

Keywords: interaction signs, cross-lingual CMC study, Wikipedia talk pages

1. Background and Motivation

Interaction signs are elements that are not integrated in the syntactic structure of utterances, but serve as devices for discourse management: they can be used to express reactions to a partner's utterances or to display emotions. The category "interaction sign" was defined in Beißwenger et al. (2012), building on the grammar framework of the "Grammatik der deutschen Sprache" (henceforth *GDS*), which already included interjections ("hm", "well", "oh my god", "oops") and responsives ("yes", "no", "okay"). This framework was expanded with categories which have similar functions as interjections and responsives but typically occur in computer-mediated communication (CMC), e.g. emoticons, addressing terms (@USERNAME), action words ("lol", "grin") etc. (cf. Beißwenger et al., 2012).

The focus of this paper is on OKAY, which is an interesting object of study because it is used in many languages with a wide range of functions (cf. Figure 2). OKAY is not a CMC-specific interaction sign (like emoticons or "lol"), but is used in both written and spoken language. In our studies, the meta-lemma OKAY represents the different variants of spelling and pronunciation. Using OKAY as an example, we want to demonstrate that comparing the usage of interaction signs in speech corpora with its usage in written CMC corpora can yield interesting results. In our cross-genre and cross-lingual studies, we also explore which spelling variants are preferred by the users and whether these variants are compliant with spelling rules.

Most of the previous work on OKAY deals with spoken language: Schegloff/Sacks (1973) investigate OKAY in pre-closing sequences of spoken conversation. The studies of Beach (1993) and Bangerter et al. (2003) examine the usage of OKAY in phone calls. Levin/Gray (1983) describe the usage of OKAY in lecturer's presentations. Condon/Čech (2007) investigate the role of OKAY in decision making processes, comparing face-to-face interaction with CMC data. All these studies deal with the usage of OKAY in English. Studies on other languages are rare, although OKAY is used in many languages: Delahaie (2009) studies the usage of OKAY as an agreement marker in the learning of French as a foreign language. Kaiser (2011) investigates the usage of OKAY in German spoken doctor-patient communication. Cirko (2016) describes the usage of OKAY in German examination talks.

In our paper, we investigate the usage of OKAY across genre types (comparing CMC with text genres), across modes (comparing the usage in spoken interaction and written CMC), and across languages (comparing the same CMC genre in German and French). The cross-genre study builds on the results of Herzberg (2016), who compared the usage of OKAY in German Wikipedia article talk pages with its usage in Wikipedia articles. These results also form the basis for contrasting the usage of OKAY in written CMC and in spoken interaction (using data from the German speech corpus FOLK). Finally, we compare the usage of OKAY in the German Wikipedia talk pages with its usage in French Wikipedia talk pages.

2. Cross-genre study

2.1 Corpus Data

For the cross-genre study we compared data from two linguistically annotated Wikipedia corpora (cf. Margaretha/Lüngen, 2014): a corpus with German Wikipedia articles (Wiki-A-de; appr. 797 million tokens) and a corpus with German Wikipedia article talk pages (Wiki-D-de; 310 million tokens). Wikipedia articles represent a text genre (monologous structure, standard language etc.), while talk pages have features of CMC genres (dialoguous structure, informal writing style with non-standard language etc., cf. Storrer, 2017). The two corpora were downloaded from the Institute for the

German Language (IDS) and queried in RAPIDMINER-KOBRA[1].

2.2 Classification: Categories and Procedure

(1) First, we analysed the frequency of different spelling variants of OKAY in both corpora. The assumption was that OKAY is quite more frequent in the CMC corpus Wiki-D-de due to the dialogical structure and conversation-like nature of Wikipedia discussions. Different spelling variants had been queried and combined to draw the samples for the article and the talk pages (cf. Herzberg, 2016 for details). Since not all spelling variants occurred equally in both corpora, the two samples differ in their totals. The procedure resulted in a Wiki-A-de sample of 6,336 OKAY occurrences in total, and in a Wiki-D-de sample of 10,554 occurrences in total. All occurrences in both samples were manually checked and the false positives were sorted out. The distribution of true and false positives is illustrated in Table 1. It shows absolute frequencies as well as normalised frequencies as *pmw* values (occurrence per million words). Three types of false positives were distinguished: a) OKAY was mentioned as a word, e.g. in an article about interjections, b) OKAY was cited, e.g. in a song title or c) spelling variants of OKAY were homographic with abbreviations of proper names, such as a volcano ("Ok [...] is a shield volcano in Iceland")[2].
(2) Second, each spelling variant had been investigated individually. The two categories "conformant vs. non-conformant" and "speedy vs. non-speedy" served as objects of study. Because CMC writing is less norm-conformant, we expected to find spelling variants that do not comply with the German spelling norm. In German *okay*, *Okay*, *o. k.* and *O. K.* are the norm-conformant spelling variants[3]. It has to be noted, that the variants "o. k." and "O. K." have to display a blank space between *O* and *K* to be norm-conformant. Therefore, the spelling variants *ok*, *OK*, *Ok*, *o.k.*, and *O.K.* are non-conformant spellings.
Another hypothesis was that CMC users prefer "speedy" spelling variants (*ok*, *Ok*, *OK*) because speed writing is a general feature of CMC. We classified *ok*, *OK* and *Ok* as "speedy" and all other variants as "non-speedy".

2.3 Results and Discussion

(1) The results of the cross-genre frequency study on OKAY are presented in Table 1.

	true positives		false positives	
	abs.	pmw	abs.	pmw
Wiki-A-de	25	0.03	6,311	7.92
Wiki-D-de	8,248	26.62	2,306	7.44

Table 1: Distribution of true and false positives of OKAY in the German Wikipedia.

As expected, OKAY is quite more frequent in the CMC corpus (talk pages) than in the text corpus (Wikipedia articles). Interestingly, the two corpora considerably differ in their number of false positives: in the CMC sample, 2,306 (21.8 %) were classified as being false positives. In the text sample 6,311 (99.6 %) occurrences of OKAY turned out to be false positives: only 25 (0.4 %) of all occurrences were true positives. As 25 items is a very small data set, we restricted our studies on the frequency of spelling variants on the CMC sample.
(2) Table 2 shows the results of the studies on norm-conformance and frequency of OKAY spelling variants in the German CMC corpus Wiki-D-de and in the French CMC corpus Wiki-D-fr. In this section, we discuss the results of the German data; the cross-lingual aspects are treated in section 4.3[4].

Spelling Variant	Norm-conformance		Frequency Wiki-D-de		Frequency Wiki-D-fr	
	DE	FR	abs.	pmw	abs.	pmw
OK			17,796	57.43	9,281	67.69
ok			16,048	51.78	5,476	39.94
Ok			15,431	49.79	7,495	54.67
okay	✓		8,421	27.17	86	0.63
Okay	✓		8,287	26.74	163	1.19
o. k.	✓		96	0.31	0	0
O. K.	✓		86	0.28	6	0.04
o.k.			80	0.26	0	0
O.K.		✓	21	0.07	3	0.02

Table 2: Frequency and norm-conformance of OKAY spelling variants in German and French.

The results in Table 2 clearly support the assumption that non-conformant variants are more frequently used than the conformant ones in the German CMC corpus. Moreover, the results support the hypothesis that the three speedy variants *ok*, *OK*, and *Ok* are preferred, although they do not conform to German spelling rules.

3. Cross-modal study

3.1 Corpus Data

There are significant differences between the usage of interaction signs in spoken and written language. In spoken interaction intonation plays a crucial role in interpreting a positive, negative, or doubting evaluation expressed by an interaction sign. Interaction signs are relevant for organizing turn-taking in spoken interaction: hearers use interaction signs to encourage the floor holder to continue (so-called "continuers", cf. Schegloff, 1982 p. 81). While these functions have been widely investigated in spoken language (see cited works in section 1), studies on CMC or cross-modal studies are still rare.

[1] Details on the queries are provided in Herzberg (2016).
[2] Cf. https://en.wikipedia.org/wiki/Ok_(volcano) [15.06.17].
[3] Cf. Duden-Rechtschreibung, 2013 p. 781.
[4] We integrated the French data in Table 2 in order to save space. The table presents absolute frequencies as well as normalised frequencies as *pmw* values.

In our cross-modal case study, we compared data from the CMC corpus Wiki-D-de (310 million tokens), with spoken interaction data taken from the German FOLK corpus (1.9 million tokens). The speech data was queried automatically via the DGD.

3.2 Classification: Categories and Procedure

(1) We distinguished between two main functional categories: OKAY as a syntactic unit (used in predicative, adverbial, attributive function or as a noun) and OKAY as an interaction sign (used in responsive, reactive, interrogative, and structural function). In Herzberg (2016), all true positives in the Wiki-D-de sample described in section 2 (8,248 occurrences in total, cf. Table 1) have been classified as follows: 5,045 (61.2 %) occurrences are used as interaction signs and 3,203 (39.8 %) as syntactic units. An interesting finding concerns the functional category "responsive", i.e. a (positive) answer to a polar question. This function is described as being the main function of OKAY in the German grammar GDS (1997) p. 63. However, the study revealed that only a very small amount (20 occurrences, i.e. 0.4 % of all interactive OKAY occurrences) in the examined Wiki-D-de data were used as responsives. We assumed that this mismatch between the Grammar description and our data was due to the fact that the classifications in this Grammar refers to the usage of OKAY in spoken interaction. We thus used data from the FOLK corpus to investigate whether the responsive function of OKAY is a main function in spoken interaction. We manually checked how often the responsive function of OKAY was used in a FOLK corpus sample with 1,500 occurrences of OKAY.

(2) In a second study we analysed the positions of OKAY in samples taken from FOLK, Wiki-D-de, and the French Wikipedia talk pages (Wiki-D-fr; 137 million tokens) with 500 occurrences in each sample. These samples only contain true positives; false positives have been manually sorted out. The data has then been classified according to four positional categories: initial (directly at the beginning of a [5]post/utterance); middle (within a post/utterance); final (end of a post/utterance) and standalone (OKAY forms post/utterance).

3.3 Results and Discussion

(1) The analysed FOLK sample had a similar outcome as the study on the CMC data: Only 15 (1 %) of the 1,500 examined occurrences are used as responsives. In both corpora, the responsive function which is claimed to be the main function in the GDS grammar description, only rarely occurs in both written CMC and spoken language. These results demonstrate that it is worthwhile to further evaluate assumptions about the functions of interaction signs on the basis of corpus data.

[5] Following the proposals of the TEI CMC group, we use the term "posts" for units in CMC interaction (cf. Beißwenger et al., 2012). The segments in spoken interaction are termed as "utterances".

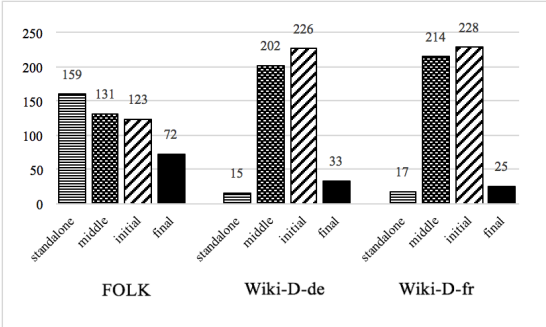


Figure 1: Positional distribution of OKAY in spoken and CMC interaction[6].

(2) The results of our comparison of positions in Figure 1 reveal significant differences between the two modes. Whereas OKAY is used variably and nearly equally often in German utterances across the three categories "standalone" (32.8 %), "middle" (27.0 %) and "initial" (25.4 %), the German CMC data presents a different picture: OKAY is preferably used at the beginning (47.5 %) and within (41.6 %) a post. These two positions make up nearly 90 % of all investigated occurrences. Interestingly, the positional distribution patterns in the German and French CMC data are quite similar[7].

There are two possible explanations for these results that have to be verified in further work: (a) The standalone position is typical for "continuers" (see above) and the final position is typical for the usage of OKAY as a tag question. Both functions are particularly relevant for organising turn-taking in spoken interaction. This may explain the lower rate of standalone and final positions in the CMC data, where turn-taking mechanism are substituted by other mechanisms of interaction management (cf. Beißwenger, 2008). (b) As it is shown in Figure 3, OKAY is mostly used as an interaction sign in the speech data from FOLK. In the two CMC corpora however, OKAY is also used as a syntactic unit. These syntactically integrated units (nouns, adverbials, predicatives) often occur in a middle position. This may be one factor to explain the higher rate of middle positions in CMC corpora in the results presented in Figure 1.

To get a clearer image of the differences in the usage of OKAY in spoken and written interaction, we want to annotate the functional and the positional categories presented in Figure 2 on two different layers and explore correlations between the positional und functional categories in more detail.

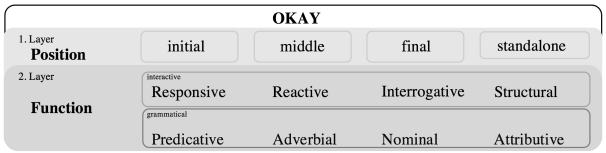


Figure 2: Formal and functional annotation categories.

[6] The figures contain absolute frequencies of true positives.

[7] Cross-lingual aspects are treated in section 4.3.

4. Cross-lingual study

4.1 Corpus Data

In our cross-lingual study we compared the corpus of German Wikipedia talk pages (Wiki-D-de; 310 million tokens) with a corpus of French Wikipedia talk pages (Wiki-D-fr; 137 million tokens). Both corpora are available within the German Reference Corpus DeReKo at the IDS. The data has been queried automatically via COSMAS II.

4.2 Classification: Categories and Procedure

(1) In a first study, we manually classified two samples of German and French, each containing 500 OKAY occurrences, in three categories: "syntactic units", "interaction signs" (cf. 3.2) and "others". We assumed that the usage of OKAY as a syntactic unit, signalling a deeper integration of the loan word in the host language system, is less frequent in the French corpus.
(2) The focus on the second investigation was again on spelling variants. We expected that the speedy and non-conformant variants are also preferred in the French CMC corpus. Similar to the study of German, *ok*, *OK* and *Ok* were classified as speedy variants whereas *okay*, *Okay*, *o.k.*, *O.K.*, *o. k.* and *O. K.* are non-speedy variants. In French, only the variant *O.K.* is conformant[8]. Therefore, the variants *okay*, *Okay*, *ok*, *OK*, *Ok*, *o.k.*, *O. K.* and *o. k.* were classified as being non-conformant.
(3) We integrated the French CMC sample in our cross-modal study on positional differences between spoken and written CMC, described in section 3.2., to investigate the distribution patterns in the CMC data of both languages.

4.3 Results and Discussion

(1) The results in Figure 3 support our assumption that OKAY is less frequently used as a syntactic unit in French than in German[9].

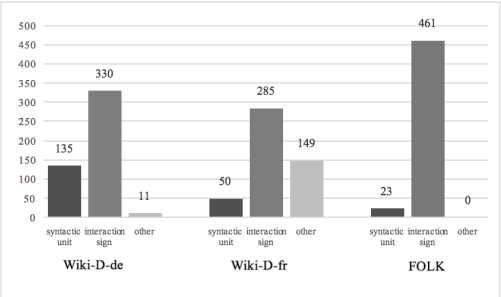


Figure 3: Functional distribution of OKAY.

In the German CMC data, 135 (28.4 %) occurrences of OKAY were tagged as syntactic units and 330 (69.3 %) occurrences as interaction signs. In French, 285 (58.9 %) occurrences were classified as interaction signs whereas 50 (10.3 %) were tagged as syntactic units. The French data included a considerably high amount of OKAY occurrences that could not clearly be classified as either interactive or syntactic (149 occurrences; 30.8 %)[10].
The aforementioned results had been achieved by manually checking and tagging the samples. Using a tagger that automatically assigns part-of-speech (POS) tags to distinguish between interactive or grammatical usages of OKAY did not achieve satisfactory results. The applied taggers either tagged all occurrences as being interactive, e.g. in FOLK, or as being grammatical, e.g. in Wiki-D-de. Studying OKAY exemplifies that there is still a need for improvement in the field of POS-tagging (cf. Lüngen et al., 2016 for details).
(2) The results of our cross-lingual study on the frequency of spelling variants are presented in Table 2 of section 2.2. The most frequent variants in the corpora are non-conformant, but support speed-writing. In both languages the non-speedy variants including a space (*o. k.* and *O. K.*) are rarely used. The variants *okay* and *Okay* are less frequent in French than in German, where these forms are norm-conformant.
(3) In terms of the positional distribution, shown in Figure 1, there is a clear distinction between speech and written CMC corpora. The distributional patterns in the French and the German CMC data do not differ to a vast extent and therefore seem to be language independent.

5. Conclusion

We investigated the usage of OKAY across genre types (German Wikipedia articles vs. talk pages), modes (German spoken vs. written interaction), and across languages (German vs. French CMC). The cross-genre study illustrated that OKAY is quite more frequently used in the CMC genre and that speedy writing variants are preferred over rule-conformant non-speedy ones. The cross-lingual study revealed that the grammatically integrated functions of OKAY occur more frequently in the German than in the French data. This may be an effect of the French language policy that recommends to avoid English loan elements. By comparing the frequency of spelling variants we found that the "speedy" variants are highly preferred in French and in German, although these variants are not rule-conformant. The cross-modal study showed that the function of a responsive, described as being the main function in the GDS grammar, is rarely used in both written and spoken corpora. It is thus worthwhile to investigate the functions of OKAY on the basis of corpus data. The results of the comparison of positional categories in Figure 2 revealed that the distribution patterns in the French and the German CMC corpora are quite similar, whereas the patterns in the CMC corpora differ considerably from the distribution in the spoken language corpus FOLK. Further work will study the usage of interaction signs in spoken and written CMC interaction on the basis of a more fine-grained annotation of functional categories.

[8] Cf. Le Petit Robert, 2017 p. 1736.
[9] The samples contain absolute frequencies of true positives.
[10] Examples are posts containing elliptical constructions like "Donc, OK pour moi" or "OK pour la date de la mort".

6. References

Bangerter, A.; Clark, H.H.; Katz, A.R. (2003). Navigating Joint Projects in Telephone Conversations. In *Discourse Processes* 37, pp. 1-23.

Beach, W. (1993). Transitional regularities for 'casual' "Okay" usages. In *Journal of Pragmatics* 19, pp. 325-352.

Beißwenger, M. (2008). Situated Chat Analysis as a Window to the User's Perspective: Aspects of Temporal and Sequential Organization. In J. Androutsopoulos, M. Beißwenger (Eds.), *Data and Methods in Computer-Mediated Discourse Analysis* (= Language@Internet 5).

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

Delahaie, J. (2009). Oui, voilà ou d'accord? Enseigner les marqueurs d'accord en classe de FL. In *Synergies Pays Scandinaves* 4, pp. 17-34.

Duden-Rechtschreibung (2013). *Duden – Die Grammatik*. 26., völlig neu erarbeitete und erweiterte Auflage. Berlin: Bibliographisches Institut GmbH. (= Band 1 – Der Duden in 12 Bänden).

Condon, S.L.; Čech, C.G. (2007). OK, next one: Discourse markers of common ground. In A. Fetzer, K. Fischer (Eds.), *Lexical Markers of Common Grounds*. London: Elsevier, pp. 18-45.

GDS (1997). *Grammatik der deutschen Sprache*. Zifonun, G.; Hoffmann, L.; Strecker, B.; et al. (Eds.). 3 Bände. Berlin/New York: de Gruyter.

Herzberg, L. (2016). *Korpuslinguistische Analyse interaktiver Einheiten: das Beispiel okay*. Master thesis. University of Mannheim.

Kaiser, J. (2011). *okay in ärztlichen Gesprächen – eine linguistische Gesprächsanalyse*. State examination thesis. Ruprecht-Karls-University Heidelberg.

Le Petit Robert (2017). *Dictionnaire Alphabétique Et Analogique De La Langue Française*. Paris: Dictionnaires Le Robert.

Levin, H.; Gray, D. (1983). The Lecture's OK. In *American Speech* 58, pp. 195-200.

Lüngen, H.; Beißwenger, M.; Herold, A.; Storrer, A. (2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In S. Dipper, F. Neubarth, H. Zinsmeister (Eds.), *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), pp. 156-164.

Margaretha, E.; Lüngen, H. (2014). Building linguistic corpora from Wikipedia articles and discussions. In *Journal of Language Technologie and Computational Linguistics JLCL* 29(2), pp. 59-83.

Schegloff, E.A.; Sacks, H. (1973). Opening up Closings. In *Semiotica* 8, pp. 289-327.

Schegloff, E.A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk*, Washington, DC: Georgetown University Press, pp. 71-93.

Storrer, A. (2017). Grammatische Variation in Gespräch, Text und internetbasierter Kommunikation. In M. Konopka, A. Wöllstein (Eds.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, Berlin/New York: de Gruyter, pp. 105-125.

Corpus tools and resources:

COSMAS I/II: Corpus Search, Management and Analysis System. Institute for the German language Mannheim, http://www.ids-mannheim.de/cosmas2/.

DeReKo: Das Deutsche Referenzkorpus. Institute for the German language Mannheim, http://www.ids-mannheim.de/kl/projekte/korpora/.

DGD: Datenbank gesprochenes Deutsch. Institute for the German language Mannheim, http://agd.ids-mannheim.de/folk.shtml.

FOLK: Forschungs- und Lehrkorpus für gesprochenes Deutsch. Institute for the German language Mannheim, http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome.

RAPIDMINER-KOBRA: RapidMiner Software, www.rapidminer.com and KobRA Plugin, http://www.kobra.tu-dortmund.de/mediawiki/index.php?title=Software.

Wiki-A-de: Corpus with all articles of the German Wikipedia (Version 17.11.2015). Institute for the German language Mannheim, http://corpora.ids-mannheim.de/pub/wikipedia-deutsch/2015/.

Wiki-D-de: Corpus with all article talk pages of the German Wikipedia (Version 17.11.2015). Institute for the German language Mannheim, http://corpora.ids-mannheim.de/pub/wikipedia-deutsch/2015/.

Wiki-D-fr: Corpus with all article talk pages of the French Wikipedia (Version 17.11.2015). Institute for the German language Mannheim, http://corpora.ids-mannheim.de/pub/wikipedia-fremdspr/2015/.

Anonymisation of the Dortmund Chat Corpus 2.1

Harald Lüngen[1], Michael Beißwenger[2], Laura Herzberg[3], Cathrin Pichler[2]
[1]Institut für Deutsche Sprache, R5 6-13, D-68161 Mannheim
[2]Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6-8, D-45127 Essen
[3]Universität Mannheim, Germanistische Linguistik, Schloss, D-68131 Mannheim
E-mail: luengen@ids-mannheim.de, michael.beisswenger@uni-due.de, lherzber@mail.uni-mannheim.de, cathrin.pichler@tu-dortmund.de

Abstract

As a consequence of a recent curation project, the Dortmund Chat Corpus is available in CLARIN-D research infrastructures for download and querying. In a legal expertise it had been recommended that standard measures of anonymisation be applied to the corpus before its republication. This paper reports about the anonymisation campaign that was conducted for the corpus. Anonymisation has been realised as categorisation, and the taxonomy of anonymisation categories applied is introduced and the method of applying it to the TEI files is demonstrated. The results of the anonymisation campaign as well as issues of quality assessment are discussed. Finally, pseudonymisation as an alternative to categorisation as a method of the anonymisation of CMC data is discussed, as well as possibilities of an automatisation of the process.

Keywords: Corpora, Computer-mediated communication, Anonymisation

1. Introduction

In the CLARIN-D curation project "Integration of the Dortmund Chat Corpus into CLARIN-D" (Lüngen et al., 2016), a legal expertise was sought to clarify issues concerning the possibility to republish the material, which had been collected between 2004-2008 partly without written consent of the participants, in the CLARIN-D infrastructures (Lüngen et al., 2016; Beißwenger et al., 2017). The legal expertise was composed by the company iRights.law (Berlin), which specialises on legal issues concerning digital media. Below is a summary of the recommendations that were given to the hosting institutions. They follow from considerations of personality and data protection rights. Other legal statuses like copyright and intellectual property rights were also considered in the expertise, but are not discussed here.

1. Remove the chats which originally came from psycho-social counselling platforms completely (10 out of 480 logfiles)
2. Grant access to chats collected from closed platforms only for authorised scientific use
3. Apply "standard measures" of anonymisation to all chat files
   a. Randomise/replace host names, nicknames, place names, and platform names
   b. Remove or permute the time stamps

Medlock (2006) distinguishes between categorisation and pseudonymisation. The latter is a procedure of permutation or replacement of the sensitive references with instances of the same ontological category (e.g. replacing occurrences of the male name *Holger* with the male name *Werner*. To get an idea of possible types and categories of sensitive references in chat and CMC, we also looked at the anonymisation in previous CMC corpus projects. Among them was no project dealing with chat data, however an email corpus (Medlock, 2006), a Facebook corpus (DiDi, 2015), two SMS corpora (Panckhurst, 2013; Ueberwasser, 2015), as well as one spoken conversation corpus (FOLK, cf. Winterscheid, 2015). In all the CMC corpora, anonymisation was realised as categorisation, only in the spoken corpus was it realised as pseudonymisation.

2. Anonymisation by categorisation

Categorisation preserves some of the information so that a corpus can still reasonably be used for linguistic analyses. It implies the replacement of a sensitive reference with a placeholder string that indicates its ontological category, such as *Person_name* or *Place_name*. Since most of the references that had to be anonymised in the chat corpus are names, we firstly included the five named entity categories PER, ORG, LOC, GPE, OTH from the TüBa-D/Z treebank (Telljohann et al., 2004), which had already been used in NER experiments with DeReKo (Bingel & Haider, 2015) in our category inventory. Because these five NER cateogories are relatively coarse-grained, and because the annotations in the original chat corpus resource contained already more specific information, we extended the set by the categories NICK (for nickname, a subcategory of PER) and ROOM (for chat room). Moreover, we added the category GEO_DE for a noun or adjective derived from a LOC or a GPE (a union of the categories _GeoNE_ and _GeoADJA_ in DiDi, 2015). Besides these, three categories for more formal references were added: URL (for a web address), email (for an email address), and NUMBER (for any kind of referencing number, see Table 1 for examples). Following Winterscheid (2015) and DiDi (2015), we also introduced the two rarer categories IMPLICIT (for an implicit reference), and CITATION (for a quote by which an individual might be identified). The 13 anonymisation categories used are shown in Table 1 with their definitions, example(s), and the source of or inspiration for the category.

#	Category short form	Category long form and definition	Examples	Source
1	PER	PERSONNAME: A first name or second name or a sequence out of first name and second name	“Erwin”, “Meike”, “Anna Hein”	TüBa-DZ (Telljohann et al., 2004)
2	NICK	NICKNAME: User name chosen by a chat participant, or a variant therof	“superman”, “lela2” ,“Tiger”, “Lan5”, “KainPech”	DO chat corpus
3	ORG	ORGANISATIONNAME: Company (e.g. the employer of a participant), sports club, institute, university etc.	“RUB”, “John Deere”, “ASV Schifferstadt”	TüBa-DZ
4	LOC	LOCATIONNAME: A place or area which is not a GPE, e.g. mountains, valleys, rivers, roads, motorways, etc.	“Augustaanlage”, “Neckar”, “Königstuhl” “A6”	
5	GPE	GEOPOLITICALENTITTYNAME: A geo-political entity, i.e. a place or area of which the borders are officially defined, i.e. cities, municipalities, countries, states, suburbs etc., including their spelling variants and abbreviations	“Mannheim”, “NRW”, “Italien” “doaaadmund” “DO”	
6	GEO_DE	GEODERIVATIONNAME: Noun or adjective that is morphologically derived from a (mostly GPE or LOC) name and which expresses an association or a quality (adjectives) or a group or inhabitants (noun)	“Mannheimer”, “Mannheimerinnen”, “Gelbfüßler”	
7	OTH	OTHERNAME: Residual category for all sensitive names and references that cannot be categorised otherwise	“unicum”	
8	ROOM	CHATROOMNAME: Name of a chatroom	“Welcome”, “blue”	DO chat corpus
9	URL	WWWURL: Web address	“http://www.ids-mannheim.de/”	
10	EMAIL	EMAIL: Email address	“fix@ids-mannheim.de”	
11	NUMBER	NUMBER: Any number or code that can be associated with a person: e.g. house number, serial numer, postal code, telephone number, passport number, account number, IP address, password	“0621/1581418”, “10.0.1.45”, “68161”	
12	IMPLICIT	IMPLICIT: Implicit reference: Revealing descriptions and pieces of information from which the identity of a chat participant or a third party can be inferred (e.g. someone’s job)	„IT-Operator“	FOLK (Winterscheid, 2015)
13	CITATION	CITATION: A quote, e.g. from a song, which can be used to identify a chat participant or a third party		FOLK

Table 1: Anonymisation categories in the Dortmund Chat Corpus 2.1.

3. Anonymisation campaign and results

The major bulk of the anonymisation by categorisation process was carried out by four student assistants of Mannheim University, Duisburg-Essen University and the Institute for the German Language (IDS), Mannheim. The sensible references that had not been pre-annotated were identified and annotated with the category inventory in Table 1 using the "author mode" of the XML editor Oxygen. The campaign lasted from August till December 2016 (five months) and took approximately 625 hours of manual annotation work. Subsequently an XSLT post-processing step was implemented to insert the replacement strings and to provide TEI annotation in terms of the elements <name> and <ref>.

Listings 1-4 contain XML code snippets that show what the result of the anonymisation looks like in CLARIN-D TEI (cf. Lüngen et al., 2016).

```
<particDesc>
    <!-- 1301005 -->
    <listPerson>
      <!-- ... -->
      <person role="celebrity" xml:id="A03">
        <persName type="nickname">Günther Beckstein</persName>
        <sex evidence="estimated">male</sex>
      </person>
      <!-- ... -->
      <person role="moderator" xml:id="A04">
        <persName type="nickname">[_MALE-MODERATOR-A04_]</persName>
        <sex evidence="estimated">male</sex>
      </person>
      <!-- ... -->
      <person role="participant" xml:id="A07">
        <persName type="nickname">[_FEMALE-PARTICIPANT-A07_]</persName>
        <sex evidence="estimated">female</sex>
      </person>
      <person role="participant" xml:id="A08">
        <persName type="nickname">[_PARTICIPANT-A08_]</persName>
        <sex evidence="estimated">unknown</sex>
      </person>
      <!-- ... -->
    </listPerson>
  </particDesc>
```

Listing 1: Anonymisation of metadata (participant list). Mentions of celebrities and politicians are from the public sphere and are not anonymised.

```
<post auto="false" rend="color:lime" type="event" who="#A14" xml:id="m487">
    <name corresp="#A14" type="nickname">
      <w lemma="[_PARTICIPANT-A14_]" type="NE" xml:id="m487.t1"> [_PARTICIPANT-
      A14_]</w>
    </name>
    <w lemma="werden" type="VAFIN" xml:id="m487.t2">wird</w>
    <w lemma="schlecht" type="ADJD" xml:id="m487.t3">schlecht</w>
</post>
```

Listing 2: Anonymisation of a nickname without role entry in participant list.

```
<w lemma="auch" type="ADV" xml:id="m40.t1">auch</w>
<w lemma="bei" type="APPR" xml:id="m40.t2">bei</w>
<w lemma="die" type="ART" xml:id="m40.t3">den</w>
<name type="GEO_DE">
    <w lemma="[_GEODERIVATIONNAME_]" type="NN"
xml:id="m40.t4">[_GEODERIVATIONNAME-4_]</w>
</name>
```

Listing 3: Anonymisation of a derivation of a name (like *Düsseldorfern*).

```
<post auto="false" rend="color:#808080" synch="#t427" type="standard" who="#A26" xml:id="m576">
    <w lemma="wollen" type="VMFIN" xml:id="m576.t1">willst</w>
    <w lemma="du" type="PPER" xml:id="m576.t2">du</w>
    <w lemma="eine" type="ART" xml:id="m576.t3">ne</w>
    <w lemma="Therapie" type="NN" xml:id="m576.t4">therapie</w>
    <ref corresp="#A31" type="addressingTerm">
        <w lemma="@" type="ADRIND" xml:id="m576.t5">@</w>
        <w type="NE" xml:id="m576.t6">[_MALE-PARTICIPANT-A31_]</w>
    </ref>
    <w lemma="ich" type="PPER" xml:id="m576.t7">ich</w>
    <w lemma="studieren" type="VVFIN" xml:id="m576.t8">studier</w>
    <ref type="IMPLICIT">
        <w type="NE" xml:id="m576.t9">[_IMPLICIT-1_]</w>
    </ref>
</post>
```

Listing 4: Anonymisation of an implicit reference.

Below are the stats of the annotation of categories in the hole corpus - remember that the chat corpus contains roughly 1 million tokens

Category	# Occurrences
NICK:	30,022
ROOM:	2,409
OTH:	1,819
URL:	1,742
GPE:	1,309
PER:	838
ORG:	741
GEO_DE:	178
NUMBER:	169
IMPLICIT:	130
LOC	107
EMAIL:	50
CITATION:	5
∑ =	39,519

Table 2: Occurrences of categories for sensitive references in Dortmund Chat Corpus 2.1.

4. Quality assessment

To ge some impression of the agreement between our coders, we asked all four of them to annotate the chat logfile with the ID 1102001 immediately after the training session. The file contains 675 chat posts, and the union of the sensitive references identified by the four coders contained 126 references. This figure was subsequently used as N (number of items to be coded) in the Kappa calculation described in the following. We calculated Fleiss' Kappa using the IRR package for the programming language R (function *kappam.fleiss*)[1]. The agreement between the four coders was κ=0.582. According to the interpretation scale by Landis & Koch (1997), this corresponds to "moderate" agreement.

A closer inspection of the disagreements revealed that

[1] Cf. https://cran.r-project.org/package=irr [22.06.2017].

Coder 1 was the source of an unusual great deal of the mismatches. For instance, in cases where the name of a person was given fully as first name + last name, Coder 1 had, contrary to the training instructions, always annotated them as separate instances. There were at least 15 such first name + last name combinations. (Coder 1 was subsequently made aware of this error i.e. before anonymising her share of the corpus.)

We additionally calculated Fleiss' Kappa among the remaining three coders Coder 2, Coder 3, and Coder 4 only. For them, Kappa was found to be κ=0.827 after all, which according to Landis & Koch (1977) can be interpreted as "almost perfect agreement".

Because of these results, we believe that our method is appropriate for achieving an anonymisation of the chat corpus that conforms to legal standards as put forth in the legal expertise. Ideally, one would have had more material annotated by all four coders, and calculated the inter-rater reliability not only at the beginning of the annotation campaign but also in the middle and at the end of it. Moreover, it would have been interesting if we had even checked for *intra*-rater reliability of each or at least some of the coders. Unfortunately, in the present campaign there was no more time for coding, coordination, and evaluation work. But for future projects, this should be kept in mind.

5. Discussion

During the campaign we noticed that for several chats, a more fine-grained category scheme would have been desirable from a discourse linguist's point of view. In some chats, for instance, many locations were mentioned, and in the anonymised version one would have wished to have more information on the kind of location discussed (e.g. restaurant, shop, school) available. On the other hand, a more complex encoding scheme usually affects inter-rater agreement to the negative. A simple solution to this could be to allow coders to add free information strings.

Another way to address all kinds of problems with the category scheme could be to aim for corpus *pseudonymisation* such as in the spoken conversation corpus FOLK. However, to achieve a full pseudonymisation is even more costly than our anonymisation by categorisation method, and besides has its own drawbacks, such as the possibility of introducing inconsistencies in the dialogue.

Finally, it seems obvious that we need an automatisation of the anonymisation process. A campaign like the one described above is simply not feasible for larger corpora, and the need for anonymisation is potentially given with many kinds of CMC, even web corpora. However, the task is non-trivial and comprises more than standard Named Entity Recognition. The anonymised Dortmund Chat Corpus 2.1 can also serve as training data for future developments of corpus anonymisation tools.

6. References

Beißwenger, M., Lüngen, H., Schallaböck, J., Weitzmann, J.H., Herold, A., Kamocki, P., Storrer, A., Wildgans, J. (2017, to appear). Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In Michael Beißwenger (Ed.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin/Bew York: de Gryuter (Empirische Linguistik).

Bingel, J., Haider, T. (2014). Named Entity Tagging a Very Large Unbalanced Corpus. Training and Evaluating NE Classifiers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. Reykjavik: European Language Resources Association (ELRA).

DiDi (2015). *Beschreibung der Anonymisierung im DiDi-Korpus*. Available online http://www.eurac.edu/en/research/autonomies/commul/Documents/DiDi/DiDi_anonymisation_DE.pdf.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. In *Psychological Bulletin* 76 (5), pp. 378–382.

Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. In *Biometrics* 33, pp. 159–174.

Lüngen, H., Beißwenger, M., Herold, A., Storrer, A. (2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In S. Dipper, F. Neubarth & Heike Zinsmeister (Eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 156-164.

Medlock, B. (2006). An Introduction to NLP-based Textual Anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*. Genoa: European Language Resources Association (ELRA). Available online http://www.lrec-conf.org/proceedings/lrec2006/.

Panckhurst, R. (2013). A large SMS Corpus in French: from Design and Collation to Anonymisation, Transcoding and Analysis. In *Procedia - Social and Behavioral Sciences* 95, pp. 96-104.

Telljohann, H., Hinrichs, E., Kübler, S. (2004). The TüBa-D/Z Treebank: Annotating German with a Contet-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Ueberwasser, S. (2015). *Anonymisation* (SMS4Science.ch). Available online https://sms.linguistik.uzh.ch/bin/view/SMS4Science/Anonymisation.

Winterscheid, J. (2015). GAIS Web: Maskierung. Document in the IDS Project. Available online http://prowiki.ids-mannheim.de/bin/view/GAIS/MasKierung.

Emoticons as multifunctional and pragmatic Resources: a corpus-based Study on Twitter

Stefania Spina
University for Foreigners of Perugia
Piazza Fortebraccio, 4 – Perugia, Italy
E-mail: stefania.spina@unistrapg.it

Abstract

Emoticons play an important role in digital written communication: they can serve as markers either of emotions or social relationship and familiarity, and they can intensify or downgrade the pragmatic force of a text.
The aim of this study is to investigate the use of emoticons in Twitter by Italian users, and to verify, by relying on corpus data and on statistical methodologies, some of the prevailing opinions on the use of emoticons: that they are technically-driven resources, that they are mostly used by young people, and more often by females, and that they are superficial and easy ways of expressing emotions using images instead of words.
A mixed-effects model analysis has shown that the use of emoticons on Twitter is affected by a complex interaction of cultural, technological, situational and sociolinguistic variables.

Keywords: emoticons, Twitter, mixed-effects models.

1. Introduction

Emoticons (the graphic signs, such as the smiley face, that often accompany digital written communication) are an integral part of digital culture since its beginnings: they have followed its development over the last decades, evolving alongside with the rapid spread of new written communication environments, such as social media or text messaging systems.
Many studies have outlined the key role of emoticons in digital written communication (e.g. Amaghlobeli, 2012; Baron, 2009; Danesi, 2016; Derks, Bos & von Grumbkow, 2007; Dresner & Herring, 2010; Spina, 2016; Vandergriff, 2014; Walther & D'Addario, 2001; Yus & Yus, 2014): as people use writing more and more instead of face to face interactions or phone calls, the need for overcoming limitations in communicating emotional tone arises. The widespread use of emoticons allows to convey nonlinguistic information that in face-to-face communication is expressed through facial expression and other bodily indicators. Emoticons, therefore, are primarily "emotion icons": additional opportunities to convey emotions through the use of graphic symbols, directly mapped onto facial expressions.
The role of emoticon in digital written communication, however, is much more nuanced and not limited to the expression of emotions. Following Dresner & Herring (2010), Vandergriff (2014), and Spina (2016), they are developing at least two other important pragmatic functions, that are not necessarily mapped onto facial expressions, or aimed at the expression of emotions:

- the function of social markers of familiarity and empathy. In this sense, they are relational icons, that promote rapport and play a social and affiliative role;
- the function of markers of the pragmatic force of a text, aimed at intensifying or downgrading its meaning. In this function, they are contextualization cues (Gumperz, 1982; Auer, 1992), that provide information on how to interpret the verbal message.

As a consequence, emoticons are multifunctional and highly context-sensitive resources, whose different functions most often tend to overlap and to occur simultaneously within the use of a single emoticon. This claim is illustrated by the examples (1) and (2):

(1)
@user2 Ci vieni in piscina domani?
[@user2 Are you coming to the pool tomorrow?]

@user1 No :-(

(2)
@user2 Hai visto la foto del mio profilo?
[@user2 Have you seen my profile picture?]

@user1 Bellissima!! :-)))
[@user1 So beautiful!! :-)))]

In example (1), the sad emoticon serves both as a mitigation resource, aimed at softening the refusal of an invitation, and as a means of expressing regret for this refusal. In example (2), the smiley is both a marker of intensification of the positive emotion expressed verbally by "so beautiful" and graphically by the exclamation

marks, and a marker of familiarity, aimed at expressing empathy and friendliness.
Emoticons, therefore, are not just a ludic and extralinguistic supplement to language, with the exclusive role of expressing emotions, but rather linguistic resources that play other important pragmatic functions in digital written communication, such as conveying the intentions of the writer (Tagg, 2012), supporting social relationships among participants, and providing new opportunities for creative expressions.

2. Motivation

The main aim of the present investigation was to examine the use of emoticons in Twitter by Italian users. More specifically, it tried to verify, by relying on corpus data and on statistical methodologies, some of the prevailing findings on the use of emoticons in digital written communication: that they are technically-driven resources, whose spread is mainly due to the diffusion of mobile devices (Baron, 2008); that they are mostly used by young people (Merchant, 2001; Tagliamonte & Denis, 2008), and more often by females (Baron, 2008; Huffaker & Calvert, 2005; Spina & Cancila, 2013; Tossel et al., 2012); finally, that they are easy ways of expressing emotions using images instead of words (Provine, Spencer & Mandell, 2007).
Conversely, the nuanced social and pragmatic functions played by emoticons in digital communication suggest that their use and distribution should be affected by a more complex interaction of technological, cultural, situational and sociolinguistic variables.
The questions that this study tried to answer were: what are the variables that, at a discourse level, affect the use of emoticons in Twitter interactions? How is the use of emoticons influenced by these variables?

3. Method

To answer these questions, a large corpus of tweets extracted from the Italian timeline was used. The Ita_twitter corpus (Spina, 2016) contains more than 550,000 tweets sent in a time span of seven months (November 2012-May 2013). The 8,842,450 tokens were pos-tagged through an ad hoc version of TreeTagger (Schmid, 1994), purposely trained to automatically detect emoticons.
From the Ita_twitter corpus, a subset written by 290 users was randomly selected. This subset consists of 4,441 tweets and contains information on the authors (sex, geographical provenance), on their level of mastery within Twitter environment (date of registration on Twitter, number of tweets sent), on the technical context (the software device from which each tweet was sent), and on the type of tweet (a simple status update, or the reply to a previous tweet written by someone else). Information on the authors' age was obtained by manually checking each of the 290 profiles.
The subset of 4,441 tweets contained 15 different types of ASCII emoticons, that are listed in table 1. Each of the types is represented in the corpus by a number of different graphic forms, depending on the combination of ASCII symbols. The classic smiley :-), for example, is represented by a number of different forms, such as :), :)), :-)), etc.

Emoticon	**Meaning**
:-)	smile
;-)	wink
:-(	sad or frown
:*	kiss
<3	heart
-	dazed
:')	tears of happiness
^-^	happy
:P	tongue sticking out
x.x	dead
:'(	crying
-.-	annoyed
:D	laughing
O.O	surprised
u.u	sarcastic

Table 1: the 15 types of emoticons used in the corpus

In addition, given that the corpus was pos-tagged and lemmatized, a range of other linguistic information could be added to the selected tweets, including the type of sentence (question, exclamation, etc.), the co-occurrence of other discourse elements relevant to Twitter interactions (hashtags, mentions), and the length of each tweet (in number of tokens).
In order to explore how Italian participants use emoticons in their Twitter interactions, a mixed-effects model analysis was performed on the selected data. Mixed-effects modeling (e.g., Baayen, Davidson, & Bates, 2008) is particularly suited to corpus data (Gries, 2015), because it can integrate multiple categorical and numeric variables (fixed effects), and, at the same time, it can address the idiosyncrasies deriving from the analysis of data produced by the same subjects (random effects).
The mixed-effect model was built using R version 3.3.3 and the R packages lme4 (version 1.1–13; Bates, Maechler, & Bolker, 2012), lmerTest (version 2.0–33), and sjPlot (version 2.3.1). The number of emoticons used in each tweet (range: 0-27, mean: 0.26, sd 0.68) was used as dependent variable, and the following predictors were initially included in the model: the age (range: 16-67, mean: 31.72, sd 10.57) and sex of participants (f: 1443; m: 2998); the device from which the tweets were sent (mobile or desktop); the level of mastery within the Twitter environment (measured as the number of tweets sent from the date of registration to the date of each tweet in the corpus), distributed in five bands, from the lowest (a) to the highest (e); the type of tweet (status update: 1842, or reply: 2599); the number of co-occurring hashtags (range: 0-8, mean: 0.23; sd 0.63); the tweet length in number of tokens (range: 1-41, mean:

13.85, sd 7.16), and the type of sentence (declarative: 3053, or non-declarative, that is exclamative or interrogative: 1388).
In order to model the individual differences in the use of emoticons, the authors of the tweets were used as random effect, by assuming different random intercepts for each author. In addition, as random slope models allow the predictors to have a different effect for each subject, random slopes were included in the model, with the aim of accounting for the different effects that the "type of tweet" and "sentence" variables have on each single author (Winter, 2013).
Finally, in the model building process, a backward selection approach was adopted, starting with a full model, including all the fixed effects mentioned above, and then dropping one variable at a time, and excluding a variable from the model if non-significant (Gries, 2015).

4. Results

Preliminary results show that four predictors affect the use of emoticons in Twitter as fixed effects (see the plot in figure 1). The predictor with the stronger effect is the type of tweet: emoticons occur far more in replies to the tweets of other participants, which are the most interactive form of tweet (Honeycutt & Herring, 2009; Schnoebelen, 2012), than in status updates. This finding clearly confirms the hypothesis that emoticons are one of the linguistic resources that participants rely on when they need support for establishing or maintaining a social relationship with their interlocutors. Replies, that automatically include a mention to the addressee, are in fact one-to-one or one-to-few interactions, even if they also presuppose the presence of the more numerous audience of followers. In this context, emoticons seem particularly suited to convey reactions to opinions or feelings expressed by others.

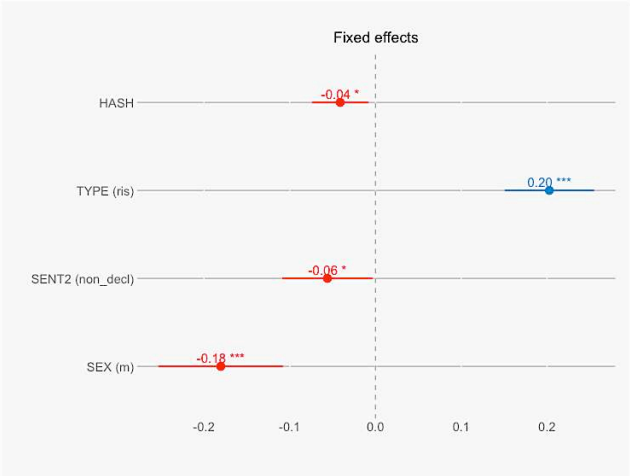


Figure 1: A plot of the estimates of the four significant fixed effects.

Another highly affecting predictor is sex: females use emoticons significantly more than males. This finding confirms previous research in computer-mediated communication (e.g. Vandergriff, 2014; Wolf, 2000; but Danesi, 2016), according to which "females are the prime users of emoticons" (Baron, 2008:65). The same is not true, however, for the age of participants: this variable is never significant in the model as a fixed effect, and cannot therefore be considered a major predictor of the use of emoticons in Twitter.
The different types of sentence used in the tweets are another significant predictor; the model shows that the presence of exclamative or interrogative sentences, signalled by the ? and ! punctuation marks, produces a decrease in the use of emoticons. This finding could seem surprising, since emoticons are a graphical means of expressing emotions, and exclamatives often convey feelings and emotions, but it can be interpreted as an evidence of the multifunctionality of emoticons: one of their roles in digital written communication is that of syntactic markers, often serving as punctuation in place of traditional punctuation marks (Amaghlobeli, 2012). We can conclude, then, that exclamative and interrogative sentences decrease where emoticons replace question and exclamation marks, as in examples (3) and (4):

(3)
@user ciao... Buon lunedì... Che bell'inizio di settimana :-D
[@user Hello… Have a good Monday! What a great start of the week :-D]

(4)
@user Ma che ci fai per DUE MESI a new york :D
[@user What are you going to do for TWO MONTHS in New York :D]

The last significant fixed effect of the model is the hashtag. Emoticons and hashtags tend to have a complementary distribution: when more emoticons are used, less hashtags are found in the tweets. This finding seems coherent with the respective functions of the two discourse elements: while emoticons express either emotions or familiarity, or mark the pragmatic force of a text, the hashtag has an informative function (marking the topic of a text), or addresses the social need of aggregating communities of participants around a common theme or interest (Zappavigna, 2015). In this sense, emoticons seem to serve the pragmatic function of supporting social relationships among few participants, whereas the hashtag plays an important role in affiliating large masses of people in flows of conversations on shared topics.
Going further with the analysis, the picture described so far gets clearer if interactions between different predictors are considered. The type of tweet and the level of mastery, for example, have an interaction effect on the use of emoticons: while replies always contain a greater number of emoticons, this effect seems to slightly increase if the users are more familiar with Twitter. In the case of status updates, the opposite is true: the less proficient the users, the less emoticons they use. In addition, the mastery of Twitter rules and conventions also interacts with the length of tweets in the effect on

the use of emoticons: as shown in Figure 2, low mastery levels (a) produce shorter tweets with less emoticons; medium and high mastery levels (b, c, d and e), conversely, tend to increase the number of emoticons, together with the text length.

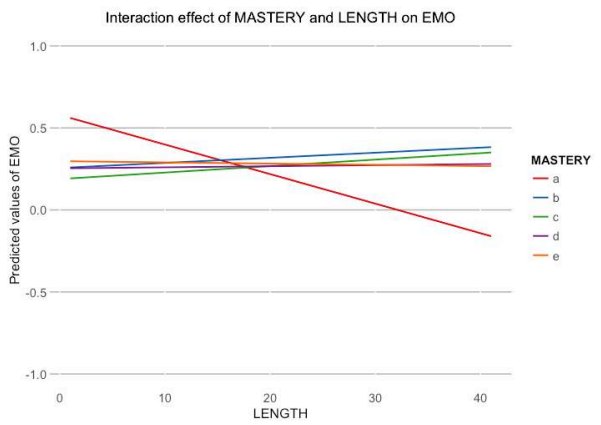


Figure 2: A plot of the interaction between mastery and length of tweets in their effect on the use of emoticons.

From the above, we can conclude that the use of emoticons in Twitter is not technically-driven (the predictor "software device" is never significant in the model, neither as a fixed effect nor in interaction with other factors), and it is not exclusive of young people; rather, it is influenced by a number of pragmatic, cultural and sociolinguistic factors interacting with each other. As a result, far from being only an add-on feature or a frivolous way of expressing emotions, emoticons are constitutive of CMC (Vandergriff, 2014), since they are assuming more and more sophisticated social and pragmatic roles.

After this quantitative investigation, a more in-depth and qualitative analysis needs to be conducted on emoticons, in order to investigate in more detail the linguistic context that favors their use, and their distribution among the previously mentioned functions. A specific attention should be paid to the linguistic features that are traditionally associated to the purpose of establishing and maintaining relationships with other participants (discourse markers and personal pronouns, for example) and of modulating the pragmatic force of texts (intensifiers, affective vocabulary, etc.).

5. References

Amaghlobeli, N. (2012). Linguistic features of typographic emoticons in SMS discourse. *Theory and Practice in Language Studies*, 2(2), pp. 348--354.

Auer, P. (1992). *The contextualization of language*. Amsterdam: Benjamins.

Baayen, R. H., Davidson, D. J. and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, pp. 390--412.

Bates, D. M., Maechler, M. and Bolker, B. (2012). *lme4: Linear mixed-effects models using Eigen and S4*.

Baron, N. (2008). *Always On: Language in an Online and Mobile World*. Oxford: OUP.

Baron, N. (2009). The Myth of Impoverished Signal: Dispelling the Spoken Language Fallacy for Emoticons in Online Communication. In J. Vincent & L. Fortunati (Eds.), *Electronic Emotion: The Mediation of Emotion via Information and Communication Technologies*. Bern: Peter Lang, pp. 107--135.

Baron, N. (2010). *Alphabet to Email: How Written English Evolved and Where It's Heading*. London: Routledge.

Danesi, M. (2016). *The Semiotics of Emoji. The Rise of Visual Language in the Age of the Internet*. London: Bloomsbury

Derks, D.; Bos, A.E.R and von Grumbkow, J. (2007). Emoticons and social interaction on the Internet: the importance of social context. *Computers in Human Behavior*, 23, pp. 842--849.

Dresner, E., Herring, S.C. (2010). Functions of the Nonverbal in CMC: Emoticons and Illocutionary Force. *Communication Theory*, 20(3), pp. 249--268.

Gries, S. Th. (2015). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95--125.

Gumperz, J. (1982). *Discourse strategies*. Cambridge: CUP.

Honeycutt, C., Herring S.C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42),* Los Alamitos: IEEE Press.

Huffaker, D. A., and Calvert, S. L. (2005). Gender, Identity, and Language Use in Teenage Blogs. *Journal of Computer - Mediated Communication*, 10(2).

Lo, S.K. (2008). The Nonverbal Communication Functions of Emoticons in Computer-Mediated Communication. *CyberPsychology & Behavior*, 11(5), pp. 595--597.

Merchant, G. (2001). Teenagers in cyberspace: An investigation of language use and language change in Internet chatrooms. *Journal of Research in Reading*, 24(3), 293--306.

Provine, R.R.; Spencer, R. and Mandell, D. (2007). Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology,* 26(3), pp. 299--307.

Schmid, H, (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Schnoebelen, T.J. (2012). Emotions Are Relational: Positioning and the Use of Affective Linguistic Resources. Ph.D. thesis, Department of Linguistics, Stanford University.

Spina, S. (2016). *Fiumi di parole. Discorso e grammatica delle interazioni scritte in Twitter*. Loreto: StreetLib.

Spina, S., Cancila, J. (2013). Gender issues in the

interactions of Italian politicians on Twitter: Identity, representation and flows of conversation. *International Journal of Cross-cultural Studies and Environmental Communication*, 2(2), pp. 147--157.

Tagg, C. (2012). *Discourse of Text Messaging: Analysis of SMS Communication*. London: Continuum.

Tagliamonte, S., Denis, D. (2008). Linguistic ruin? LOL! Instant Messaging and teen language. *American Speech*, 83(1), pp. 3--34.

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H., Rahmati, A., and Zhong, L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior*, 28(2), pp. 659--663.

Vandergriff, I. (2014). A pragmatic investigation of emoticon use in nonnative/native speaker text chat. Language@ Internet, 11.

Walther, J., D'Addario, K. (2001). The impacts of emoticons on message interpretation in computer mediated communication. *Social Science Computer Review*, 19, 324--347.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf].

Wolf, A. (2000). Emotional Expression Online: Gender Differences in Emoticon Use. *CyberPsychology & Behavior*, 3(5), pp. 827--833.

Yus, F., and Yus, F. (2014). Not all emoticons are created equal. *Linguagem Em (Dis) Curso*, 14(3), pp. 511--529

Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics*, 25(3), pp. 274--291.

Corpus-Based Analysis of Demonyms in Slovene Twitter

Taja Kuzman, Darja Fišer
Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana, Slovenia
E-mail: kuzman.taja@gmail.com, darja.fiser@ff.uni-lj.si

Abstract

This paper reports on a corpus-based analysis of demonym mentions in the corpus of Slovene tweets. First, we analyze the frequency of demonym mentions for the inhabitants of the European and G8 countries. Then, we focus on the representation of demonyms for residents of Slovenia's neighboring countries: Austria, Italy, Hungary and Croatia. The main topic of the tweets mentioning Croatians, Austrians and Italians is sport, whereas Hungarians occur most often in relation to the Eurovision. Some economic and political issues are also represented, such as the selling of Slovene companies to foreign firms, the refugee crisis and the arbitration procedure between Slovenia and Croatia. A collocation analysis revealed a highly stereotypical treatment of the neighboring nations and hostility of some Slovene Twitter users to inhabitants of Slovenia's neighboring countries.

Keywords: demonyms, nationalities, Twitter, discourse analysis, Slovene

1. Introduction

A corpus of user-generated content, especially of tweets, offers an insight into people's beliefs, opinions and attitudes, including attitudes towards residents of other countries. This paper presents an analysis of demonyms (i.e. nouns, used to denote inhabitants of a particular city, country etc.) for the nations which are members of the European Union and of non-European G8 nations that are mentioned in the corpus of Slovene tweets Janes Tweet v4.0 in order to analyze how often Slovene Twitter users talk about other nationalities and in which contexts. Next, a detailed analysis of the representation of the neighboring nationalities was performed in order to establish the general attitude of Slovene Twitter users towards their neighbors.

2. Related Work

Phrases that appear together multiple times provide cultural information and analyzing them can "provide empirical evidence of how the culture is expressed in lexical patterns" (Stubbs, 1996: 169). It is therefore not surprising that many corpus-based discourse analyses have been conducted to observe how people present other nations in written text.

For instance, Bang (2008) examined the representation of foreign countries in the corpus of US news reports. The premodifiers of the keywords 'country', 'countries', 'nation' and 'nations' were analyzed, and collocates indicating verbal and mental actions of Arab and European leaders were examined. Furthermore, the lexical collocates of 'said' and the grammatical collocates of keywords 'China', 'North Korea', 'South Korea' and 'Japan' were analyzed. The study revealed that the representation of foreign counties in US news reports is characterized by stereotyping and asymmetry (ibid.).

Similarly, Tarasheva (2009) used critical discourse analysis to study the representation of Bulgaria in a corpus of articles, published on the BBC website. Articles about Bulgaria were compared to the ones about Belgium, Portugal, Finland and Denmark in a comparable corpora. The research examined the topics of the articles, most frequent keywords and collocations. The results showed that events from Bulgaria are presented differently than those from the other examined countries: articles about crime appear much more often and the most frequent keywords indicate that Bulgarians are mainly portrayed as crime victims. Tarasheva (2009) concludes that "negative coverage for Bulgaria is deliberately sought and achieved".

Our study differs from other corpus-based work mentioned above in that it does not examine texts ordered, authored and edited by professionals but rather unsolicited user-generated content posted by the general public.

In contrast to the abundance of corpus-based studies of representations of countries, representations of inhabitants have not yet received much attention. However, the complex topic regarding the Slovenes' attitudes towards their neighbors has been the subject of many academic works. Throughout history, Slovenes lived in multicultural countries – until the early 20th century in the Austro-Hungarian Empire and then in Yugoslavia until the 1990s (Zupančič & Arbeiter, 2016). Furthermore, during the world wars, they were occupied by Italians, Austrians and Hungarians. Hence, Slovenes began to perceive themselves as inferior to their neighbors. Moreover, they perceived them as their enemies and felt threatened by them (Romih, 2013). Thus, Slovenes have become introverted and developed negative attitudes towards their neighbors in order to feel superior to them as well as to strengthen their nationalistic feelings (Šabec, 2007; Zupančič & Arbeiter, 2016). The growth of negative attitudes has also been influenced by the media in former Yugoslavia which tended to portray other nations as crude and violent (Zupančič & Arbeiter, 2016). Today, Slovenes still distrust their neighbors, especially Croatians, who are perceived to be the least trustworthy peoples from former

Yugoslavia, according to surveys in 2009 and 2010 (Salihović, 2012).

3. The Janes v4.0 Tweet Corpus

The Janes v4.0 Tweet corpus is a subcorpus of Slovene user-generated corpus Janes (Fišer et al., 2016), which contains tweets, written by Slovene Twitter users in the period June 2013–July 2016. The corpus contains 107 million tokens and has been richly linguistically annotated (rediacritization, word-form normalization, part-of-speech tagging and lemmatization) and enriched with metadata, obtained directly from the Twitter API (author, title, time of post, number of retweets and favorites), but also through specialized processes, e.g. sentiment ("neutral", "positive" or "negative"), the gender of the author, the type of the user ("private" for individuals or "corporate" for companies, news agencies etc.) and the linguistic and technical level of (non)standardness of the text.

4. Demonyms in the Slovene Twittosphere

4.1 Subcorpus

The study was performed in the Sketch Engine concordancer. For the purposes of our study, we constructed a subcorpus of tweets, written by individuals (annotated as "private") in the Slovene language. The subcorpus contains 77,250,014 tokens.

Since we were interested in opinions of the general public, we only examined private users' tweets in order to exclude tweets from companies or news outlets that often have a persuasive function, trying to influence the readers' opinion or attract customers.

4.2 Methodology

In the first part of the study, we examined the frequency of demonym mentions for inhabitants of all European nations that were part of European Union in April, 2017 (including Great Britain) and of non-European members of the G8 (Canada, Japan, Russia and the USA). Due to length restrictions of this paper, only official demonyms as they occur in the Slovene orthography manual *Slovenski pravopis* (Toporišič et al., 2014) were analyzed. We examined the occurrence of both masculine and feminine form of demonyms.

4.3 Results

As can be seen from Figure 1, Slovene Twitter users most frequently mention their southern neighbors, Croatians, much more often than inhabitants of other neighboring countries. After Croatians, Slovene tweets most frequently feature residents from the most influential nations of the world—Germany, Russia and the United States of America—which is not surprising as the actions of these countries have a profound influence on the rest of the world. Interestingly, Greeks also occur frequently: regarding a random sample of tweets, we could presume that Slovene Twitter users mostly mention Greeks in connection with the economic crisis in Greece and when commenting their decisions regarding the European Union, as they have an important impact on the economic and political situation in the whole European Union. The least frequent demonyms represent residents of smaller European nations, such as Luxembourg, Cyprus, Malta etc.

Feminine forms of all nationality names are rather rare, which is not surprising as in Slovene the masculine form of the demonym is used as the generic noun that includes both men and women. The only feminine form that stands out is the form for 'Slovene woman' *Slovenka*. It must also be taken into account that when users generalize actions of members of their own nation, they likely substitute 'Slovenes' by 'us'. That could be the reason why the frequency of the demonym 'Slovenes' (*Slovenci*) is lower than frequency of 'Croatians', 'Germans' etc.

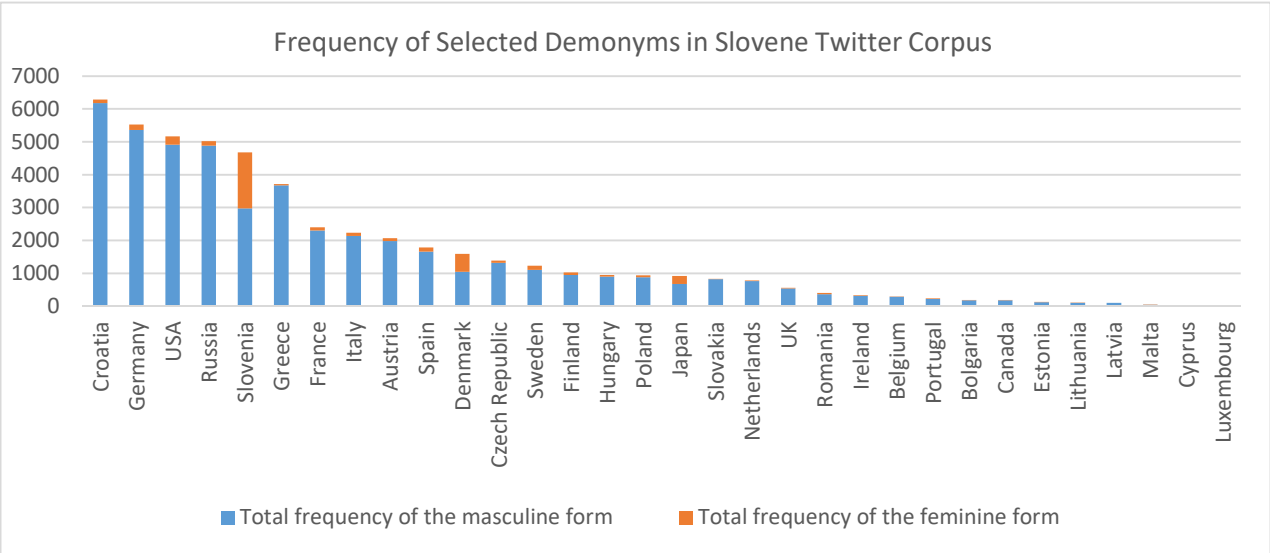


Figure 1: Frequency of Selected Demonyms in Slovene Twitter Corpus

5. Representations of the Neighboring Nations

5.1 Methodology

In the second part of the study, the representations of demonyms for Slovenia's neighboring nations were compared. The keywords *Avstrijec*, *Avstrijka* (masculine and feminine form for 'Austrian'), *Italijan*, *Italijanka* (masculine and feminine form for 'Italian'), *Madžar*, *Madžarka* (masculine and feminine form for 'Hungarian') and *Hrvat*, *Hrvatica* (masculine and feminine form for 'Croatian') were examined in terms of the users that mentioned them (frequency of different users, their gender), the annotated sentiment of the tweets, the number of retweets and favorites and the topics of the tweets. Furthermore, the collocations of these keywords with nouns, adjectives and verbs were analyzed. The aim of this part of the study was to analyze how often Slovenes mention their neighbors and in connection with which topic, whether these tweets receive a lot of attention and what the general attitude of Slovenes towards their neighbors is.

The same subcorpus and concordancer were used as in Section 4. The analysis was conducted using the metadata in the corpus. Collocation analysis was performed with the Word Sketch feature in the Sketch Engine. However, it was limited to masculine forms of the nationality names as the frequency of the feminine forms was too low. All collocations that appeared five times or more were examined. Topics of the tweets in which the relevant nationalities were mentioned were deducted from the accompanying hashtags. Tweets without hashtags were not considered in this final step.

5.2 Results

5.2.1 Metadata

As was already shown in Section 4, Slovene Twitter users most often talk about Croatians: 6,285 tweets were found that contain either masculine or feminine form of this demonym. The second most mentioned neighbors of Slovenes are Italians with 2,231 hits, closely followed by Austrians with 2,068 hits, while Hungarians are mentioned in only 952 tweets. Tweets mentioning Hungarians had the highest frequency of different users, which means that a user rarely wrote about Hungarians more than once; whereas there is the largest number of users who recurrently wrote about Croatians.

A very small amount of all tweets containing demonyms was retweeted (7) or favorited (20) more than 20 times. Tweets mentioning Croatians were retweeted or favorited the most, which is not unusual, given that 'Croatian' is the most frequent demonym mentioned by Slovene Twitter users. Interestingly, the most retweeted (47 times) and favorited (119 times) tweet does not refer to Croatians, but to Austrians. As feminine forms of demonyms rarely appear, it is not surprising that the most retweeted post, containing feminine form of a demonym, was retweeted only 17 times, mentioning Italian women. Approximately three quarters of tweets, featuring masculine forms, were written by men. Surprisingly, feminine forms are much more frequently used by feminine users (*Hrvatica* 'Croatian woman'—39% of authors of tweets were females, *Italijanka* 'Italian woman'—52%, *Avstrijka* 'Austrian woman'—45% and *Madžarka* 'Hungarian woman' —24%).

While the sentiment of tweets containing the masculine form of demonyms are mostly negative, tweets with the feminine form are mostly neutral, except tweets containing *Avstrijka* 'Austrian woman', which are mostly negative. However, in comparison with others, the demonym for Austrian women also has the largest percentage of positive tweets (25%). Positive tweets, which represent 17%–25% of all hits, are the least frequent for all examined nationality names. Interestingly, more than half of the tweets about Hungarians are negative, which makes this nationality the most negatively presented, according to the sentiment annotation.

5.2.2 Collocations

The analysis of demonym + noun collocations revealed not only that analyzed demonyms almost exclusively appear in coordination with other demonyms, mostly with demonyms for residents of Slovenia's neighboring countries. Interestingly, Croatians co-occur with Slovenes much more often than the other three demonyms. Croatians also frequently occurs in coordination with demonyms for inhabitants from the Balkans (Serbians and Bosnians). Italians and Austrians frequently co-occur with Germans. Hungarians appear more often in connection with Italians, Croatians, Austrians, Czechs and Slovaks than with Slovenes.

Due to a low frequency count, no adjective + demonym collocations that pass the frequency threshold (5) were found for *Madžar* 'Hungarian'. On the other hand, *Hrvat* 'Croatian', *Avstrijec* 'Austrian' and *Italijan* 'Italian' collocate with various different adjectives, which indicate how differently they are represented in Slovene Twitter.

'Croatian' collocates with adjectives that are otherwise associated with Slovenes: *podalpski* 'sub-alpine', *alpski* 'alpine' and *brdski* 'from the hills'. These adjectives are used in order to shock readers and to declare that Slovenes are becoming Croatians, or acting as them, as in the tweet "Unfortunately, too many Slovenes are actually Alpine Croatians." (*Žal je preveč Slovencev v resnici Alpskih Hrvatov.*). Furthermore, most adjectives that collocate with 'Croatian' are used ironically. Such adjective is 'poor' (*ubog*) as in "Poor Croatians are left with only 1,000 km of coast…" (*Ubogim hrvatom ostane samo še 1000 km obale ...*) Positive adjectives 'grand' (*veliki*) and 'dear' (*dragi*) are also used ironically. Another adjective that also occurs frequently with this keyword is 'guilty' (*kriv*). It mostly appears in ironic

tweets in which users mock Slovene tendency to blame Croatians for everything, for example "Listening to the news reports, one would say that Croatians are guilty for the unpreparedness of our government." (*Po poročanju medijev bi človek rekel, da so za nepripravljenost naše vlade krivi Hrvati*.) Furthermore, 'Croatian' also collocates with 'true' (*pravi*) in tweets from which stereotypes about Croatians can be easily presumed. Such example is "Refugees are not true Croatians. True Croatians never flee" (*Begunci niso pravi Hrvati. Pravi Hrvat nikoli ne beži*).

In contrast, 'Italian' and 'Austrian' do not appear in collocations with adjectives that are used ironically. 'Italian' mostly collocates with 'loud' (*glasen*), which is generally perceived as a negative trait. That can be seen from the following example: "So a coffee in peace changed into 'a coffee in a coffee shop, filled with loud Italians.' Yay" (*In kava v miru se je spremenila v 'kava v kafiču polnem glasnih italijanov'. Yay*). Interestingly, a collocation 'old Italian' (*star Italijan*) also occurs quite often, mostly with a negative connotation, as in "What, can these old Italians smell in which sauna is a woman. Suddenly, a whole bunch of them is next to her" (*Kva ti stari italijani zavohajo v keri savni je ženska. Naenkrat jih je cel kombi ob njej*).

The keyword 'Austrian' frequently collocates only with one adjective, which is 'rich' (*bogat*). It appears mostly in its superlative form, for instance as in "This year, the richest Austrian is 80 times richer than the richest Slovene" (*Najbogatejši Avstrijec je letos 80-krat bogatejši od najbogatejšega Slovenca*).

As the direct object, the nouns 'Croatian' and 'Italian' frequently appear with the verb 'to defeat' (*premagati*). This collocation appears in connection with sport and Slovene Twitter users mostly hope that their team or foreign teams would beat Croatians or Italians and tweet about it with excitement when it happens. A collocation 'to have a Croatian for neighbor' (*imeti Hrvata za soseda*) also occurs quite often. Generally, there is not enough context to determine whether this is meant in a positive or negative way. However, there are some very telling examples which are clearly negative, such as "Who needs an enemy when you have a Croatian for a neighbor!" (*Kdo rabi sovražnika če imaš Hrvata za soseda!*). Furthermore, the verb 'to hate' also co-occurs relatively frequently with Croatians, but it was discovered that it actually occurs in only one sentence that had been then retweeted by different users: "Who doesn't hate Croatians, ain't Slovene" (*Kdor ne sovraži Hrvatov, ni Sloven'c*—an allusion to a Slovene popular soccer fan slogan "Who doesn't jump, ain't Slovene").

The keywords 'Croatian' and 'Austrian' often collocate with the verb 'to sell' (*prodati*), as Slovene Twitter users mention or disapprove the fact that many Slovene firms have been sold to Croatian and Austrian companies.

The keyword 'Italian' frequently collocates with the verb 'support' (*navijati*), connected with sport. In most analyzed tweets, Twitter users declare (sometimes surprised) that they support the Italian team. Such example is "I see that you support the Italian football team. And I support the Croatians. Who would thought so?!" (*navijaš za italijane u fuzbalu, vidim. Jst za hrvate. Kdu bi si mislil..?!*) In contrast to that, 'Italian' also quite often appears with *ne marati,* meaning 'dislike'. However, these tweets seem much less negative than tweets about Croatians and some have a positive turn, as in "I don't like Italians, but today I supported them" (*ne maram italijanov ampak danes sem za njih navijal*).

5.2.3 Topics

As can be already presumed from the collocations of the keywords with verbs, frequency analysis of hashtags shows that the topic of a majority of the tweets mentioning 'Croatians', 'Austrians' and 'Italians' is sport (e.g. #eurobasket, #sochi…) The only exception are Hungarians, for which the most frequent topic is Eurovision, also a popular topic in tweets with the other three demonyms. In terms of politics, a number of hashtags relate to the arbitration procedure to define border between Slovenia and Croatia, as well as to refugees.

6. Conclusion

In this paper we examined demonym mentions in the corpus of Slovene tweets. The results showed that Slovene Twitter users mostly talk about their southern neighbors, Croatians. According to sentiment annotation, tweets comprising masculine forms of demonyms for Slovenia's neighboring countries are mostly negative, while feminine forms mostly occur in neutral tweets. The collocation analysis revealed that Croatians are generally disliked by the Slovene Twitter users, occurring in ironic or negative context that presents them as unwanted neighbors and reveals deeply rooted stereotypes. Italians are presented as being sometimes unpleasant, but still more likeable than Croatians. When referring to Austrians, Slovene Twitter users mostly connect them with being rich (or richer than Slovenes). Hashtag analysis revealed that Slovenes predominantly mention these nationalities in connection with sport. Some events and political issues are also represented, such as the selling of Slovene companies to Croatian and Austrian firms, the refugee crisis and the arbitration procedure between Slovenia and Croatia.

The analysis was sometimes difficult as some errors in annotation occurred due to polysemy and multilinguality issues (e.g. *Danka* 'Danish woman' or *danka* 'rectum', *Japonka* 'Japanese woman' or *japonka* 'slipper', *Maltežan* 'Maltese man' or *maltežan* 'Maltese dog'). The results are also limited because there were included only official demonyms.

The corpus offers numerous opportunities for extending the research, e.g. the usage of derogatory or discriminatory terms for nationalities, representation of the peoples from the Balkans, as well as comparison between the representation of various nationalities by private and corporate Twitter accounts.

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

Bang, M. (2008). *Representation of Foreign Countries in the US Press: A Corpus Study*. A thesis for the degree of doctor of philosophy. The University of Birmingham.

Fišer, D., Erjavec, T., Ljubešić, N. (2017). The compilation, processing and analysis of the Janes corpus of Slovene user-generated content. *Corpus de communication médiée par les réseaux: construction, structuration, analyse*. Collection Humanités Numériques. Paris: L'Harmattan.

Romih, K. (2013). Humor in slovenska nacionalna identiteta. Bachelor thesis. University of Ljubljana.

Šabec, K. (2007). Conflicting memories and stereotypical images of Italians in the Slovenian collective perception: the case of Slovenian literature from Trieste. *Družboslovne razprave*, 23(55), pp. 95–113.

Salihović, S. (2012). Odnos Slovencev do pripadnikov nekdanje Jugoslavije: diplomska naloga visokošolskega univerzitetnega študijskega programa prve stopnje. Bachelor thesis. School of Advanced Social Studies.

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford, Cambridge: Blackwell.

Tarasheva, E. (2009). The Image of Bulgaria in International Media: A Critical Discourse Analysis and Corpus Study. In *Year Book of the Department of Foreign Languages and Cultures NBU*. http://ebox.nbu.bg/cel/cult07_en.html

Thomas, A. L. (2016). The Agent Across the Border: "Russia" and "Ukraine" as Actors in the News, 2013-2015. In *Theses and Dissertations – Linguistics*. Paper 15. http://uknowledge.uky.edu/ltt_etds/15

Toporišič, J. and Fran Ramovš Institute of the Slovenian Language (2014): *Slovenski pravopis*. Ljubljana: Založba ZRC SAZU. http://bos.zrc-sazu.si/sp2001.html.

Zupančič, R., Arbeiter, J. (2016). Primitive, cruel and blood-thirsty savages: stereotypes in and about the Western Balkans. *Teorija in praksa*, 53(5), pp. 1051–1063.

European Language Ecology and Bilingualism with English on Twitter

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

Abstract

Societal and demographic changes have contributed to increasing bi- and multilingualism in European countries in recent years, and communication on social media platforms such as Twitter reflects this linguistic diversity. While high rates of English use online have been attested for many European countries by survey research, relatively little work has quantified the extent to which English is used on social media in European contexts. In this study, English use and bilingualism with English in Europe are investigated on Twitter. A large corpus of Twitter messages with geographical metadata was created by accessing the Twitter APIs. After language detection and filtering, linguistic profiles for European countries were created and the behavior of bi- and multilingual users examined. The analysis supports some previous findings that suggest that a large-scale language shift towards English may be ongoing in Europe in some communicative domains. Geographical differences shed light on the dynamics of this process.

Keywords: Bilingualism, social media, Twitter, corpus linguistics, quantitative methods

1. Introduction and Background

Recent years have seen an increase in the relative prominence of computer-mediated communication (CMC) modalities such as texting, instant messaging, or posting on social media, and platforms such as Twitter have become multilingual sites with global representation (Mocanu et al. 2013; Leetaru et al. 2013). At the same time, population movements and changes in education and media consumption have contributed to an increasing bi- and multilingualization of local environments, particularly with English – trends that are particularly evident in online communicative domains in some European societies. Although national languages continue to receive reinforcement in education and media, bilingualism with English has become the norm for many within Europe, particularly for young people.

In this study, bi- and multilingualism with English are investigated by means of a quantitative analysis of Twitter messages with location metadata in order to establish a language ecology (Haugen 1972). The research poses the following questions: Which languages are favored by multilinguals on Twitter in Europe? How linguistically diverse are European societies on the platform, and what role does English play? And to what extent do national languages play a role in the discourse of European Twitter users? Addressing these questions may allow us to characterize European Twitter discourse in terms of a language ecology that can "tell us something about where [a] language stands and where it is going in comparison with other languages of the world" (Haugen 1972, p. 337).

In a first step, the linguistic behavior on Twitter of users who can be reliably located within European countries is examined according to country in order to provide an overview of the language ecoloy of Europe. In a second step, the aggregate network behavior of bi- and multilingual users is examined more closely: Which languages do multilinguals favor in which places? The structure of the network of multilinguals between languages can shed light on the relative status of English and national languages and, due to the prevailing demographics of Twitter users, perhaps provide an indication of middle- to long-term language shift for European societies.

2. Previous Work: Twitter Language and Multilingualism

A number of studies of CMC and Twitter language have investigated aspects of English, including phenomena such as the discourse functions of hashtags (Wikström 2014; Squires 2015), lexical innovation in American English (Eisenstein et al. 2014), African-American Vernacular English dialect on Twitter (Jørgensen, Hovy, and Søgaard 2015), grammatical variation in English-language Twitter from Finland and the Nordic countries (Coats 2016a; Coats 2016b), or the interaction between demographic parameters such as gender with lexical and grammatical features in American English (Bamann, Eisenstein, and Schnoebelen 2014).

Ronen et al. (2014) found that English plays an important central role in multilingual networks of Wikipedia editors, book translations, and Twitter users. Hale (2014) investigated global multilingual networks on Twitter, including the network associations of retweets and user mentions, and found that while most interaction networks are language-based and English is the most important single mediating language, other languages collectively represent a larger bridging force. Eleta and Golbeck (2014) examined the tweets of 92 multilingual Twitter users and showed that their language choice on the Twitter reflects the predominant language of their social networks. Kim et al. (2014) used Shannon Entropy to quantify linguistic diversity on Twitter in Switzerland, Quebec and Qatar. They created networks of mono-, bi- and multilinguals, and demonstrated that while English mediates between language communities, users of local languages have more influence. Topic selection may also influence language choice. Such findings have confirmed the status of English as the global *lingua franca*, but the dynamics of multilingualism in a large social media data

set from all of Europe has to our knowledge not yet been subject to research attention.

Other studies have used surveys to investigate online exposure to and use of languages, their relative status in various media or communicative contexts, and attitudes towards them in Europe (e.g. the *Eurobarometer* surveys conducted by the European Commission or Leppänen et al. 2011 for Finland). Increasing knowledge of English has cemented the language's "hypercentral" position within the language ecology of Europe (Swaan 2001; Soler-Carbonell 2016), and there may be evidence that English has now displaced some local languages in certain functional domains in some European societies (Görlach 2002, p. 16; for a discussion see the contributions in Linn 2016).

Few large-scale studies of aggregate online language use in Europe, however, have been based on documented usage, and empirical research into aggregate use on Twitter has typically offered only an overview of language frequencies. Additionally, while language-use profiles at country level for Twitter data exist (e.g., Mocanu et al. 2013; Leetaru et al. 2013; Magdy et al. 2014; Graham, Hale, and Gaffney 2014), relatively few studies focus specifically on bi- or multilingualism.

3. Methods

Corpus-based and NLP methods were employed in the study. They comprised the collection of data online, filtering of data, quantification of multilingualism, and the construction and visualization of language networks.

3.1. Data Collection

Over 140 million tweets with `place` attributes from European countries or territories were collected from the Twitter Streaming API from November 2016 until June 2017 using the *Tweepy* library in Python (Roesslein 2015). From this "seed" dataset of tweets by 2.9 million users, the tweets of those with at least 20 tweets and at least 50% of tweets from a single country (654,676 users) were retained for analysis.[1] In total, the data used for analysis comprised over 69.8 million tweets from 55 European countries or territories.

3.2. Data Filtering and Language Detection

Not all tweet user messages are composed by humans: A substantial proportion of tweets is generated automatically by apps or bots that interact with the Twitter API (Haustein et al. 2016). Because many apps post content that is not user-composed but rather consists of automatically-generated text, filtering tweets by the `source` value can reduce the amount of noise in the data set. A manual analysis of a selection of tweets showed that widely-used Twitter apps such as "Twitter for iPhone" or the Twitter Web Client (i.e. www.twitter.com) were less likely to broadcast automatically-generated text than were some infrequently-used apps. For this reason, the data was filtered to retain only those tweets broadcast by the following apps: Twitter Web Client, Twitter for iOS, Twitter for iPhone, Twitter for Android, Twitter for Windows Phone, Twitter for Instagram, Tweetbot for iOS, and Tweetbot for iPhone. Tweets with these sources collectively comprised over 87% of all those by European users.

A consideration of bi- and multilingualism on the Twitter platform critically depends on accurate characterization of the language of individual tweets, but automatic language detection of tweets can pose difficulties. Character sequences present in URL addresses, usernames, hashtags, emojis, and non-standard orthography can create problems for automatic language detection algorithms, as they rarely correspond to items in the lexicons of natural languages. Even after removing such sequences, very short texts are not handled well by language detection algorithms (Figure 1). To increase detected language accuracy, the data was therefore filtered to include tweets that exhibited three-way agreement between the native Twitter language detection algorithm and the algorithms `langid` (Lui and Baldwin 2014) and `compact language detector 2` (Sites 2014) after removal of URLs, usernames, hashtags, and emojis. For some less-widely-used languages not identified by all three algorithms, such as Faroese, Nynorsk, Albanian, or Somali (among others), two-way language identification or identification by a single algorithm with a high probabilistic accuracy value was used to assign languages to tweets.[2]

	text	langlangid
12260	normalee, ili dzeperice :)	it
12272	luicrede divincere mann saràcosì! referendum trascu...	it
12397	gol dla 2:1!	it
12459	koje crno malo :))	it
12736	no hej	it
154	ssshes so besyitful	de
251	popieram:-)	de
344	gesehen? fantastische serie.	de
518	i am bored	de
773	a legend	de

Figure 1: Language misidentification on short texts by `langid`

3.3. Quantification of Bilingualism Strength

A user in the dataset was determined to be bilingual for languages i, j if he or she had authored at least 10% of the total number of tweets in each of the two languages. The connection strength between languages i, j was quantified on the basis of all users with the phi coefficient, calculated from a contingency table (Table 1 and Equation 1).

[1] For this data, correlation between the center of the `place` bounding box and the precise GPS coordinates from the `coordinates` object, if both were present, was found to be quite high (= 0.992). For this reason, the `place` field was considered an accurate indication of true user location when posting a tweet.

[2] The presence of unique vocabulary markers (Ljubešić, Fišer, and Erjavec 2014) can be used to collect tweets in less-used languages, but the method is not applicable to the detection of already-collected tweets.

Table 1: Contingency Table for Number of Bilinguals

	$language_i$	$\sim language_i$	
$language_j$	O_{11}	O_{12}	$= R_1$
$\sim language_j$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

$$\phi_{ij} = \frac{(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{R_1 R_2 C_1 C_2}} \qquad (1)$$

ϕ is equivalent to Pearson's product-moment correlation coefficient for occurrence frequence of two binary variables, and ranges in value from -1 to 1. Positive values indicate the language pairs are more strongly connected than would be expected based on the prevalence of the languages in the multilingual dataset.

A t-statistic was calculated to test the significance of the correlation between languages according to the formula

$$t_{ij} = \frac{\phi_{ij}\sqrt{D-2}}{\sqrt{1-\phi_{ij}^2}} \qquad (2)$$

where $D = max(R_1, C_1)$.

A multilingualism network for Europe was created in which node size corresponds to the number of multilingual users for a language and edge width corresponds to the strength of the connection (number of bilinguals) for a language pair. The network for those nodes and edges with at least 10 bilinguals and t-test p-values < 0.05 was retained (Figure 2). Network relationships were visualized using the R packages *igraph* and *visNetwork* (Csardi and Nepusz 2006; Thieurmel 2016). Additionally, nodes and vertices were annotated with information about total number of users, total number of bilinguals, and average message length.

4. Results

In terms of overall language representation in the European Twitter bi- and multilingualism network, English is the most prevalent language, with approximately 30% of all tweets in English. For Europe as a whole, a network of 42 languages and 68 edges describes the statistically significant bilingual links (Figure 2). English clearly plays the most important role: it is connected to almost all of the languages in the network. Other languages with large numbers of users, such as French, Spanish, Turkish, and Russian, have multiple connctions to other languages.

It should be remarked that the bi- and multilingual network only accounts for productive language use (i.e. authorship of discourse in a particular language), not passive understanding of languages. In an additional step, follower and friend statistics will be used to estimate these values as well.

Multilingual networks were also created for individual European countries. In them, English serves as as a bridge between linguistic communities, but the principal national

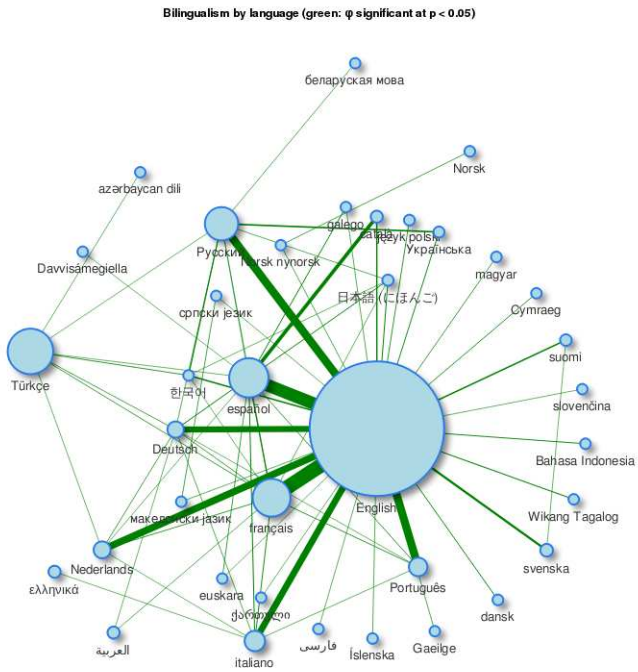


Figure 2: Bilingualism Network

language(s) figure prominently. Nonetheless, high levels of bilingualism with English, when considered in light of previously reported demographic statistics (European Commission 2012), may indicate a shift towards English and away from traditional languages in Europe, particularly for some minority languages with official status.

References

Bamann, David, Jacob Eisenstein, and Tyler Schnoebelen (2014). "Gender Identity and Lexical Variation in Social Media". In: *Journal of Sociolinguistics* 18.2, pp. 135–160.

Coats, Steven (2016a). "Grammatical feature frequencies of English on Twitter in Finland". In: *English in Computer-mediated Communication: Variation, Representation, and Change*. Ed. by Lauren Squires. Berlin: De Gruyter, pp. 179–210.

– (2016b). "Grammatical frequencies and gender in Nordic Twitter Englishes". In: *Proceedings of the 4th conference on CMC and social media corpora for the humanities*. Ljubljana, pp. 12–16. URL: http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-conference-proceedings-2016.pdf.

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems, p. 1695. URL: http://igraph.org.

Eisenstein, Jacob et al. (2014). "Diffusion of Lexical Change in Social Media". In: *PLoS ONE* 9.1.

Eleta, Irene and Jennifer Golbeck (2014). "Multilingual use of Twitter: Social networks at the language frontier". In: *Computers in Human Behavior* 41, pp. 424–432.

European Commission (2006). "Europeans and their languages: Special Eurobarometer 243". In: URL: http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.

– (2011). "User language preference online: Flash Eurobarometer 313". In: URL: http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

– (2012). "Europeans and their languages: Special Eurobarometer 386". In: URL: http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_sum_en.pdf.

Graham, Mark, Scott A. Hale, and Devin Gaffney (2014). "Where in the World Are You? Geolocation and Language Identification in Twitter". In: *The Professional Geographer* 66.4, pp. 568–578. URL: http://dx.doi.org/10.1080/00330124.2014.907699.

Görlach, Manfred (2002). *Still More Englishes*. Amsterdam: John Benjamins.

Grosjean, François (2008). "Studying bilinguals: Methodological and conceptual issues". In: *Handbook of bilingualism*. Ed. by Tej K. Bhatia and William C. Ritchie. Malden, MA: Wiley-Blackwell, pp. 32–63.

Hale, Scott (2014). "Global connectivity and multilinguals in the Twitter network". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. New York, pp. 833–842.

Haugen, Einar (1972). "The ecology of language". In: *The ecology of language*. Ed. by Einar Haugen and Anwar Dil. Palo Alto: Stanford University Press, pp. 325–339.

Haustein, Stefanie et al. (2016). "Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter". In: *Journal of the Association for Information Science and Technology* 67.1, pp. 232–238. ISSN: 2330-1643. DOI: 10.1002/asi.23456. URL: http://dx.doi.org/10.1002/asi.23456.

Jørgensen, Anna Katrine, Dirk Hovy, and Anders Søgaard (2015). "Challenges of studying and processing dialects in social media". In: *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Association for Computational Linguistics. Stroudsburg, PA, pp. 9–18. URL: http://aclweb.org/anthology/W15-4302.

Kim, Suin et al. (2014). "Sociolinguistic Analysis of Twitter in Multilingual Societies". In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. HT '14. Santiago, Chile: ACM, pp. 243–248. URL: http://doi.acm.org/10.1145/2631775.2631824.

Leetaru, Kalev H. et al. (2013). "Mapping the global Twitter heartbeat: The geography of Twitter". In: *First Monday* 18.5/6.

Leppänen, Sirpa et al. (2011). *National Survey on the English Language in Finland: Uses, meanings and attitudes (= Studies in Variation, Contacts and Change in English, Volume 5)*. Helsinki: Varieng.

Linn, Andrew, ed. (2016). *Investigating English in Europe: Contexts and agendas*. Berlin and Boston: De Gruyter Mouton.

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec (2014). "TweetCaT: a tool for building Twitter corpora of smaller languages". In: *LREC*.

Lui, Marco and Timothy Baldwin (2012). "Langid.py: An off-the-shelf language identification tool". In: *50th Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics. Stroudsburg, PA, pp. 25–30.

– (2014). "Accurate language identification of Twitter messages". In: *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) EACL 2014*. Association for Computational Linguistics. Stroudsburg, PA, pp. 17–25.

Magdy, Amr et al. (2014). "Exploiting geo-tagged tweets to understand localized language diversity". In: *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*. GeoRich'14. Snowbird, UT, USA: ACM, 2:1–6. URL: http://doi.acm.org/10.1145/2619112.2619114.

Mocanu, Delia et al. (2013). "The Twitter of babel: Mapping world languages through microblogging platforms". In: *PLoS ONE* 8.4.

Roesslein, Joshua (2015). *Tweepy*. Python programming language module. URL: https://github.com/tweepy/tweepy.

Ronen, Shahar et al. (2014). "Links that speak: The global language network and its association with global fame". In: *PNAS* 111.52, E5616–E5622.

Sites, Dick (2014). *Compact language detector 2*. R package version 1.0.2. URL: https://github.com/CLD2Owners/cld2.

Soler-Carbonell, Josep (2016). "English in the language ecology of Europe". In: *Investigating English in Europe: Contexts and agendas*. Ed. by Andrew Linn. Berlin and Boston: De Gruyter Mouton, pp. 53–58.

Squires, Lauren (2015). "Twitter: Design, discourse, and implications of public text". In: *The Routledge Handbook of Language and Digital Communication*. Ed. by Alexandra Georgakopoulou and Tereza Spilioti. London and New York: Routledge, pp. 239–256.

Swaan, Abram De (2001). *Words of the world: The global language system*. Cambridge: Polity.

Thieurmel, Benoit (2016). *visNetwork: Network Visualization using 'vis.js' Library*. R package version 1.0.2. URL: https://CRAN.R-project.org/package=visNetwork.

Wikström, Peter (2014). "#srynotfunny: Communicative functions of hashtags on Twitter". In: *SKY Journal of Linguistics* 27, pp. 127–152.

Reliable Part-of-Speech Tagging of Low-frequency Phenomena in the Social Media Domain

Tobias Horsmann, Michael Beißwenger and Torsten Zesch
Language Technology Lab
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen, Germany
{tobias.horsmann,michael.beisswenger,torsten.zesch}@uni-due.de

Abstract

We present a series of experiments to fit a part-of-speech (PoS) tagger towards tagging extremely infrequent PoS tags of which we only have a limited amount of training data. The objective is to implement a tagger that tags this phenomenon with a high degree of correctness in order to be able to use it as a corpus query tool on plain text corpora, so that new instances of this phenomenon can be easily found. We focused on avoiding manual annotation as much as possible and experimented with altering the frequency weight of the PoS tag of interest in the small training data set we have. This approach was compared to adding machine tagged training data in which only the phenomenon of interest is manually corrected. We find that adding more training data is unavoidable but machine tagging data and hand correcting the tag of interest suffices. Furthermore, the choice of the tagger plays an important role as some taggers are equipped to deal with rare phenomena more adequately than others. The best trade off between precision and recall of the phenomenon of interest was achieved by a separation of the tagging into two steps An evaluation of this phenomenon-fitted tagger on social media plain-text confirmed that the tagger serves as a useful corpus query tool that retrieves instances of the phenomenon including many unseen ones.

Keywords: Part-of-speech, Social Media, CMC, Rare Phenomena

1. Introduction

This paper reports on experiments on adapting part-of-speech (PoS) taggers for tagging rare phenomena found in genres of computer-mediated communication (CMC). Our work is motivated by a use case in which a linguist wants to study a rarely occurring CMC phenomenon using Twitter data from the social media domain. The central problem here is how to find such rare instances of the phenomenon under observation without spending hours of screening through plain text. A filtering tool would be desirable that facilitates the retrieval process for the linguist. The tool should find many instances of the phenomenon and at the same time achieve reasonably correct results in order to decrease the workload considerably. The project we present here investigates how to adapt a PoS tagger for tagging a certain rarely occurring phenomenon in order to use the PoS tagger as a filtering tool that linguists can use to query a corpus.

The main challenge of adapting a PoS tagger to the language use in the social media domain lies in dealing with the notorious lack of training data and many out-of-vocabulary words. This problem becomes even more severe when the tagger shall be adapted for dealing with a phenomenon that is under-represented in the already small training data sets. We will, thus, investigate methods to improve tagging of under-represented phenomena while laying emphasis on avoiding manual annotation as much as possible. We aim on detecting a German verb-pronoun contraction phenomenon that the linguist wishes to study in detail on the basis of a broad set of instances automatically retrieved from social media data. To deal with the lack of training data, we experiment with (i) adjusting the frequency weight of the under represented phenomena by under- and oversampling and (ii) adding automatically tagged but new data in which only the tag of the phenomenon of interest is manually corrected. In a concluding case study, we optimize a tagger towards finding this contraction phenomenon and evaluate how well the filtering works in a real world setup on plain text Twitter messages.

wiederholen (to repeat) + es (it)	1st person
ich **wiederhols** nochmal, ihr redet hier öffentlich!	
I repeat it [repeat-it] again, you're talking in public!	
kommen (to come) + du (you)	2nd person
wieso? wo **kommste** denn her?	
why? where do you come [come-you] from?	

Table 1: Full verb + pers. pronoun (VVPPER) contraction

2. German Verb-Pronoun Contraction

We are interested in a particular phenomenon in which a verb and a following personal pronoun are contracted into a single form. Table 1 shows examples of this type of contractions taken from the Dortmund Chat Corpus (Beißwenger, 2013). Verb-pronoun contractions belong to the class of phenomena that are not unique for CMC discourse but typical for spontaneous - spoken or 'conceptually oral' - language in colloquial registers. Phenomena of this type are of special interest for linguists who want to use corpora to compare written discourse from the social media domain to the language of edited text and the language found in informal, spoken interactions. If we use a tagger as a filtering tool, we need a high precision to avoid screening through countless false positive instances. At the same time, we want to find new lexical forms unknown from the training set, which requires a high recall (i.e., high generalization).

We have a data set of 23k tokens of German social media discourse that was annotated for a shared task on PoS tagging for German CMC and social me-

dia data (Beißwenger et al., 2016). The data are annotated with an extended version of the Stuttgart-Tübingen tagset (STTS) (Schiller et al., 1999) that has been expanded by tags needed for tagging social media phenomena (Beißwenger et al., 2015). In this tagset, verb-pronoun contractions are labelled by an own tag, VVPPER, which occurs 13 times in total. Results of the shared task showed that this infrequency prevents taggers from learning the phenomenon in a reliable manner (Horsmann and Zesch, 2016b). Since the VVPPER tag is not included in the canonical STTS, as those contractions do not occur in the domain of edited text, existing STTS annotated corpora (e.g., newspaper corpora) cannot be used to obtain additional instances of the phenomenon for training. Although there is a small amount of annotated data that we can build upon, there are still not enough VVPPER instances to make the phenomenon recognizable when training PoS taggers.

3. Dealing with Infrequency

In this experiment, we test different strategies to improve the tagging of VVPPER instances. With a total of 13 instances in our data set, we have to annotate at least some additional data in order to train the tagger but also to arrive at meaningful results during evaluation. At the same time we keep the manual annotation effort at a minimum.

3.1. Data Set

We base our experiment on the data set from the aforementioned shared task. We enrich the data by selecting 230 user posts that contain this phenomenon from the Dortmund Chat Corpus. We automatically tagged these additional data by using the Stanford tagger that assigns PoS tags of the canonical STTS and manually corrected the tag for the verb-pronoun contraction that only exists in the extended STTS. This is the most minimalistic amount of manual annotation one can possibly perform which - as we will see soon - suffices. Of the additional 230 instances, we add one half to the testing set and one-sixth to the training set. The remaining two-sixths are our development set in the following experiments and are held back for the moment. The enhanced training set now contains 45 (38+7) sequences with the phenomenon and the testing set 121 (115+6) sequences. This should be enough instances for learning and evaluating the phenomenon.

3.2. Frequency Weight vs. Lexical Knowledge

An option to circumvent annotation of a larger amount of data is boosting the signal for a certain PoS tag in the already existing data. This can either be done by oversampling (Daumé III, 2007) the few instances one has by adding them N times to the training set, or by downsampling, i.e. removing sequences *without* the PoS tag of interest i.e. VVPPER. Both approaches lead to an increased frequency weight of the phenomenon relative to the other PoS tags in the corpus. We experiment with both strategies in the following setup: *Downsampling:* We remove 25, 50 and 75 percent of the training data instances that do not contain any verb contractions. *Oversampling/new Instances:* We choose oversampling rates that add a number of instances which we can also provide from the held back annotated sequences. This allows a direct comparison between oversampling instances and adding fresh ones. We will, thus, oversample two and three times, and compare this to adding the same amount of instances from the set of new sequences in the held back development set.

We conduct these experiments with the following taggers to learn about the empirical differences between tagger implementations for our objective:

Stanford (Toutanova et al., 2003) a PoS tagger that is frequently used in the community due to its good reputation and high accuracy.

HunPos (Halácsy et al., 2007), a tagger with a good reputation based on Hidden-Markov models and a re-implementation of the TNT tagger (Brants, 2000).

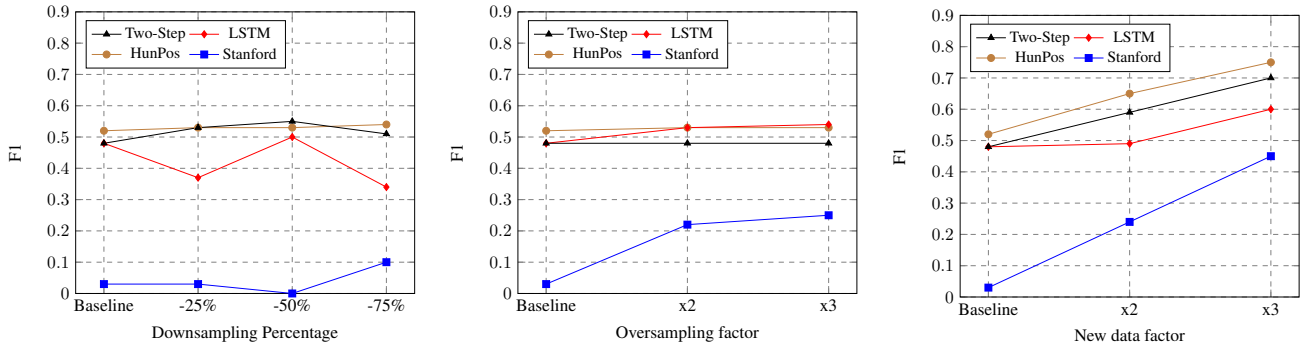
LSTM A deep learning PoS tagger by Plank et al. (2016), that is based on Long-Short-Term-Memory (Hochreiter and Schmidhuber, 1997) neural networks. We use the same parametrization as Plank et al. (2016) and self-trained German word embeddings trained on German Twitter messages with $195 \cdot 10^6$ tokens.

Two-Step Horsmann and Zesch (2016a) proposed a tagger architecture for social media data that first uses a highly generalized *coarse-grained* tagger, and as a second step applies a specialized non-sequential tagger for *fine-grained* tagging. The second tagger is tailored towards recognizing the tag of interest while the first tagging step constraints the application of the second tagger.

We implement this approach by using a CRF tagger (Lafferty et al., 2001) in the first step and an SVM in the second step. For training the coarse-grained sequence model, we map the extended-STTS tags of the training data to the coarse-grained tagset used by the Universal Dependency project and map VVPPER to *verb*. We include a PoS dictionary and Brown (Brown et al., 1992) clusters created over German Twitter messages to compensate for the lack of training data. This coarse-grained tagger reaches a F_1 of 0.93 on the tag *Verb* in the test data, which means that some VVPPER instances will be missed because the coarse model did not predict *verb*.

Results In Figure 1, we show the results of the three strategies on the VVPPER tag. We focus on *out-of-vocabulary* instances which perform considerably poorer than *in-vocabulary* instances (F_1 between 0.96 to 0.99), and thus offer more opportunities for improvements. We see that neither downsampling nor oversampling helps to reach a substantial improvement on the tag. Furthermore, downsampling shows that the anyway low amount of training data becomes a large problem for the LSTM if further reduced. The Stanford tagger stays behind the other taggers with both sampling methods. The only effective method is, without much surprise, providing new data. The LSTM needs considerably more data to improve while the other taggers improve linearly with each new data set.

Discussion Table 2 shows details of the two best taggers HunPoS and Two-Step. Once again, we focus on the out-of-vocabulary instances, while also showing precision (P) and recall (R). The F_1 score shows that both taggers reach



(a) Downsampling (b) Oversampling (c) Annotated Data

Figure 1: Results on *unknown* VVPPER word forms with various methods

	Setup	All F1	Out-Vocabulary P	R	F1
HunPos	Baseline	.78	.80	.38	.52
	Downs. 75%	.78	.63	.48	.54
	Downs. 50%	.79	.74	.41	.53
	Downs. 25%	.79	.81	.40	.53
	Overs. x2	.79	.78	.40	.53
	Overs. x3	.79	.74	.41	.53
	Annotated x2	.83	.80	.56	.65
	Annotated x3	**.88**	.81	.70	.75
Two-Step	Baseline	.77	.95	.32	.48
	Downs. 75%	.78	.85	.37	.51
	Downs. 50%	.80	.96	.38	.55
	Downs. 25%	.79	.92	.38	.53
	Overs. x2	.77	.95	.32	.48
	Overs. x3	.77	.95	.32	.48
	Annotated x2	.81	.93	.43	.59
	Annotated x3	**.85**	.92	.56	.69

Table 2: F_1 on all and on out-of-vocabulary instances

a rather similar overall performance. When looking at precision and recall for adding annotated data, highlighted in grey, we see that Two-Step is considerably more precise than HunPos, which has a better recall. Because oversampling showed barely any effect, we suspect that the added lexical knowledge is mostly accountable for the improvements, which also means that the word context seems to be neglected for making decisions. If the tagger focuses too much on lexical forms, it will find mostly instances known from training which is in particular a problem for finding new instances. Hence, an increased weighting of the local word context should support finding new instances and enable a better generalization.

3.3. Experiment: Forced Generalization

In this experiment, we try to improve generalization of the Two-Step tagger by forcing the tagger to rely more on the local word context and, thus, improve the recall. We chose Two-Step, as we have implemented this tagger ourselves which facilitates adaptation. We alter the feature space of the SVM and exclude all features that contain the lexical form of the *positive* instances. Thus, the SVM is not aware of any lexical forms that can occur with the tag VVPPER,

Configuration	All F_1	Out-of-Vocabulary P	R	F_1
Baseline	.81 (+.04)	.93 (+.02)	.41 (+.09)	.57 (+.09)
Annotated x3	.86 (+.01)	.89 (−.03)	.62 (+.06)	.73 (+.04)

Table 3: Results of the contextualised Two-Step

and must now rely more on the word context.

Results In Table 3, we show the changes in performance of the contextualised Two-Step tagger. In parentheses, we show the differences to the not contextualized tagger in Table 2. For both setups we see an improved F_1, but especially the recall increases for out-of-vocabulary instances. The overall F_1 reached by HunPos (.88) in Table 2 is still superior but the trade off between precision and recall of Two-Step better supports the use case in which the tagger functions as a precise filtering tool with decent recall.

4. Field Trial in Social Media

So far, we have only simulated our use case of a linguist who uses a tagger as a filtering tool, while now, we turn to a real setting and apply a tagger to plain text Twitter messages for finding verb-pronoun contractions.

Working on plain text means that the ground truth of how many instances there are in the data is unknown, thus, the recall cannot be computed. Consequently, we focus on evaluating the precision of the tagging, and evaluate how many new instances are found. We choose the Twitter domain for its ease of obtaining data but also for its linguistic diversity that ranges from tweets using informal, interactional language to tweets that are close to the written standard. This domain provides us with a challenging test bed that should allow to determine a conservative, lower-bound performance for our approach. We will use the contextualized Two-Step tagger for its higher precision while providing a reasonable high recall.

Twitter Data We use a random subsample of 50k tweets (about 1.7 million tokens) crawled between 2011 and 2017 from the public Twitter API that we language-filtered for German. All occurrences of user-mentions, hashtags and URLs are replaced by a text constant and the tweets are tokenized by Gimpel et al. (2011)'s ArkTools tokenizer.

Strict
Da **lernste** pragmatisch zu sein .
Ich **sachs** dir noch .
Relaxed
Wer **häts** gedacht .
Ich **wills** nicht ich will aber auch nicht [...]
All
Warum einfach , **wenn's** auch kompliziert geht ? URL
Ich beschränke mich **auf's** nicht im Weg stehen .
Frequent Confusion Cases
Und keiner **weiss** warum .
Ich **weiss** gar nicht , was du beruflich machst .

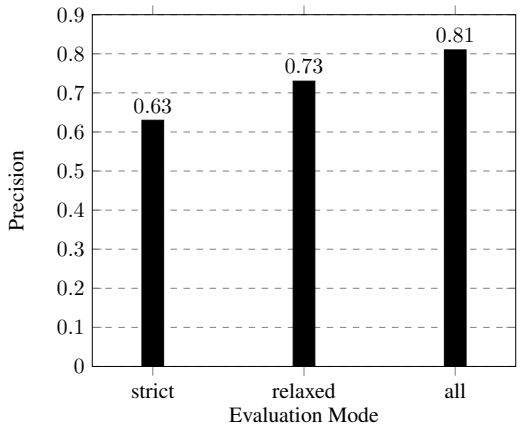
Table 4: Examples of tagged instances

Tagger Setup We train the coarse model and the SVM on the full shared task data set including the additionally annotated data. To provide more lexical knowledge and increase the robustness when facing standard language text, we also add 100k tokens of the German newswire Tiger (Brants et al., 2004) corpus to both tagging steps.

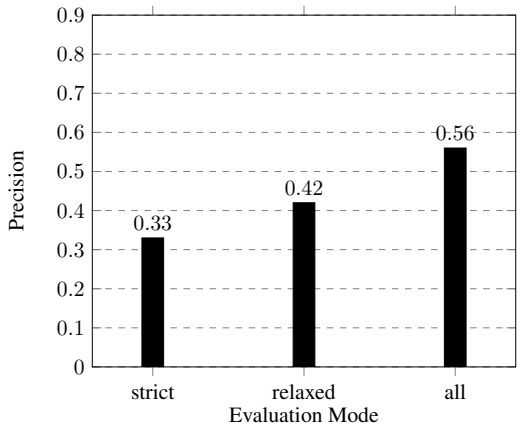
Evaluation setup We evaluate the tagged instances with two annotators. The annotators make four distinctions: *strict*, *relaxed*, *all* and *none*. *Strict* are full verb contractions with personal pronoun, the exact phenomenon we intended to tag. *Relaxed* counts all verb contractions with personal pronoun as correct, this includes also modal and auxiliary verbs. *All* counts all contractions phenomena as correct, this additionally includes, for instance, contractions of conjunctions with personal pronouns. The remaining cases are no contractions and are, thus, false positives.
We will evaluate two setups. The first one selects the first 250 of all found instances, which will be the overall evaluation. The second evaluation focuses on out-of-vocabulary instances in which we remove all tagged instances that are known from the training set until we gather 250 instance and, thus, evaluate how reliably new instances are found.

Results In total, we found 1091 instances in 50k tweets tagged as *VVPPER*. The two annotators reached a perfect agreement on the subset of the first 250 instances that we evaluated manually. Figure 2a shows the precision of the overall evaluation. The *strict* result shows that the majority of found instances are the targeted full verb contractions. Including modal and auxiliary verbs in the *relaxed* mode, even three-quarter are verb contractions. Including also miscellaneous contractions in *all*, almost all instances are contractions.
In Figure 2b, we take a closer look on the performance of detecting new contractions, e.g. out-of-vocabulary instances. We focus our discussion on the *strict* results. The precision is drastically decreased to almost half the value that we reach when including all instances. We also computed the type/token ratio which is at 0.69 almost twice as high as in the overall evaluation in Figure 2a. This confirms that the tagger is able to recognize many new instances of the phenomenon. Furthermore, when ignoring the known instances almost every correct instance is a new lexical form.

Discussion Table 4 depicts examples of each of the three contraction classes (bold face) and additionally presents a



(a) In- and out-of-vocabulary contractions



(b) Out-of-vocabulary contractions

Figure 2: Results of manual evaluation

frequent confusion case, which is erroneously tagged as contraction. Of the VVPPER training data we provided, many instances end on *s* or *'s*, which is a common morphological property of contractions in German. On the one hand, this bias introduces a substantial amount of false positives - for instance the verb *weiß (to know)* occurs frequently in a misspelled form *weiss* in social media. On the other hand, this enables the SVM to also tag similar contraction cases of other word classes in *relaxed* or *all*.

5. Conclusion

We presented experiments that investigated how a PoS tagger can be designed that works as a corpus querying tool to find instances of rare phenomena. We experimented with altering the frequency weight of rare instances but found that adding relatively small amounts of additionally labelled data is unavoidable. By machine tagging data in which only the phenomenon of interest is manually corrected, we keep the effort minimal but yet achieve considerable improvements on detecting the phenomenon. We showed how recall is easily improved when forcing a tagger to focus more on the local word context. In a field study on plain text, we confirmed that our tagger works well as corpus query tool which finds accurately instances of the phenomenon of interest including many new ones. For future work, we plan to improve our method and also study the applicability to other under-represented phenomena.

Beißwenger, M., Bartz, T., Storrer, A., and Westpfahl, S. (2015). Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation / Tagset and guidelibes for the PoS tagging of langauge data from genres of computer-mediated communication. In *EmpiriST guideline document (German and English version)*.

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 44–56, Berlin, Germany.

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. In *Zeitschrift für germanistische Linguistik 41*, volume 1, pages 161–164.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. pages 597–620. Journal of Language and Computation.

Brants, T. (2000). TnT: A Statistical Part-of-speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, P. F., DeSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18:467–479.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, pages 256–263, Prague, Czech Republic.

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*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, pages 1735–1780.

Horsmann, T. and Zesch, T. (2016a). Assigning Fine-grained PoS Tags based on High-precision Coarse-grained Tagging. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 328 – 336, Osaka, Japan. Dublin City University and Association for Computational Linguistics.

Horsmann, T. and Zesch, T. (2016b). LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 120–126, Berlin, Germany. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of ACL2016*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

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

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 NACCL: HLT*, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Defining a protocol and assessing natural language metadata for a Databank of Oral Teletandem Interactions (DOTI)

Paola Leone
University of Salento
P.zzetta A. Rizzo, 1
Lecce (Italy)

paola.leone@unisalento.it

Abstract

The current study addresses the definition of a protocol for collecting, storing data and describing (in a simple and generic way) a repository. Particularly, the transparency of a form aimed at gathering information about the pedagogical context of oral telecollaboration for language learning named Teletandem (TT; Telles, 2006) will be tested before it is spread more widely. To uncover problems in submitting information, data-input-triggers quality and reliability have been tested interviewing professors and language instructors who will be involved in a preliminary phase of Teletandem corpus implementation. General goals of the study are to enlarge the research group, to increase data and to improve efficiency in data collection.

Keywords: corpora, data collection, metadata, telecollaboration, language learning, protocol

1. Teletandem as learning context

Teletandem (TT) is a learning context, in which pairs of native and non-native speakers of different languages talk alternatively in their L1 and L2 in order to learn each other's native language (www.teletandembrasil.org). The communication between partners is multimodal, via video calls and chat (e.g. by employing Skype, MSN). Students "virtually meet" during sessions that normally last one hour. From a pedagogical perspective, Teletandem has a positive impact on students' learning experience for different reasons: i) it is a form of learning from peers (Hanushek, Kain, Markman, & Rivkin, 2003), ii) it is extensible to perform and develop different forms of plurilingualism (e.g. to foster L2 production and/or to enhance interlocutors' ability to mutually understand each other when speaking different languages). Furthermore, since communication is enabled by VoIP technology, TT is a context for practicing computer mediated interactions which are becoming more and more used for job application and for university admission. The existence of websites which give "essential Skype interview tips" is a proof of the future relevance of this technology transmitted interaction. The positive impact of Teletandem in language learning (e.g. Leone & Telles, 2016) is a good basis for expecting a gradual, but sustained increase of its practice in higher education. This trend calls for further empirical research and implies a high demand for video/audio data.

2. Background information

Bearing the above in mind, two universities, in which for several years TT has been experienced, have undertaken a project which aims at organizing already existing Teletandem data and at defining natural language metadata. Teletandem databank is currently named Databank of Oral Teletandem Interaction (DOTI; Aranha & Leone, 2016; forthcoming) and it is build out of video recordings of Teletandem sessions thus data includes participants' moving visual images as well as voice and chat texts. Metadata allow the description of the context in which video recordings have been done and they are organized hierarchically. They stem from computer mediated interaction standardized metadata (Chanier, et al., 2014) and from pedagogical research (Mangenot, 2008); the former displays general characteristics of computer mediated interaction, the latter concerns the learning context and are currently not yet in a standardized form.

Following the general framework of "interaction space" (Chanier, et al., 2014), which includes comprehensively guidelines for describing all CMC genres in a multimodal perspective, teletandem as learning context is characterized in terms of: a) participants (i.e. n.2); b) location, meaning online location that is b.1) how interactions are transmitted (e.g. via VoIP technology, via mail etc.), b.2) where data are originally recorded (e.g. university server), b.3) the place where teletandem sessions happen (i.e. at the university or outside the institution); c) time frame which shows the beginning and the end of each session, but also the length of the telecollaboration project with specific information (days of the week, year, month). Concerning the technological environment, Teletandem is multimodal - i.e. diverse communicative modalities are used: audio-video via

VOIP technology and Internet Relay Chat (IRC)- ; communication is synchronous; it is a dyadic exchange since two people are involved; it implies oral, written, gestural and iconic modes of communication; language used can be different.
For recording characteristics related to the pedagogical implementation of Teletandem, the concepts of "learning scenario" and task have been referred to. The hierarchical organization of the learning scenario encompasses as well concepts of macrotasks (e.g. described in terms of number, typology) and microtasks.

3. Research

The general goal of the current research is to enlarge the research group involved in data collection, to increase data and to improve efficiency in data collection. Hinging on the above mentioned studies (Aranha & Leone, 2016, forthcoming), we aim at developing a protocol which offers guidelines and rules for forthcoming data collection (Aranha & Leone, forthcoming). In the protocol two main implementation levels are laid down. Level 1 consists in collecting Teletandem interactional video recordings as "raw data", storing them and describing video file contents using natural-language definition metadata (Aranha & Leone, 2016, forthcoming). Professors and/or language instructors will carry out this activity uploading video files in a cloud storage system and submitting as well - in a digital form - information which highlights properties of the specific TT learning context - i.e. information range from the institutional profile for the course in which TT has been institutionalized (Aranha & Cavalari, 2014) to the tasks assigned for the whole activity-. In level 1, researchers, professors and language instructors must agree on which terms to be used for property description and on which definition terms have. In level 1 DOTI will be enriched. Level 2 encompasses all those actions aimed to create an interoperational databank. It consists in implementing metadata interoperability by checking the compatibility of level 1 metadata with standardized metadata (e.g. Text Encoding Initiatives and Dublin Core Metadata Initiative). In this phase of corpora development, metadata will give access to digital data and they will be targeted to applied linguistic researchers.
The current research sticks into level 1 actions and it is a test of the clarity of the form which will be used to collect information about Teletandem as learning context and then for describing Teletandem databank. Informants are colleagues who already employed Teletandem in teaching. Research questions are: Are language instructors and professors familiar with terms employed as natural language metadata for describing Teletandem as a learning context in Higher Education? Does terminology facilitate entry of information? How can metadata be improved so to be more user friendly? The methodology for collecting colleagues' opinions is an in depth interview supported by a written questionnaire. Interviews have been recorded. Informants have to fill the form in (i.e. the assumption was "the form is well defined if colleagues can submit proper information in the right place") and then discuss if each question and/or concept description was/were clear (or not) and, if unclear, how it could be improved. The form contains data-input-triggers which are originated from the metadata provided for describing already collected data.

4. Results and conclusion

A first analysis of the findings shows that for some form fields (and subfields) there is coherence between required and given information (e.g. for the concept of learning scenario), whereas in other entries there was no match between purposes and information (e.g. macrotask and microtask are not easy to understand). To place interface terminologies in context, results will be discussed considering the kind of background informants show in the field of foreign language teaching. Suggestions on how natural language metadata could be improved and how the form interface layout could be implemented will be given.

References

Aranha, S., & Cavalari, S. (2014). A trajetória do projeto Teletandem Brasil: da modalidade institucional não-integrada à institucional integrada. *The ESPecialist*, 35(2), 70-88.

Aranha, S., & Leone, P. (2016). DOTI: Databank of Oral Teletandem Interactions. In S. Jager, & Kurek, M. (Eds.), *New directions in telecollaborative research and practice: selected papers from the second conference on telecollaboration in higher education* (pp. 327-332). Research-publishing.net.

Aranha, S., & Leone, P. (forthcoming). State of the arts and DOTI-Databank of Oral Teletandem Interaction.

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

Hanushek, E., Kain, J., Markman, J., & Rivkin, S. (2003). Does peer ability affect student achievement? *Applied econometrics*, 527-544.

Leone, P., & Telles, J. (2016). The Teletandem network. In T. Lewis, & R. O'Dowd (A cura di), *Online Intercultural Exchange: Policy, Pedagogy, Practice* (p. 243-248). London: Routledge.

Mangenot, F. (2008). La question du scénario de communication dans les interactions pédagogiques en ligne. *Jocair (Journées Communication et Apprentissage Instrumentés en Réseau*), 13-26 .

Telles, J. A. (2006). *Projeto Teletandem Brasil: Línguas Estrangeiras para Todos* – Ensinando e Aprendendo línguas estrangeiras in-tandem via MSN Messenger. Faculdade de Ciências e Letras de Assis, UNESP.

The #Idéo2017 Platform

Julien Longhi[1], Claudia Marinica[2], Nader Hassine[1,2], Abdulhafiz Alkhouli[2], Boris Borzic[2]
[1] University of Cergy-Pontoise, AGORA
[2] ETIS Lab UMR 8051 University of Paris-Seine, University of Cergy-Pontoise, ENSEA, CNRS
33 Boulevard du Port, 95000 Cergy-Pontoise, France
julien.longhi@u-cergy.fr, claudia.marinica@ensea.fr, nader2hassine@gmail.com,
abdulhafiz.alkhouli@ensea.fr, boris.borzic@ensea.fr

Abstract

The #Idéo2017 platform allows citizens to analyze the tweets of the 11 candidates at the French 2017 Presidential Election. #Idéo2017 processes the messages of the candidates by creating a corpus in almost real time. By using techniques from linguistics supplied with tools, #Idéo2017 is able to provide the main characteristics of the corpus and of the employment of the political lexicon, and allows comparisons between the different candidates.

Keywords: NLP for social media, NLP applications, textometry, tweets mining

1. Introduction

Social networks are becoming an important source for citizen's information, concerning mainly their "consumption" of information (Mercier, 2014). Twitter, the most known micro-blogging platform (Kaplan and Haenlein, 2010), by allowing the publication of short messages (140 characters), gives to social networks a new dimension. Indeed, Twitter can be used to assess how users react to social (Longhi and Saigh, 2016), political (Longhi et al., 2014; Conover et al., 2011), or economic issues. Therefore, the textual data (messages) sent on Twitter can be used to extract emotions, feelings, opinions, etc., of the users (Johnson and Goldwasser, 2016).

The analysis of political tweets during the election campaigns, or specific events, is increasing and can be seen as a specific type of political discourse (Longhi, 2013). Among the studies on this subject, Roginsky and Cock (2015) propose a qualitative analysis of interactions on Twitter, but they are limited to "the discursive and communicational analysis of the types of expression on Twitter that we can observe, with a particular interest in the way studied actors present and put forward themselves". Johnson and Goldwasser (2016) propose a classification of positionings based on the most frequent words, while the analysis of Vidak and Jackiewicz (2016) focuses on emotions. Moreover, many studies have proposed approaches to predict the result of the presidential elections (or to explain why the prediction is not possible) by analyzing the tweets (Tumasjan et al., 2010; Gayo-Avello et al., 2011; Metaxas et al., 2011).

Thus, there is an extensive literature on the analysis of political tweets, but these works are difficult to gather because they come either from the computer sciences, either from the humanities and social sciences (communication sciences, linguistics). Moreover, despite the unquestionable interest in outlining political facts, these results are not accessible by citizens interested in this subject.

In this context, this article presents #Idéo2017, a new and innovative platform making analytic information available to citizens. #Idéo2017 proposes a tool for analyzing tweets and speeches (relayed on Twitter) of the 11 candidates at the presidential election in France in 2017. #Idéo2017 analyzes the messages of the candidates by creating a corpus in almost real time (updated every 24 hours) with the tweets published in candidates' official accounts (from September 1st 2016 to May 7th 2017). Using techniques and metrics derived from linguistic tools, the new platform provides the main characteristics of the corpus and allows comparisons between the different candidates.

The rest of the paper is structured as follows. In Section 2., we provide a general description of the tool, and in Section 3. we present the analyses that can be carried out. Then, in Section 4. we detail the tool's development, as well as the technological choices that we made. Section 5. concludes the papers and provides a set of perspectives.

2. Description of #Idéo2017

#Idéo2017 is a web platform available online allowing to analyze the messages, posted on Twitter, related to political news (meetings, debates, television broadcasts, etc.). Its objective is to make available on the web for average citizens a set of statistical analyses and data visualization tools applied on the Twitter messages. The choice of a web platform rather than software to be installed comes from the fact that we want citizens that are non-specialists of tools and software to be able to have access to the analyses' results without going through the phases of corpus formation, tagging, etc. Thus, the citizens can make their own queries (based on linguistic and textometric criteria, more precisely, the most used words by political personalities, analyses of similarities, ALCESTE algorithm, etc.) and obtain comprehensible result.

The #Idéo2017 platform follows the processing chain shown in Figure 1: (1) retrieving the set of tweets of the candidates, (2) setting up a backup of tweets, (3) indexing tweets to facilitate the search process, (4) applying a set of linguistic analyses on tweets, (5) setting up a search engine on the tweets, and (6) displaying the results on a web page. In this processing chain, we are firstly interested in extracting candidates' tweets: we want to extract the tweets daily and to propose to the users to analyze the current database; for example, on April 4th, 2017 the users are able to analyze

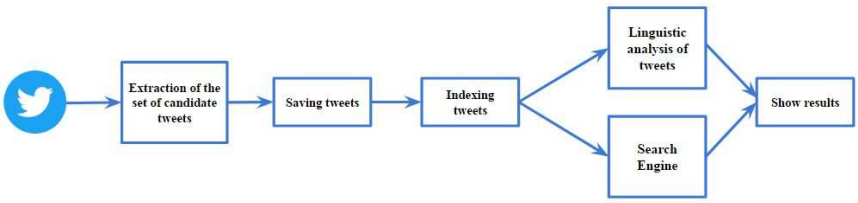


Figure 1: Processing chain in the #Idéo2017 platform.

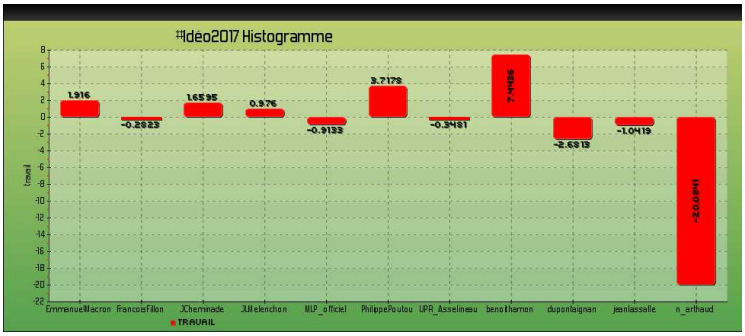


Figure 2: Analyses for the word *work*: under- / over- use.

the tweets sent by the candidates until April 3rd, 2017. To this end, we took advantage of the work that we carried out in the context of the extraction of the *Polititweets* corpus (Longhi et al., 2014), which is available and documented on the Ortolang platform[1].

In the tweets backup and indexing steps, we are interested in the issues of tweets storage, and, respectively, in the implementation of an indexing system. These two steps facilitate the access to the tweets and the development of an intelligent search engine.

In the linguistic analysis step, we propose to the user a set of analyses to be carried out on the set of tweets. These analyses, described in the next section, concern: the use of a specific word and its derivatives by the different candidates, the words associated with a specific word, the word cloud, themes, relations between words, and the specificities of the different candidates. In addition, we have developed an intelligent search engine based on the faceted search process that allows to the user to perform searches on tweets using complex filters.

3. Analyses and Search Engine Description

The #Idéo2017 platform, available at http://ideo2017.ensea.fr/plateforme/, proposes two types of analyses and the search engine. These three elements are described below.

3.1. The Analysis "I Analyze the Tweets that Contain the Word [Word]"

The analysis "I analyze the tweets that contain the word [word]" allows to the user to choose a word among the 13 words that are often used in political debates (Alduy, 2017).

[1] https://repository.ortolang.fr/api/content/comere/v3.3/cmr-polititweets.html

Our choice on limiting the user to 13 words is related to the computation time of analyses and graphics which would be too high if performed in real time. So, all the computations for the 13 words are performed at the same time (during the night) and the results are kept on the drive and displayed when requested. In a new version of our platform we plan to improve this aspect. On the other side, if the user wish to search for a word in the tweets he/she can use the search engine.

The list of 13 selected words is: *France*, *state*, *Republic*, *people*, *law*, *work*, *freedom*, *democracy*, *security*, *immigration*, *terrorism*, *Islam* and *secularism*.

Once the word is chosen, the user has access to four analyses:

- The first analysis allows to identify the use of the chosen word by the different candidates, and the results are presented in the form of two graphs: one for the computation of specificities (the under- / over- use of the word by the candidates), and the other one for the frequency of use of the word by the candidates.
- The second analysis detects the words associated with the chosen word for all candidates. This analysis of co-occurrences is presented in the form of a graph of associated words.
- The third analysis consists in computing the use of the chosen word and its derivatives (on contrary to the first analysis) by the different candidates.
- The last analysis creates a word cloud that allows to display graphically the lexicon.

To exemplify these analyses, let us consider the word *work* (*travail* in French). We can see in Figure 2 the under- / over- use of the word *work* (computation of specificities),

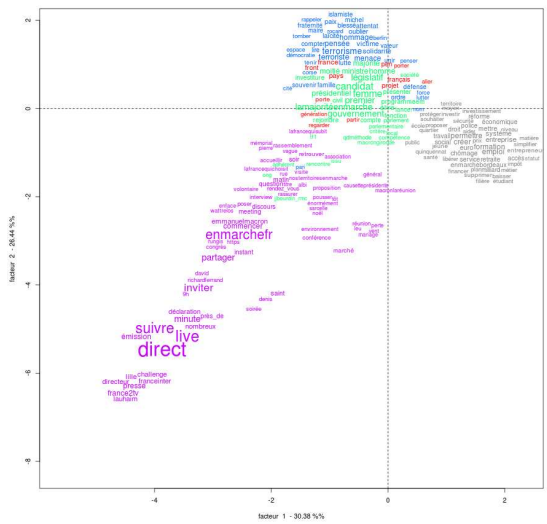


Figure 3: Analysis of the tweets of the candidate Emmanuel Macron: the themes (by lexical categories).

and in Figure 5 in Appendix Section the frequency of this word in all candidates' tweets.
The Figure 6 in Appendix Section allows us to see the words associated with the word *work* (analysis of similarities), and the Figure 10 in Appendix Section the word cloud.
It is important to outline that the analyses are performed after lemmatizing each word (for example, for the lemma liberty, we can have the forms liberty, liberties, etc.).

3.2. The Analysis "I Analyze the Tweets of [Candidate]"

The analysis "I analyze the tweets of [candidate]" allows to the user to perform a set of analyses on each of the 11 candidates: *N. Arthaud*, *F. Asselineau*, *J. Cheminade*, *N. Dupont-Aignan*, *F. Fillon*, *B. Hamon*, *J. Lassalle*, *M. Le Pen*, *E. Macron*, *J.-L. Mélenchon* and *P. Poutou*.
Once a candidate is chosen, the possible linguistic analyses are the following:

- The first analysis allows to detect and analyze the words the most used by the candidate;
- The second analysis, called "Themes", proposes, for the chosen candidate, to group together the words that are semantically close in order to outline the important themes discussed by the candidate;
- The third one analyzes the similarity between the words of each candidate in a graphical form;
- The forth analysis displays the lexicon of the tweets in the form of a word cloud;
- The last one is devoted to the analysis of the tweets of all the candidates. It allows to identify the specific words and categories of the different candidates and to compare them.

For example, let us consider the candidate Emmanuel Macron; the Figure 8 in Appendix Section presents the most frequent words used in his tweets, and the Figure 3 exposes the themes of his tweets.

3.3. The Search Engine

In the previous 2 sections we presented two type of analyses available in the #Idéo2017 platform; the third feature of #Idéo2017 is an intelligent search engine in real time (the graphical interface is shown in Figure 9 in Appendix Section) which offers a faceted search over the tweets: by candidate, by hashtag or by mention. Moreover, it provides complete search flexibility allowing to the user to compare the candidates using complex queries, and it also allows to sort the results by the date or by the commitment.
Twitter includes already a classical search engine in its interface; in order to propose a richer search experience, we built our search tool as a hybrid system bringing together the results of real time queries on the tweets and the synthesis of several tweets by aggregating the information via several facets (filters) and linguistic computations or word clouds. Thus, for a specific word or theme, our goal is to provide an access to the original tweets for each candidate, but also to compute the exact distribution of tweets per candidate and per theme.
The tweets' distribution information permits to contextualize each query, because our objective is at the same time to build a search engine, but also to propose a business intelligence system allowing to study the communication strategies of the candidates. It is important to outline that the two features (search engine and business intelligence system) have completely different goals: meanwhile the search engine struggles against the noise (all the answers should be the most pertinent), a business intelligence system, as a benchmark, aims to reduce the silence (all pertinent tweet should be shown to the user). However, when we try to reduce the silence, we increase the noise, and the more we fight against the noise, the more the silence becomes loud. In computer science, this complexity is assessed by two complementary metrics named the precision and the recall. To overcome this challenge, we took advantage of the applications in business intelligence (BI), the tools in reporting and the systems of knowledge management. Generally dedicated to dashboarding or back-office tools, we propose to the citizens to extend the queries results with a synthetic information provided by the linguistic analyses and integrated visually and progressively.

4. Tool Development

For the development of the tool, we had to tackle different technological problems. We will present the solutions, shown in Figure 11 in Appendix Section, that we have chosen for each problem. First, we used the Twitter API to retrieve the tweets directly from the official accounts. Then, we stored these tweets in the MongoDB[2] NoSql database; its advantage consists in a flexible, document-oriented structure that does not require complex queries to access the data. Then, we decided to use Elasticsearch to

[2] www.mongodb.com

store the tweets; Elasticsearch (Kononenko et al., 2014) improves the response time of our tool especially when using the search engine.
Given that Elasticsearch's standard method performs a classical search without dealing with the derivatives of a word, before sending the data from MongoDB to Elasticsearch, we prepare a data index that takes into account the derivatives of the words. All the communication between these tools uses the Java language. For the analysis part, we compared several software packages for the linguistic analysis of corpus (Hyperbase, Lexico3, Trameur, TXM and Iramuteq). We studied their analyses but also their availability in open source and/or API. After our study, we decided to use several features of Iramuteq[3] that are implemented in PHP and available in open source. To this end, several modifications were needed in the implementation of Iramuteq. We also used PHP Word Cloud[4] for word clouds and pChart[5] for graphics.
After the election, a second version of the platform was released with new features: the first one, for data visualizations (as shown in Figure 7 in Appendix Section), and the second one, for sub-corpus extraction from the complete corpus by the candidate and the period (as shown in Figure 4). The latter is very important and useful for the humanities and social sciences community. Indeed, researchers in this area are interested in specific political issues, but they do not have access to tools/platforms allowing them to extract and structure their corpus before usage.

Figure 4: Corpus generator feature.

5. Conclusion and Perspectives

#Idéo2017 combines different technologies and inputs, which give to citizens the opportunity to grasp a part of the discursive issues of the election. This development, which can be enriched, allows to easily use a set of features usually accessible by software requiring different transformations of the data.

5.1. Creation of the #Idéo2017 Corpus

For the period from September 1st 2016 to May 7th 2017, 42290 tweets were extracted for the 11 candidates. These tweets were gathered in a collection that will be published in a TEI corpus in the standards of the Ortolang platform. The publication of this corpus under the requested standard is founded with the support of the CORLI consortium[6]. This process will follow the guidelines listed in the acquisition report (Longhi, 2014) written when we released the Polititweets corpus. Concerning the juridical issues, the creation of this corpus is legal. The position of Twitter is the following:

- "Please review the Twitter Rules (which are part of these Terms) to better understand what is prohibited on the Service. We reserve the right at all times (but will not have an obligation) to remove or refuse to distribute any Content on the Services, to suspend or terminate users, and to reclaim usernames without liability to you. We also reserve the right to access, read, preserve, and disclose any information as we reasonably believe is necessary to (i) satisfy any applicable law, regulation, legal process or governmental request, (ii) enforce the Terms, including investigation of potential violations hereof, (iii) detect, prevent, or otherwise address fraud, security or technical issues, (iv) respond to user support requests, or (v) protect the rights, property or safety of Twitter, its users and the public."

- "Except as permitted through the Services, these Terms, or the terms provided on dev.twitter.com, you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Content or Services."

Thus, Twitter does not disclose personally identifying information to third parties except in accordance with their Privacy Policy. Moreover, Twitter encourages and allows broad re-use of content. The Twitter API exists to enable this.

5.2. #Idéo2017 as a Prototype: the Reproducibility of the Platform

#Idéo2017 allowed to the French electors to analyze the discourse of the candidates by means of their tweets. But, the utility of this platform is not limited to the French election, because it can be modified to different needs and usages, and adapted to other contexts. Thus, after the presidential election, we released two other versions of the platform: #législatives2017 (http://ideo2017.ensea.fr/legislatives2017/) and #quinquennat (http://ideo2017.ensea.fr/quinquennat/) which allow, in the first case to analyze the tweets of the main political parties during the election of deputies to the French National Assembly, and in the second case to daily analyze the beginning of Emmanuel Macron's presidential mandate via the tweets of the current political protagonists.
A set of new features were also proposed such as statistical analyses of the hashtags and mentions with Kibana. As a perspective, the #Idéo2017 prototype could be adapted to be used as a competitive intelligence, a measurement and a visualization tool analyzing people's opinion on Twitter. We can imagine dealing with political, social or cultural subjects through the integration of influential accounts, but also individual accounts of the users.

[3] www.iramuteq.org
[4] github.com/sixty-nine/PHP_Word_Cloud
[5] www.pchart.net
[6] https://corli.huma-num.fr/

6. Acknowledgements

UCP Foundation, Vo tv/Vo News, IUT of Cergy-Pontoise, University of Cergy-Pontoise, ETIS Lab, AGORA Lab.

7. References

Alduy, C. (2017). *Ce Qu'Ils Disent Vraiment. Decoder Le Discours Des Presidentiables*. SEUIL.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Gayo-Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In Lada A. Adamic, et al., editors, *ICWSM*. The AAAI Press.

Johnson, K. and Goldwasser, D. (2016). "all I know about politics is what I read in twitter": Weakly supervised models for extracting politicians' stances from twitter. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2966–2977.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68.

Kononenko, O., Baysal, O., Holmes, R., and Godfrey, M. W. (2014). Mining modern repositories with elasticsearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, pages 328–331, New York, NY, USA. ACM.

Longhi, J. and Saigh, D. (2016). A textometrical analysis of french arts workers "fr.intermittents" on twitter. In *4th conference CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenia, September.

Longhi, J., Marinica, C., Borzic, B., and Alkhouli, A. (2014). Polititweets : corpus de tweets provenant de comptes politiques influents. In Chanier T., editor, *Banque de corpus CoMeRe*. Ortolang.fr.

Longhi, J. (2013). Essai de caractérisation du tweet politique. *L'information grammaticale*.

Longhi, J. (2014). Proposition pour l'acquisition d'un corpus de tweets (cmr-polititweets-teiv1-manuel.pdf). In J. Longhi, et al., editors, *Polititweets : corpus de tweets provenant de comptes politiques influents*. Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets-tei-v1].

Mercier, A. (2014). Twitter l'actualité : usages et réseautage chez les journalistes français. recherches en communication. *Recherches en communication 40*, pages 111–132.

Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. (2011). How (not) to predict elections. In *SocialCom/PASSAT*, pages 165–171. IEEE.

Roginsky, S. and Cock, B. D. (2015). Faire campagne sur twitter. modalités d'énonciation et mises en récit des candidats à l'élection européenne. *Les Cahiers du numérique*, 11(4):119–144.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vidak, M. and Jackiewicz, A. (2016). Les outils multimodaux de twitter comme moyens d'expression des émotions et des prises de position. *Cahiers de praxématique*, (66).

Appendix

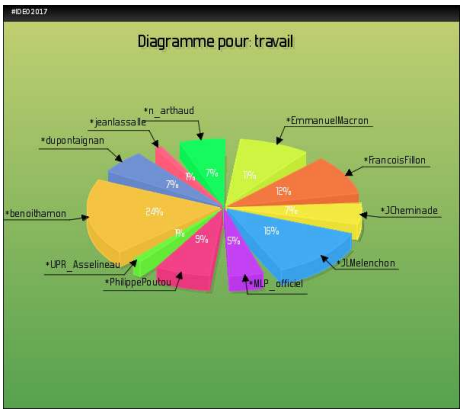


Figure 5: Analyses for the word *work*: frequency.

Figure 6: Analyses for the word *work*: associated words.

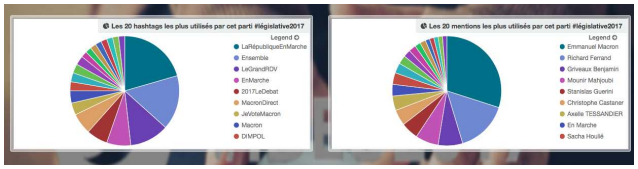


Figure 7: Data visualization.

Les mots les plus utilisés de @EmmanuelMacron

Forme	Fréquence	Type
france	223	Nom
français	184	Adjectif
projet	155	Nom
politique	153	Nom
aller	136	Verbe
pays	132	Nom
europe	125	Non reconnue
donner	89	Verbe
ensemble	85	Adverbe

Figure 8: Analysis of the tweets of the candidate Emmanuel Macron: the most frequent words.

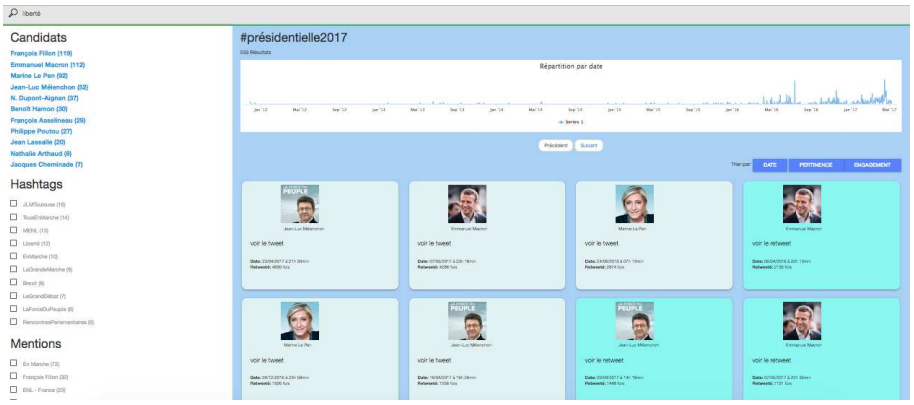


Figure 9: The user interface of the search engine.

Figure 10: Analyses for the word *work*: word cloud.

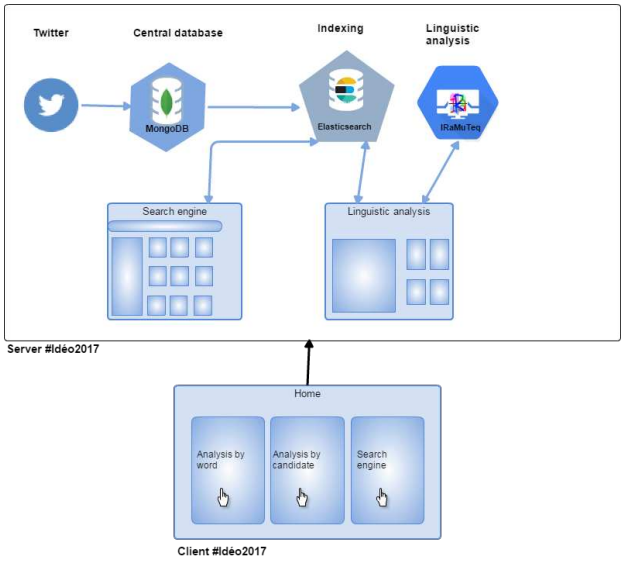


Figure 11: Schema of the tool.

Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages?

Michael Beißwenger[1], Ciara Wigham[2], Carole Etienne[3], Darja Fišer[4], Holger Grumt Suárez[5], Laura Herzberg[6], Erhard Hinrichs[7], Tobias Horsmann[1], Natali Karlova-Bourbonus[5], Lothar Lemnitzer[8], Julien Longhi[9], Harald Lüngen[10], Lydia-Mai Ho-Dac[11], Christophe Parisse[12], Céline Poudat[13], Thomas Schmidt[10], Egon Steml[14], Angelika Storrer[6], Torsten Zesch[1]

[1] University of Duisburg-Essen, Germany [2] University Clermont-Auvergne, France [3] ICAR Laboratory Lyon, France
[4] University of Ljubljana, Slovenia [5] Justus-Liebig-Universität Gießen, Germany [6] University of Mannheim, Germany
[7] Eberhard-Karls-Universität Tübingen, Germany [8] Berlin-Brandenburg Academy of Sciences, Germany
[9] Université de Cergy-Pontoise, France [10] Institute for the German Language, Mannheim, Germany
[11] Université Toulouse 2, France [12] Université Paris Nanterre, France
[13] Université Nice Côte d'Azur, France [14] Eurac Research, Bolzano, Italy

michael.beisswenger@uni-due.de, ciara.wigham@uca.fr, carole.etienne@ens-lyon.fr, darja.fiser@ff.uni-lj.si, Holger.H.Grumt-Suarez@germanistik.uni-giessen.de, lherzber@mail.uni-mannheim.de, erhard.hinrichs@uni-tuebingen.de, tobias.horsmann@uni-due.de, Natali.Karlova-Bourbonus@zmi.uni-giessen.de, lemnitzer@bbaw.de, julien.longhi@u-cergy.fr, luengen@ids-mannheim.de, hodac@univ-tlse2.fr, cparisse@u-paris10.fr, celine.poudat@unice.fr, thomas.schmidt@ids-mannheim.de, egon.stemle@eurac.edu, astorrer@mail.uni-mannheim.de, torsten.zesch@uni-due.de

The paper reports on the results of a scientific colloquium dedicated to the creation of standards and best practices which are needed to facilitate the integration of language resources for CMC stemming from different origins and the linguistic analysis of CMC phenomena in different languages and genres. The key issue to be solved is that of interoperability – with respect to the structural representation of CMC genres, linguistic annotations metadata, and anonymization/pseudonymization schemas. The objective of the paper is to convince more projects to partake in a discussion about standards for CMC corpora and for the creation of a CMC corpus infrastructure across languages and genres. In view of the broad range of corpus projects which are currently underway all over Europe, there is a great window of opportunity for the creation of standards in a bottom-up approach.

Keywords: corpora, research infrastructures, annotation, anonymization

1. Background and Motivation

The paper reports on the results of a scientific colloquium (https://sites.google.com/view/dhcmc2017/) dedicated to the creation of standards and best practices which are needed as a prerequisite for the exchange, interconnection, and combined analysis of CMC corpora of different origins, for different languages and different genres. The goal of the colloquium which was held with funding from the French Embassy in Germany was to determine open issues which have to be solved to represent CMC corpus data (including metadata and annotations) using interoperable formats. From a wider perspective, the colloquium addressed not only issues of interoperability of CMC corpora between one another but also the interoperability of CMC corpora with corpora of other types, namely text corpora and spoken language corpora.

To make the goal of interoperability and of the development of standards more tangible for the community of CMC researchers and corpus creators, the colloquium outlined the scenario of creating a multilingual and genre-heterogeneous demo corpus that would include samples from existing CMC corpora in different languages and on different CMC genres and which would also present CMC in the context of other discourse domains through the inclusion of samples from text and spoken language corpora.

In the following sections, we summarize results and remaining open issues towards the creation of standards and towards an interoperability of corpora as were determined during the colloquium. The overview is based on input from representatives of the following corpus projects and language resource infrastructure projects:

a) *CMC corpora:*

- *CoMeRe:* a collection of 14 French corpora for 9 different CMC genres (including multimodal genres) represented in TEI (Chanier et al., 2014) and available for download (CC BY, OpenData) via ORTOLANG (Longhi and Wigham, 2015)[1]
- *DeReKo-News:* Corpus of German Newsgroups in DeReKo, since 2013, 98 million tokens (Schröck and Lüngen, 2015).[2]
- *DeReKo-Wikipedia:* Wikipedia corpora in DeReKo: German language article, talk and user talk (Margaretha and Lüngen, 2014), 581 million tokens, available for online querying via COSMAS II.
- *DiDi corpus:* The CMC corpus from the DiDi project with 570.000 tokens of German, Italian and South Tyrolean Facebook posts and interactions, available

[1] http://hdl.handle.net/11403/comere
[2] https://cosmas2.ids-mannheim.de/

for online querying via ANNIS (Frey et al., 2016).[3]

- *Dortmund Chat Corpus 2.0:* corpus of German chat discourse represented in TEI, available as part of the CLARIN-D corpus infrastructure (Lüngen et al., 2016, Beißwenger et al., 2017).
- *DWDS blog corpus:* The blog corpus in the corpus collection of the DWDS project: 103 million tokens from CC-licensed, mainly German blog entries, available for online querying (Barbaresi, 2016).[4]
- *Gießen scienceblog corpus:* ongoing project at the University of Gießen; goal: creation and annotation of a corpus of German science blogs. (Grumt Suárez et al., 2016)
- *Janes corpus:* The Corpus of Nonstandard Slovene comprising >200 million tokens from tweets, forum posts, blogs, comments on news articles and Wikipedia discussions (Fišer et al., 2016, 2017).[5]
- *MoCoDa:* ongoing project at University of Duisburg-Essen (2017–), goal: creation of a database with a web frontend for repeated, donation-based collection of mobile CMC (whatsapp, sms & co.).
- *Wikiconflits:* TEI-CMC-encoded corpus of French Wikipedia talk pages associated to conflicts capturing 7 topics related to (pseudo-)science with 4456 posts (Poudat et al., 2017).
- *WikiTalk:* TEI-P5-encoded corpus of French Wikipedia talk pages (365,612 pages, >1M threads).
- *DiscoWiki:* corpus of an ongoing project with the goal of annotating relevant characteristics for conflict detection and description at the thread level; corpus based on a selection of talk pages extracted from *Wikiconflits* and *WikiTalk* (Ho-Dac and Laippala, 2017).

b) *Corpora of other types:*

- Text corpus collection at the BBAW, Berlin (DWDS corpora) (Geyken et al., 2017).
- German reference corpus at the IDS Mannheim (DeReKo) (Lüngen, 2017).
- The French spoken language corpora *Colaje* (Morgenstern and Parisse, 2012) and *Orfeo*[6].
- Research and Teaching Corpus of Spoken German (FOLK) at the IDS Mannheim (Schmidt, 2016).[7]

c) *Corpus infrastructure projects:*

- *CLARIN-D*, the German national branch of the Eurpoean CLARIN initiative.[8]
- *ORTOLANG*, the French infrastructure for Open Resources and TOols for LANGuage.[9]

The objective of presenting this summary at CMC-Corpora17 is to convince more people with corpus projects in the field to join the discussion about standards for CMC corpora and – probably – for transforming the idea of a CMC demo corpus into a cooperative project with the participation of a broad range of projects and researchers.

2. State of the art and open issues

2.1 Basic representation format

Since 2012, the special interest group (SIG) "computer-mediated communication" in the Text Encoding Initiative (TEI) has created three TEI extensions for the representation of CMC data and tested these extensions with different CMC genres and corpora for French and German (Beißwenger et al., 2012, Chanier et al., 2014, Lüngen et al., 2016). All three extensions are available in the form of RNG schemas and ODD documents and ready to be used for annotation in other projects.[10] Current work of the SIG is focused on the transformation of the available extensions into a "feature request" which is necessary to make an official suggestion for extending the TEI guidelines with models for CMC. Discussions at the colloquium showed that for further dissemination of the TEI extensions for CMC is desirable

- to document practices, tools, and guideline documents from projects that have already converted raw data into TEI and make them available as Open Access resources to facilitate the conversion of corpus data into TEI for colleagues who have not worked with TEI before;
- to diffuse information/documentation concerning toolchains that can be used on CMC data as currently there is little support for users to help them process resources in one way or another.

One possible outlet for the diffusion of resources of this type would be the CLARIN Language resource switchboard (Zinn, 2016) that is currently being developed within CLARIN-PLUS as a means to link linguistic resources with the tools that can process them. It aims to create a single point of access where users can find the tools that fit their needs and their language resource.

2.2 Natural language processing

Different creators of CMC corpora are using a different types of linguistic annotations and different (e.g., language-specific) tagsets. For mapping annotations in existing corpus resources without the need to perform a complete re-annotation, the resources of the Universal Dependencies Initiative (UD)[11] may provide CMC researchers with "a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary". Nevertheless, UD does not provide any tags for the

[3] http://www.eurac.edu/didi
[4] https://www.dwds.de
[5] http://nl.ijs.si/janes/
[6] http://www.projet-orfeo.fr/
[7] http://agd.ids-mannheim.de/folk.shtml
[8] https://www.clarin-d.de/de/
[9] http://ortolang.fr

[10] These resources are available via the SIG space in the TEI wiki: https://wiki.tei-c.org/index.php?title=SIG:CMC
[11] http://universaldependencies.org/introduction.html

description of CMC-specific phenomena. In order to determine the feasibility of mapping tagset extensions used in CMC corpora in different languages onto each other, the group agreed that it will be helpful to compare how CMC-specific phenomena are treated in the DiDi, CoMeRe, CLARIN-D and Janes (and other CMC) corpora and check the extent to which they are already compatible with each other / analyse the effort needed to transform them into a compatible structure. This could be the subject of a workshop or short-term project in the near future. Further work on this topic should also take into account the tagsets and resources both from the PoSTWITA shared task on PoS tagging Italian social media data[12] and from the EmpiriST shared task on PoS tagging of German CMC and web corpora data[13].

2.3 Anonymization

In order for CMC corpora to conform to the restrictions of national data protection rights (DPR), existing corpora projects have developed different strategies and practices for removing or masking personal data. From a DPR perspective, it is required that the data included in corpora which are made available to the scientific community should be represented in a way that the expenses that one would have to invest to identify a certain person are so high that it seems unrealistic that anybody would invest them. Different annymization approaches have been adopted by the different projects. It may be a fruitful topic for further investigation (i.e. in form of a workshop or short-term project) to compare, in detail, the results of anonymization vs. pseudonymization approaches adopted in different projects to determine the best balance between preserving as much of the semantics of the original data as possible (which is an important resource especially for qualitative analyses) while on the other hand staying as feasible as possible in terms of time-cost factors. One action point could be to make anonymization guidelines from different projects available to other colleagues so that procedures employed can be re-used. Examples for best practices for anonymizing CMC corpora have been developed and employed in the DiDi project (Frey et al., 2015), in the CLARIN-D curation project ChatCorpus2 CLARIN (Lüngen et al., 2017) and in CoMeRe (Chanier and Jin, 2013).

2.4 Metadata

The creation and representation of metadata for CMC corpora and for single interactions preserved in them is a huge and urgent open issue. One key point is for metadata to include information about the version of the communication platform, given the rate at which communication platforms evolve. In the future, should access to older version no longer be available, at least prose descriptions of the communication platform at the time of data collection would be practical for future corpora end-users and mandatory to guarantee corpus data sustainability. A comparison of the metadata captured in different corpus projects (be it in the form of annotations, be it in the form of prose descriptions) could be a fruitful contribution to a more precise discussion of (i) what types of metadata are needed in and for CMC corpora, (ii) to what extent these metadata are specific to individual CMC genres or for CMC in general, (iii) the preservation of which types of metadata could be standardized, and (iv) how a basic representation schema for CMC metadata (e.g., in the TEI Header) could look like.

3. Outlook

In view of the broad range of corpus projects which are currently underway all over Europe (Beißwenger et al., 2017a), there is a great window of opportunity for the creation of standards for CMC corpora in a bottom-up approach. The discussions obtained on this issue at CMCCorpora17 shall be included in the creation of a white paper giving a more precise outline of future work for the issues addressed in this paper. The creation of a demo corpus including samples from different existing CMC corpora could support the further investigation of open issues and provide valuable feed-back for existing best practices in the field. A prerequisite would be a “critical mass” of resources and researchers who are willing to contribute to the creation of such a corpus. As a first step of preparatory work it is planned to set up a platform for the exchange of tools, tips and case studies between projects in order to facilitate the dissemination of knowledge and best practices.

4. References

Barbaresi, A. (2016). Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop, Association for Computational Linguistics*, pp. 7-16. https://hal.archives-ouvertes.fr/hal-01371704v2/document.

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* 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Beißwenger, M., Lüngen, H., Schallaböck, J., Weitzmann, J.H., Herold, A., Kamocki, P., Storrer, A., Wildgans, J. (2017, in press). Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In: M. Beißwenger (Ed.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin/New York: de Gruyter (Empirische Linguistik / Empirical Linguistics).

Beißwenger, M., Chanier, T., Erjavec, T., Fišer, D., Herold, A., Lubešic, N., Lüngen, H., Poudat, C., Stemle, E., Storrer, A., Wigham, C. (2017a). Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In: L. Borin (Ed.), *Selected papers from the CLARIN Annual*

[12] http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=guidelines

[13] Tagset/guidelines: https://sites.google.com/site/empirist2015/; results: WAC-X/EmpiriST (2016).

Conference 2016, Aix-en-Provence, 26–28 October 2016 (Linköping University Electronic Conference Proceedings 136), pp. 1-18. http://www.ep.liu.se/ecp/contents.asp?issue=136

Chanier, T., Jin, K. (2013). *Defining the online interaction space and the TEI structure for CoMeRe corpora. Projet CoMeRe (Communication Médiée par les Réseaux).* https://corpuscomere.files.wordpress.com/2014/01/tei-cmc-comere-interactionspace_131231.pdf

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. *Journal of language Technology and Computational Linguistics*, 29(2), pp. 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf.

Fišer, D., Erjavec, T., Ljubešić, N. (2017). The compilation, processing and analysis of the Janes corpus of Slovene user-generated content: In C.R. Wigham, G. Ledegen (Eds.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse*. Paris: L'Harmattan (Humanités numériques), pp. 125–138.

Fišer, D., Erjavec, T., Ljubešić, N. (2016). JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2), pp. 67–99.

Frey, J.C., Glaznieks, A., Stemle, E.W. (2015). The DiDi Cor-pus of South Tyr-olean CMC Data. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015), Essen, Germany*.

Frey, J.-C., Glaznieks, A., Stemle, E. (2016). The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*. ceur-ws.org/Vol-1749/paper27.pdf.

Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., Lemnitzer, L. (2017, in press). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, 45 (2).

Grumt Suárez, H., Karlova-Bourbonus, N., 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 (cmc-corpora2016), Ljubljana, Slovenia*. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Grumt_et_al_Compilation-and-Annotation.pdf.

Ho-Dac, L.-M., Laippala, V. (2017). Le corpus WikiDisc, une ressource pour la caractérisation des discussions en ligne. In C.R. Wigham, G. Ledegen (Eds.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse.* Paris: L'Harmattan (Humanités numériques), pp. 107–124.

Longhi, J., Wigham, C.R. (2015). *Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow.* Poster at Corpus Linguistics 2015, Lancaster, United Kingdom. https://halshs.archives-ouvertes.fr/halshs-01176061.

Lüngen, H. (2017). DeReKo – Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim. *Zeitschrift für germanistische Linguistik*, 45 (1), pp. 161–170.

Lüngen, H., Beißwenger, M., Ehrhardt, E., Herold, A., Storrer, A. (2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), Bochum, Germany*, pp. 156–164. https://www.linguistics.rub.de/konvens16/pub/20_ konvensproc.pdf.

Lüngen, H., Beißwenger, M., Herzberg, L., Pichler, C. (2017). Anonymisation of the Dortmund Chat Corpus 2.1. In *Proceedings of the 5th Conference on CMC and Social Media corpora for the Humanities, Bolzano, Italy.*

Margaretha, E., Lüngen, H. (2014). Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics*, 29(2), pp. 59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf.

Morgenstern, A. & Parisse, C. (2012). The Paris corpus. *Journal of French Language Studies*, 22, pp. 7–12.

Poudat, C., Grabar, N., Paloque-Berges, C., Chanier, T., Juin, K. (2017). Wikiconflits: un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In C.R. Wigham, G. Ledegen (Eds.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse.* Paris: L'Harmattan (Humanités numériques), pp. 19–36.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. *International Journal of Corpus Linguistics*, 21(3), pp. 396–418.

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), Essen, Germany*, pp. 17–22. https://sites.google.com/site/nlp4cmc2015/program

[WAC-X/EmpiriST 2016] Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26). http://aclweb.org/anthology/W/W16/W16-26.pdf.

Zinn, C. (2016). The CLARIN Language Resource Switchboard. In *Proceedings of the CLARIN Annual Conference*, *Aix-en-Provence, France*, https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf.

"You're trolling because…" – A Corpus-based Study of Perceived Trolling and Motive Attribution in the Comment Threads of Three British Political Blogs

Márton Petykó
Department of Linguistics and English Language, Lancaster University
County South, Lancaster University, Lancaster LA1 4YL, United Kingdom
E-mail: m.petyko@lancaster.ac.uk

Abstract

This paper investigates the linguistically marked motives that participants attribute to those they call trolls in 991 comment threads of three British political blogs. The study is concerned with how these motives affect the discursive construction of trolling and trolls. Another goal of the paper is to examine whether the mainly emotional motives ascribed to trolls in the academic literature correspond with those that the participants attribute to the alleged trolls in the analysed threads. The paper identifies five broad motives ascribed to trolls: emotional/mental health-related/social reasons, financial gain, political beliefs, being employed by a political body, and unspecified political affiliation. It also points out that depending on these motives, trolling and trolls are constructed in various ways. Finally, the study argues that participants attribute motives to trolls not only to explain their behaviour but also to insult them.

Keywords: troll(ing), motive attribution, blog

1. Introduction

This corpus-based case study investigates a prominent social phenomenon of computer-mediated communication: trolling. It aims to identify the linguistically marked motives that participants attribute to those whom they call trolls in 991 comment threads. These threads were published on three British political blogs, *Guardian Politics Blog*, *Guido Fawkes*, and *LabourList*. The paper is also concerned with how these motives affect the way trolling and trolls are discursively constructed in the threads. Another goal of the paper is to examine to what extent the motives attributed to trolls in the academic literature correspond with those that the participants attribute to the alleged trolls.

The analysis focuses on 2,036 motivation-related metapragmatic comments taken from these 991 threads. In these comments, participants call other users trolls or identify comments as trolling and also discuss the possible reasons why the alleged trolls are trolling. The study first presents a taxonomy of the linguistically marked motives in these comments and then it applies this taxonomy to annotate the comments. Thus, it develops a discursive-pragmatic annotation system for linguistically marked motive attribution in computer-mediated interactions.

This study can be situated within the fields of corpus-based discourse analysis (Baker, 2006) and pragmatics (Culpeper & Hardaker, 2016). Beyond trolling, the paper has relevance to the pragmatics of computer-mediated communication (Herring, Stein, & Virtanen, 2013) and within that, to the study of metapragmatic comments in computer-mediated interactions (Tanskanen, 2007).

2. Literature Review

'Trolling' is usually described as a set of goal-driven behaviours, while 'troll' is deemed a behaviour-based identity (Hardaker, 2013). The most often mentioned goals attributed to trolls are: attracting other users' full attention (Hardaker, 2010), triggering intense unpleasant emotional reactions (Thacker & Griffiths, 2012), eliciting potentially offensive responses from others (Morrissey, 2010), causing, perpetuating or escalating conflict (Galán-García et al., 2014), disrupting the ongoing interaction (Binns, 2012), and deceiving or manipulating others (Donath, 1999).

The discursive actions perceived as acts of trolling are: repeating the same utterance (Shachaf–Hara, 2010), posting irrelevant or meaningless information (Morrissey, 2010), posting misleading or factually incorrect information (Hardaker, 2010), disseminating bad and/or dangerous advice (Donath, 1999), ignoring, despising, rejecting or attacking the core values of the interaction (Utz, 2005), (hypo)criticising others (Hardaker, 2013), and directly insulting, threatening or otherwise attacking others (Herring et al., 2002).

Although the motives for trolling are also often mentioned in the literature, most studies do not attempt to empirically examine them but they instead treat them in a speculative manner (Hopkinson, 2013). This is a clear gap in the literature, to which this study is related.

Trolling is usually approached as an emotionally motivated individual behaviour. The most often mentioned motive is that trolls engage in this behaviour because they simply enjoy it or its consequences (Hardaker, 2010). Further emotional motives are also mentioned, such as boredom (Baker, 2001), a need for attention or achievement, revenge (Shachaf & Hara, 2010), loneliness, curiosity, malevolence (Fichman & Sanfilippo, 2015), a desire for control and self-empowerment, hate towards specific participants, and hostility to the purpose of the interaction (Herring et al., 2002). It is also suggested that trolls can be motivated by specific political goals and (political) ideologies (Özsoy, 2015). A key aim of this study is to examine whether the above-mentioned motives correspond with those that the participants attribute to the alleged trolls.

3. Data and Method

3.1. Data collection

The corpus consists of 991 comment threads of three British political blogs, *Guardian Politics Blog* (GP), *Guido Fawkes* (GF), and *LabourList* (LL). In this paper, a 'thread' refers to the comments of a blog post. These 991 comment

threads thus include 617,782 comments of 991 blog posts. The size of the corpus is around 21.9 million tokens.

GP is the political blog section of a major British newspaper, *The Guardian*. GP can be characterised as a liberal centre-left political blog with more permanent contributors and a highly diverse readership. The blog posts are written by professional journalists in a neutral manner while the commenters represent the entire political spectrum.

GF is an independent libertarian and anti-establishment political blog, which was founded by Paul Staines. Whilst GP has more authors, Staines most likely remains the main contributor. The blog posts are often sarcastic and overtly criticise or mock the major British political parties, such as the Conservative Party, the Labour Party, and the Liberal Democrats, and their leading politicians. Similarly to most political blogs, the commenters do not form a homogeneous community. However, many of them explicitly support the right-wing UK Independence Party (UKIP). This strongly relates to GF's anti-establishment stance as many perceive UKIP as an anti-establishment party.

LL overtly supports the centre-left Labour Party and aims to provide a forum for debate within the Labour Party. The blog posts are written by numerous contributors. While LL itself is said to be independent from the Labour Party, many of the contributors are Labours MPs or are otherwhise affiliated with the Labour Party. Rather unsurprisingly, most commenters support the Labour Party and have left-wing leanings.

The threads were selected based on two criteria: (1) The thread had to be published on GP, GF or LL between 1 January and 31 December 2015. (2) The thread had to include at least one comment in which a participant called at least one other participant a troll and/or described at least one comment as an act of trolling at least once (hereon referred to as a 'troll comment'). That is, at least one participant had to use a word form of the lexeme TROLL, such as *troll*, *trolling* or *troller* to refer to another participant or comment as illustrated in example (1).

(1) [guardian_65_22345]
stop posting rubbish, ***troll****!*

Data collection included the following steps:

(1) A list of 50 British political blogs active in 2015 was compiled. I considered a blog to be any website appearing on a blog hosting platform, such as blogspot.com and/or that called itself a blog. They were deemed to be active in 2015 if at least one post was published between 1 January and 31 December 2015. Finally, I classified political blogs as those whose main topic is politics, i.e. the acquisition, distribution and practice of power in human communities, societies and states. Four sources for collection were used:

(a) **Teads list of top 100 British political blogs in September 2015.** Teads is a French technology company expert in video advertising solutions. It regularly publishes a list of top 100 British political blogs on its website.

(b) **Vuelio list of top 10 UK political blogs in October 2015.** Vuelio is a leading global provider of PR and Political Services Software. It publishes a list of the top10 UK political blogs.

(c) **Google search.** The search terms were *British political blog*, *"British political blog"* 'British political blog as exact term', *UK political blog* and *"UK political blog"* 'UK political blog as exact search term'.

(d) **The political blogs recommended on the already collected ones** were also considered.

(2) I gathered all those threads from these 50 blogs in which at least one participant/comment was deemed to be a troll/trolling. I manually searched 26,804 threads from 2015 for the *troll* character string, and found 1,712 relevant threads. Then I saved each thread in a separate txt file.

(3) For the purposes of this case study, I selected the first three blogs, GP, GF, and LL since these had the highest number of qualifying threads. I decided to focus on only these three blogs in this paper since although the original list consisted of 50 political blogs that cover the entire political spectrum from far right to far left, 58% of the collected troll threads come from these three blogs. Thus, GP, GF, and LL are the key British political blogs for analysing perceived trolling in the British political blogosphere and their troll threads consitute an adequate sample of the more comprehensive corpus that includes all the 1,712 troll threads of the 50 blogs. Furthermore, the aim of this paper is not to draw general conclusions on perceived trolling in the British political blogosphere but to provide a context-sensitive analysis of the motives attributed to trolls on three British political blogs where participants call others trolls considerably more often than on other British political blogs.

(4) Four versions of the corpus were created. Version 1 consists of complete comment threads with blog posts and metadata (nicknames, dates, URLs etc.). Version 2 also includes complete comment threads but without the blog posts and any metadata. The troll comments (<tc></tc>) and the troll tokens within them (<tt></tt>) are also annotated in this version. Version 3 has only the troll comments while Version 4 contains all non-troll comments.

Table 1 includes the number of blog posts, comments, tokens, troll comments, and troll tokens in the second version of the corpus.

	Overall	GP	GF	LL
Threads	991	167	391	433
	100%	16.9%	39.5%	43.7%
Comments	617,782	374,604	170,610	72,568
	100%	60.6%	27.6%	11.7%
Tokens (million)	21.9	14.5	3.9	3.5
	100%	66.2%	17.8%	16%
Troll comments	4,477	1,738	900	1,839
	100%	38.8%	20.1%	41.1%
Troll tokens	4,884	1,894	955	2,035
	100%	38.8%	19.6%	41.7%

Table 1: Threads, comments, and tokens in the corpus

The majority of the data comes from GP as 60.6% of the comments and 66.2% of the tokens were published on this blog. However, LL has the most troll comments.

3.2. Data analysis

Data analysis involved a corpus-based qualitative-interpretative analysis of the collected troll comments:

(1) Using the concordance lines of the search term *<tt>*troll*</tt>* in AntConc (Anthony, 2016), I selected and annotated those troll comments from Version 2 in which the assumed motives for trolling were discussed (hereon referred to as 'troll motive comments'). This is illustrated in example (2).

(2) [labourlist_333_21]
The Tories must be *really* ***panicking*** *if* ***they hired A to troll*** *the way he does here. You just can't get decent* ***staff*** *these days.*

(2) I identified the linguistically marked motives that participants attributed to those they called trolls and created a taxonomy from them.

(3) I described how the different linguistically marked motives affect the discursive construction of trolling and trolls in the comments.

(4) To determine how often the participants explicitly attribute the identified motives to the alleged trolls, I used the motives as descriptive categories and provided each troll motive comment with motive-related annotations.

(5) To make this discursive-pragmatic annotation process more transparent and systematic, I studied the n-grams and collocates of the search term *<tt>*troll*</tt>* in Version 2 and the positive keywords in Version 3 against Version 4 as a reference corpus using AntConc. (Settings for n-grams: search term: both on the left and on the right, cluster size: between 2 and 6, min frequency: 5 and min range 1. Settings for collocates: window span: 5L5R, statistic: Mutual Information (MI), min MI score: 3.0, min frequency: 5. Settings for keywords: keyness statistic: log-likelihood (LL), min LL score: 3.84, min frequency: 5.) The aim of this step was to identify those words and expressions that mark a motive for trolling on their own.

(6) I summarised the quantitative results of the annotation.

4. Results

4.1. A taxonomy of the motives attributed to trolls

2,037 troll motive comments were identified in the corpus. 866 in GP, 279 in GF, and 892 in LL threads. Five motives for trolling emerged during the analysis of these comments: (1) various emotional mental health-related/social reasons, (2) financial gain, (3) unspecified political affiliation, (4) political beliefs, and (5) being employed by a political body.

The first motive covers various, often inter-related emotional states (e.g. boredom, loneliness or enjoyment), mental health issues, such as OCD, and social deprivation as reasons for trolling. When users suggest this motive, trolling is constructed as an emotionally motivated individual behaviour and trolls are portrayed as miserable individuals with emotional, mental health-related, and social problems.

(3) [guido_40_308]
No wonder A keeps ***trolling*** *here. He must be* ***bored*** *witless.*

The second motive refers to those cases where users imply that others are trolling because they are paid for it. However, it is not mentioned who pays the trolls and why. Here, trolling is constructed as a financially motivated individual activity and trolls are represented as rational but immoral and dishonest individuals.

(4) [guardian_48_3718]
He/she might be an individual expressing their own opinion, legitimate in a democracy whether you or I agree with it. Whereas you could be described as a ***paid troll****.*

The third motive represents those comments where users indicate that others are trolling due to their political affiliation. However, it remains unspecified whether the trolls merely support a political body or they work for it. Thus, the way trolling and trolls are constructed in these comments is ambiguous.

(5) [labourlist_432_1761]
Tory troll *hanging around Labour sites. Why?*

The fourth motive stands for those occasions when users imply that others are trolling since they support a political party or an ideology. Thus, trolling is constructed as an ideologically motivated individual activity and trolls are depicted as irrational political fanatics.

(6) [guido_90_573]
FFS we have an unusually high number of stupid ***socialist trolls*** *in this thread. Are they seriously trying to tell us that Bin Laden wasn't a murderous butcher who had declared war on the western world? Keep it up you* ***lefty trolls*** *so everyone realises how vile and stupid you are.*

The fifth motive is that certain users are trolling because a political body, namely a British political party, another country (Russia or Israel) or the European Union employs them and has ordered them to do so.

(7) [guardian_129_6462]
Nice ***trolling*** *from a* ***Tory Party Central Office intern****.* ***Hopefully****, come the 11th,* ***you'll be*** *signing on as* ***unemployed****.*

It is also repeatedly suggested that as part of their employment, these political bodies (5a) send the trolls to these blogs, (5b) tell them how to troll, (5c) sponsor their trolling and (5d) train them. Consequently, trolling is constructed as a financially and politically motivated and centrally organised collective activity while trolls are portrayed as unskilled and low-paid employees of low prestige who simply follow orders but do not necessarily support the political body that employs them.

4.2. Annotation of the Troll Motive Comments

The above-presented motives were used as descriptive categories to annotate the 2,037 troll motive comments in the corpus. Table 2 displays the n-grams and collocates of the troll tokens and the keywords in the troll comments that were used to make the annotation process more consistent.

N-gram	Collocate	Keyword	Motive tag
paid troll(s)	paid	paid	2/5c
–	pay(ing)	–	2/5c
–	sponsored	sponsored	2/5c
–	funded	–	2/5c
Tory troll(er)(s)	Tory	Tory, torytroll	3/4/5
trolling Tory	Tory	Tory	3/4/5
	conservative	–	3/4/5
Labour troll(s)	Labour	–	3/4/5
–	Corbynista(s), corbynite	–	3/4/5
–	Corbytrolls	Corbytroll(s)	3/4/5
–	Blairite	–	3/4/5
establishment troll	establishment	establishment	3/4/5
–	liblabcon	–	3/4/5
UKIP troll(er)(s)	UKIP	–	3/4/5
	kipper	cyberkipper	3/4/5
Green (Party) troll	Green	Green	3/4/5
SNP troll	–	SNP	3/4/5
–	BNP	–	3/4/5
EU troll	EU	EU	3/4/5
right(-)wing troll(s)	right(-)wing	–	4
left(-)wing trolls	–	–	4
lefty/leftie troll	lefty, leftie(s)	leftie	4
leftard troll	leftard	leftard	4
Central Office Troll(s)	Central, office	central	5
CCHQ troll(s)	CCHQ, HQ	CCHQ, HQ	5
–	Lynton	Lynton	5
–	employed	–	5
troll army	army	–	5
–	Kremlin	Kremlin	5
Hasbara troll	Hasbara	Hasbara	5

Table 2: The n-grams, collocates and keywords marking a motive attributed to trolls

Table 3 presents the proportion of those troll motive comments that were provided with a particular motivation-related tag. Note that as one comment could receive multiple tags, the sum of the percentages in the same column is not necessarily 100%.

Motive	Tag	Overall	GP	GF	LL
Emotional reasons	1	5.9%	6.8%	10%	3.8%
Financial gain	2	1.9%	2.3%	5.4%	0.3%
Unspecified political affiliation	3	65.3%	56.2%	38%	82.7%
Political beliefs	4	12.3%	15.1%	16.5%	8.2%
Being employed by a political body (PB)	5	17.7%	24.1%	33.7%	6.5%
Being sent by a PB to troll	5a	0.8%	1.3%	0.7%	0.3%
Being told by a PB how to troll	5b	1.3%	2.2%	1.4%	0.4%
Being paid by a PB to troll	5c	5.3%	6%	16.5%	1%
Being trained by a PB for trolling	5d	0.3%	0.2%	0.7%	0.2%

Table 3: The proportion of troll motive comments provided with a particular motivation-related tag

The results demonstrate that the most prevalent linguistically marked motive for trolling is an unspecified political affiliation, which is followed by being employed by a political body, and political beliefs. Meanwhile, emotional/mental health-related and social reasons as motives ascribed to trolls only occur in 5.9% of the troll motive comments.

The most striking difference in the distribution of the motives attributed to trolls between the three blogs is that unspecified political affiliation is much more prevalent whereas being employed by a political body is considerably less frequent on LL than on GP or GF. This is because there was a single commenter on LL who frequently used the expression *Tory troll* and consequently, his/her comments were provided with the unspecified political affiliation motive tag.

This shows that since only a small minority of the commenters call others trolls, the individual habits of those who do so can have a major impact on the general distribution of the motives on a blog. Thus, the quantitative differences between the blogs can be better explained by these context-dependent individual practices than by abstract variables, such as the political position of the blogs.

5. Conclusions

The main conclusions of this study are as follows:

(1) Although the relevant academic literature regards trolling as a chiefly emotionally motivated behaviour, in the context of online political discourse, participants attribute other motives to trolls as well, including financial gain, unspecified political affiliation, political beliefs, and being employed by a political body.

(2) In the examined corpus of comment threads from British political blogs, an unspecified political affiliation, being employed by a political body and political beliefs are more frequently mentioned motives for trolling than emotional reasons.

(3) A local conspiracy theory has been developed around trolling on the investigated blogs as some participants repeatedly suggest that various British political parties, other countries or the European Union secretly employ trolls. Thus, trolling is perceived as part of the online political warfare, a means that is believed to be used to manipulate public opinion.

(4) Whilst the concept of trolling can be constructed in different ways in the analysed troll motive comments, a

common trait of these comments is that the alleged trolls are portrayed in a strongly negative manner. Thus, when participants call others troll, they do not only attribute motives to the trolls to explain their behaviour but also to insult them.

6. References

Anthony, L. (2016). *AntConc* (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Baker, P. (2001). Moral Panic and Alternative Identity Construction in Usenet. *Journal of Computer-mediated Communication*, (7)1.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Binns, A. (2012). "Don't Feed the Trolls! Managing Troublemakers in Magazines' Online Communities." *Journalism Practice*, 6(4), pp. 547–562.

Culpeper, J. & Hardaker, C. (2016). Pragmatics. In P. Baker & J. Egbert (Eds.), *Triangulating Methodological Approaches in Corpus-linguistic Research*. New York/London: Routledge, pp. 124–137.

Donath, J.S. (1999). Identity and Deception in the Virtual Community. In P. Kollock & M.A. Smith (Eds.), *Communities in Cyberspace*. London/New York: Routledge, pp. 27–58.

Fichman, P. & Sanfilippo, M.R. (2015). The Bad Boys and Girls of Cyberspace: How Gender and Context Impact Perception of and Reaction to Trolling. *Social Science Computer Review*, 33(2), pp. 163–180.

Galán-García, P. et al. (2014). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. In Á. Herrero et al. (Eds.), *Proceedings of International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*. New York: Springer, pp. 419–428.

Hardaker, C. (2010). Trolling in Asynchronous Computer-mediated Communication: From User Discussions to Academic Definitions. *Journal of Politeness Research. Language, Behaviour, Culture*, 6(2), pp. 215–242.

Hardaker, C. (2013). "Uh….not to be nitpicky,,,,,but...the past tense of drag is dragged, not drug.". An Overview of Trolling Strategies. *Journal of Language Aggression and Conflict*, 1(1), pp. 58–86.

Herring, S.C. et al. (2002). Searching for Safety Online: Managing „Trolling" in a Feminist Forum. *The Information Society*, 18(5), pp. 371–384.

Herring, S.C., Stein, D., & Virtanen, T. (Eds). (2013). *Pragmatics of Computer-Mediated Communication*. Berlin/New York: Mouton De Gruyter.

Hopkinson, C. (2013). Trolling in Online Discussions: From Provocation to Community-building. *Brno Studies in English*, 39(1), pp. 5–25.

Morrissey, L. (2010). Trolling Is an Art: Towards a Schematic Classification of Intention in Internet Trolling. *Griffith Working Papers in Pragmatics and Intercultural Communications*, 3(2), pp. 75–82.

Özsoy, D. (2015). Tweeting Political Fear: Trolls in Turkey. *Journal of History School*, 8(22), pp. 535–552.

Shachaf, P. & Hara, N. (2010). Beyond Vandalism: Wikipedia Trolls. *Journal of Information Science*, 36(3), pp. 357–370.

Tanskanen, S.-K. (2007). Metapragmatic Utterances in Computer-mediated Interactions. In A. Hübler & W. Bublitz (Eds.), *Metapragmatics in Use*. Amsterdam: John Benjamins, pp. 87–106.

Utz, S. (2005). Types of Deception and Underlying Motivation. What People Think. *Social Science Computer Review*, 23(1), pp. 49–56.

Fear and Loathing on Twitter: Attitudes towards Language

Damjan Popič,* Darja Fišer*†
*Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
†"Jožef Stefan" Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
E-mail: damjan.popic@ff.uni-lj.si, darja.fiser@ff.uni-lj.si

Abstract

The paper deals with the sociolinguistic concept of prestige imbued in the notion of standard language, and the social status connected to the inherent language skill (or lack thereof). To this end, we analyse Slovenian tweets pertaining to language use and the (in-)correctness of other users' use of language, propose a typology, especially in cases where language use is used as an argument against someone's qualifications or beliefs.

Keywords: orthography, (linguistic) prestige, computer-mediated communication

1. Introduction

The paper deals with a corpus-enhanced sociolinguistic analysis of attitudes towards language of Slovenian users on Twitter. In the Slovenian culture, the ability to produce linguistically correct texts has long been a test of a person's sophistication, education, and social class. This stance was further entrenched by the purist efforts, perpetuated by the country's constant struggle against imperialist forces, and the notion persists to this day.[1]

For this reason, we analyse the attitudes of Slovenian Twitter users towards language, and the types of discourse that language-related debates feature in. By using the keywords *comma*, *orthography*, *grammar*, *Slovene*, and *language* we isolate tweets in which users comment on language from the corpus JANES-Tweet v1.0 (Fišer et al. 2017). We analyse these examples from a sociolinguistic standpoint of prestige. This kind of attitude towards language is very common in linguistically conservative communities, and is often accompanied by a shared belief that the attitude towards language is a moral issue. This is visible in language use as (supposed) improper use of language is often used as an *ad hominem* argument aiming to suppress the relevance of that person's beliefs (on any matter).

Furthermore, recent developments have shed a new light on the attitudes towards language proficiency on social media all over the world, especially with Donald Trump. His tweets have been subject to much scrutiny, linguistic and otherwise.[2] This shows that the language on Twitter, although supposedly informal in nature, is very much considered to be public.

[1] This is perhaps best described by the latest development in which the Slovenian Parliament voted on amending the War Grave Act (with 54 ayes and 6 nays). Following the parliamentary conclusion, a (superfluous) comma will be removed from the first stanza of Oton Župančič's poem that will be inscribed on the memorial of national reconciliation.

[2] See a comprehensive R-based statistical analysis of his tweets in relation to the device used for posting (Robinson 2016; http://varianceexplained.org/r/trump-tweets/).

2. Background

The Slovenian normative language tradition has been built around the German syntactic system and is highly prescriptive in nature. This means that the syntax in Slovene is highly structured, and this combination of a highly complex syntax and strict language rules imposed by the normative language guide (i.e. *pravopis*, from Ger. *Rechtsschreibung*) makes it notoriously difficult to write "correctly." For instance, the dreaded comma is the bane of students' lives, with many recent studies showing that Slovenes in general are underperforming severely when it comes to commas (Popič et al. 2016).

3. Analysis

In this section, we present the results of the analysis of tweets containing at least one of the selected keywords: *comma*, *orthography*, *grammar*, *Slovene*, and *language*. The analysis involved a manual content analysis of the entire extracted sub-corpus and drafting a typology that allows us to cover the attitudes towards language imbued in the extracted tweets, for all the keywords. As the following tables indicate, the attitudes differ in number and tenor, however, scathing and disparaging tweets are present in each of the categories. Table 1 gives the frequencies of the selected keywords in the corpus.

Keyword	Frequency
vejica (the comma)	1,978
slovnica (grammar)	930
pravopis (orthography)	432
jezik (language)	11,202
slovenščina (Slovene)	4,607
Total	**19,149**

Table 1: Keyword frequencies.

As Table 1 demonstrates, by far the most frequent of the five keywords is *language*, with *Slovene* being second. However, the expression *language* is relatively general and also used in other senses (as in 'tongue' in the physical sense, in phrasemes, etc.). Considering the much narrower meaning of *vejica* (it almost exclusively

pertains to the comma as a means of punctuation), it is evident that the comma plays an immensely significant role with 1978 occurrences. For this reason, we devote the most attention to this keyword and deal with the rest of the keywords in pairs, focusing above all on the differences in the attitudes between the keywords.

3.1 *Vejica* 'the comma'

The comma holds the most prestigious place in the Slovenian linguistic tradition, and performs several functions in discourse. Considering the examples extracted from our corpus, we can categorize the extracted examples in the categories given in Table 2 (including a free and standardised translation for each).

Attitude	Example
inquisitive	@strankaDL Zakaj je v ministrovi izjavi vejica? ;-) Šoltes: Aktualno politico *Why is there a comma in the minister's statement?*
informative	@drVinkoGorenak Spoštovani, manjka vam vejica za prvim ne. Prav je: "Ne, ne *Dear Sir, you are missing a comma following the first 'ne'. It should read: "Ne, ne*
lamenting	nekoc v eseju s "PS" dodal deset vejic in pripisal vstavi po potrebi *I once put ten commas in the postscript in an essay, with a comment: "Insert where appropriate."*
jocular	Če bi Krpan imel vejico v žepu, bi jo gotovo postavil *If Krpan* [a Slovenian folk hero] *had a comma in his pocket, he'd certainly use it.*
dismissive	@finance_si @NovaSlovenija Še vejice porihtajte. Da ne boste kot Pojbič *Get your commas in order. So you aren't like Pojbič…*
defensive	čaki mal.. A zarad vejic je pa človk nepismen?? :-) #svašta *Wait a second… So* [the misuse of] *commas make*[s] *you illiterate? #whatever*
apologetic	tudi jaz bi napisal z vejicami, če bi mi omejitev 140 znakov *I'd use commas if I had more than 140 characters*
idiomatic	Najboljši članek! Podpišem vse do vejice in pike! Desnica naj zakoplje *The best article ever! I agree with every comma and full stop!*

Table 2: Tweets relating to the comma.

As Table 2 demonstrates, we have identified eight different attitudes towards the comma in our dataset. The inquisitive attitude pertains to actual questions regarding the use of the comma. The second, i.e. the informative attitude, is complimentary to the inquisitive one as it provides explanation(s) on the use of the comma and/or answers to questions on the use of the comma. The example we provided above includes a detailed (and quite possibly condescending) explanation directed at a Slovenian MP explaining why his use of the comma in his tweet was incorrect.

The lamenting attitude covers examples of exasperation over the perceived difficulty over the use of the comma. The example above laments about the school experience of a former pupil who asked his teacher to insert commas for him if required. The jocular attitude covers examples that apply irony on the Slovenian situation and the obsession with commas, whereas dismissive tweets use (alleged) misused commas to portray a person or institution as incompetent in general. For instance, in our example in Table 2, a person directs a comment at the Twitter account of the Slovenian conservative party NSi, saying that they should get their writing in order lest they should "go full Pojbič", referring to a Slovenian MP famous for his poor grammar and spelling in his tweets. It is of note that the Slovenian politicians (especially conservative) and official institutions are under constant scrutiny when it comes to language (in social media), especially regarding the comma.

On the other hand, the apologetic and defensive attitudes aim to explain or justify one's "transgressions" with the comma, each in its own way. While the apologetic approach involves providing reasons for a particular example of comma misuse (either not knowing the rules or more pragmatic excuses like haste, typos, etc.), the defensive attitude conveys a stronger reaction, either against the user(s) exposing the misuse or against the relevance of the comma itself (normally in contrast to the meaning). The final category involves examples containing idiomatic references to the comma, i.e. using the comma as a metonymic expression for 'language' or 'text'.

3.2 *Slovenščina* 'Slovene' and *jezik* 'language'

As both keywords display very similar attitudes, we deal with them in a single section, with the obvious intention of displaying the attitudes relating to the "mother tongue". A classification of tweets containing the two keywords is given in Table 3.

The tweets containing *Slovene* and *language* can be classified in five categories. The category General covers tweets providing general information and inquiries on both keywords. It is interesting to note that the example containing a purist tweet is again directed at the conservative MP Vinko Gorenak, who faced criticism, as recorded in Table 2, for his comma transgressions. This time he is the target of two purist comments, one lambasting him for improper grammar, and one for using a word that does not exist in Slovene on TV.

Attitude	*Jezik* 'language'	*Slovenščina* 'Slovene'
General/ inquisitive	A je slovenščina res najtežji jezik na svetu? *Is Slovene really the hardest language to learn in the world?*	Kako prevedemo˝small talk˝v slovenščino ? *How do you say "small talk" in Slovene?*
nationalist	s slovenščino, našim maternim jezikom, delamo kot svinja z mehom. *We treat Slovene, our mother tongue, like dirt.*	UPORABLJA ISKLJUČNO SLOVENSKE BESEDE. Slovenščina to smo SLOVENCI,ALI NIMAMO SVOJIH *Use only Slovene words, we are Slovenes after all.*
purist	@drVinkoGorenak Prosim za podnapise v slo jeziku: kaj pomeni "paložek"? @PlanetTV *Subtitles please: what does "paložek" mean in Slovene?*	@drVinkoGorenak "bodite točna" v slovenščini ne obstaja. Če že vikate, vikajte *"Bodite točna"* [i.e. the mixing of plural and singular forms] *doesn't exist in Slovene. If you mean to address someone formally, stick to it.*
jocular	casov " the day after yesterday". Jeziki ji ful laufajo *"The day after yesterday." She's great with languages.*	Ha odkar delam s kamionarji, mi slovenscina neki sepa. *Ever since I started working with truckers, my Slovene has deteriorated.*
idiomatic	vidva takorekoč govorita isti jezik. In kaj bi bilo *You speak the same language, as it were.*	vidim, da si bil zgolj v službi slovenščine. ČE ČM! *I see you were only in Slovene's service.*

Table 3: Tweets relating to language and Slovene

The tweets that employ a jocular perspective either expose the difficulties people have with Slovene or are meant as a ridicule of the Slovenian hard-line stance on language, whereas, in idiomatic tweets, the keyword *language* is very common, and *Slovene* much less so. The most significant difference in attitudes towards the comma and towards language/Slovene lies in the generality – the tweets belonging to the latter category are much more general (as well as less vicious and dismissive), but at the same time more nationalist.

3.3 *Pravopis* 'orthography' and *slovnica* 'grammar'

As in section 3.2, we combine the related keywords into a single analysis. This is especially relevant because most people confuse or substitute orthography and grammar, the latter most often being an umbrella term for all language-related matters, but in the Slovenian linguistic tradition, the two are distinct entities.

Attitude	*slovnica* 'grammar	*pravopis* 'orthography'
purist	slovnica pa pravopis 1,vsebina 5 *You get an F for grammar and orthography, and an A for content.*	Vzgled? Pravopis pod vzglavnik, nekompetentnež *"Vzgled"?* [an ungrammatical form of *zgled] Put the orthography under your pillow, you incompetent…*
apologetic	domače. Tle ni prostora za visoko slovnico. Če ne morem na #hodok, sem pa *This is no place for elevated grammar.*	tw je za vsebino, ne pravopis :) mene je bolj strah *Twitter is meant for content, not orthography.*
informative	slavisti. Ali pač? Po slovenski slovnici je desni prilastek pri kraticah *The Slovenian grammar on modifiers with acronyms says that…*	osebno se tudi zdi tako. Ampak pravopis je stvar konsenza. Greste na *It may seem like that but orthography is all about consensus.*
inquisitive	po ;-) Saj menda "greve" ni po slovnici? #samvprašam #slovnica *Surely "greve" is ungrammatical?*	inovatorke. Ali cestarji obvladajo pravopis? via *Do road workers know orthography?*
idiomatic	Lukšič: Ko gre za slovnico političnega delovanja, je treba *When it comes to the grammar of political action…*	le za demokracija, črn dan za pravopis! #krivopisje #domoljupi *Not just democracy, today is a sad day for orthography.*
nationalist	slabo znanje slovnice se imenuje "čefurščina" ;) *Poor grammar is indicative of čefurščina* [a pejorative term for the dialect of Slovene spoken by immigrants from the countries of the former Yugoslavia]	slovenski narod, niti osnovnega pravopisa ne pozna. Kak veš, da si v *Alas, the nation of Slovenia – they don't even know the basic orthography.*
jocular	uvede policijska ura in odpravi slovnica. Just sat through Simeon ten *Let's implement curfew and abolish grammar altogether.*	crknu!?! Kako se reče po noven slo pravopisu beseda KLON? PODOBNIK ! *What is a clone called in Slovene? It's PODOBNIK* [a surname of two Slovenian MPs, as well as a very literary word for a lookalike in Slovene].

Table 4: Tweets relating to grammar and orthography

The attitudes contained in the tweets relating to orthography and grammar are given in Table 4. The attitudes in tweets pertaining to orthography and grammar can be categorized in 7 different categories. As can be seen in our purist examples, users on Twitter employ a rather harsh tone when it comes to these keywords. The same seems to hold true for nationalist tweets, which, as in our examples, usually employ a hard-line stance or one of a exasperation over the use of language of people writing in Slovene.

With apologetic tweets, we can see an interesting trend that users themselves, upon being faced with accusations of having "poor grammar", are trying to discredit Twitter as a means of "formal" communication and instead establish it as a means of fast, direct, and informal communication. The concepts of "orthography" and "grammar" are thus used as synonyms for "form", whilst the significance of "meaning" is pointed out.

4. Conclusion

In this paper, we were interested in discovering whether or not the stereotypical attitudes towards written Slovene have been preserved and transformed in computer-mediated communication, in our case on Twitter. The analysis shows that, in spite of the belief that informality is an integral part of new media, all traditional notions pertaining to Slovene are still very much alive in the digital age. What is more, some forms of prejudice seem to be flourishing online due to the lack of oversight. For instance, personal attacks with dismissive tweets aiming to destroy one's credibility are very common, especially in regard to the comma, whereas more general topics attract more general attitudes, but just as vicious (purist and nationalist).

5. 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).

6. References

Fišer, D., Erjavec, T., Ljubešić, N. 2017. The compilation, processing and analysis of the Janes corpus of Slovene user-generated content. Wigham, C.R. & Ledegen, G. (ur.). *Corpus de communication médiée par les réseaux: construction, structuration, analyse*. Collection Humanités Numériques. Paris: L'Harmattan (in print).

Popič, D., Fišer, D., Zupan, K., Logar, P. (2016). Raba vejice v uporabniških spletnih vsebinah. *Proceedings of the Language Technologies and Digital Humanities Conference*. Ljubljana, Slovenia: 149–153.

Robinson, D. (2016). Text analysis of Trump's tweets confirms he writes only the (angrier) Android half. (http://varianceexplained.org/r/trump-tweet

A Comparative Study of Computer-mediated and Spoken Conversations from Pakistani and U.S. English using Multidimensional Analysis

Muhammad Shakir, Dagmar Deuber

Department of English, University of Münster
Johannisstr. 12 - 20, 48143 Münster, Germany
muhammad.shakir@uni-muenster.de, deuber@uni-muenster.de

Abstract

The present study compares four computer-mediated conversational registers (comments, Facebook (FB) groups, FB status updates and tweets), and spoken conversations from Pakistani and U.S. English using Biber's Multidimensional Analysis framework on three dimensions of variation, i.e. (i) Interactive versus Descriptive Explanatory Discourse, (ii) Expression of Stance, and (iii) Informational Focus versus 1st Person Narrative. Spoken conversations have a high score on dimension 2, while CM conversations show register and regional variation on dimension 1 and 3. FB groups are significantly different between both regional varieties, followed by FB status updates, comments and tweets. Pakistani FB groups discuss self-help related topics, and appear to be slightly interactive and highly informational, while the U.S. ones are interactive and narrative discussing community and political issues. Pakistani FB status updates and tweets use English mainly for informational purposes, while the U.S. counterparts have an interactive and personal orientation indicating a wider functional role of English.

Keywords: Register Variation, Multidimensional Analysis, World Englishes, Conversations

1. Introduction

Language users converse with each other to exchange news, views and ideas in an informal way (Oxford Online Dictionary, 2017). Traditionally conversations have been spoken only. With the advent of the internet, another medium has been added, i.e. the written medium. Spoken and newly emerging computer-mediated (CM) conversations are different in ways like turn-taking (Herring, 2011) or synchronicity (Bieswanger, 2016), but at the same time they are linguistically similar to each other (Jonsson, 2015). Though extensively studied, CM conversations need to be studied using a comparative and multi-dimensional approach (Herring, 2011) like Biber's (1988) Multidimensional Analysis (MDA) model, which combines an analysis of situational context with lexico-grammatical features, and interprets them functionally (Biber & Conrad, 2009). Present research paper aims to study emerging CM registers – i.e. comments, Facebook (FB) groups, FB status updates and tweets – in relation to spoken conversations. MDA studies on CM registers have, until now, largely focused either on U.S. English (Grieve et al., 2010) or native varieties of English (Biber and Egbert, 2016). Pakistani English is an outer circle variety in Kachru's (1992) three circle model, which is an important tool in the linguistic repertoire of Pakistani internet users, but not widely studied in relation to the internet. On the other hand, U.S. English, is an inner circle and globally dominant variety, which may be influencing varieties like Pakistani English due to contact and technological influence on the internet. Previous research (e.g. Hardy and Friginal, 2012) suggests that there might be differences between inner and outer circle varieties of English regarding the use of CM registers. Hence further aim is to combine the study of register and regional variation.

1.1 Previous Research

Spoken conversations are generally involved and interactive (Biber, 1988). However, later studies also show additional dimensions like narrativity, informational focus (Biber, 2004), and expression of stance (Biber, 2006). Various types of CM conversations have been studied using MDA. Collot and Belmore (1996) applied Biber's (1988) MDA to study bulletin boards – an ancestor of today's FB groups – and found them nearer to public interviews in spoken conversations. FB status updates and tweets have been said to be CM equivalents of spoken conversations but quite different (Sardinha, 2014), and to be highly informational and descriptive instead of being involved and interactive (Titak & Roberson, 2013). Similarly, comments have been found to be involved, personal and past oriented (ibid). Lastly, studies using MDA to find out regional variation, e.g. Xiao (2009) and Coats (2016), do not involve the comparison of CM and spoken conversations.

2. Material and Methods

Categories	Pakistani English		U.S. English	
	Words	Texts	Words	Texts
Comments	334,447	794	342,517	747
FB groups	163,940	502	163,158	426
FB S. U.	67,737	104	68,819	108
Tweets	58,771	115	62,086	103
Conv.	158,521	85	487,476	111

Table 1: Description of the corpus

Table 1 describes the data for both varieties. Four registers were selected for CM conversations as they were publicly accessible on the internet. Comments were

collected from various blogs (single- and multi-writer blogs, newspaper blog posts, and technology blogs) using the website downloader software DarcyRipper and a custom software written in C#. The data for FB groups was manually copied and cleaned after identifying groups originating from Pakistan (mostly closed groups) and the U.S.A. (mostly open groups). Status updates were also manually collected by identifying user profiles from member lists of already scraped groups. Twitter profiles were identified from real-time tweets originating from Pakistan and the U.S.A. The tweets were downloaded using a custom software. The CM conversations data was reviewed and edited for spam, automatic messages, Roman Urdu code switching, and non-standard spellings to facilitate the tagger. However, the spoken data was not reviewed. The data for Pakistani English was extracted from an under-development corpus of Pakistani English (ICE-PK), and for the U.S. variety from the Corpus of Contemporary American English (COCA). The entire data was collected for the time period of 2009-15.

The data was then tagged using Biber Tagger (Biber, 1988; 2006). The tagger tags only approximately 140 specific lexico-grammatical features. Hardy and Friginal (2012) have reported up to 93% accuracy of Biber Tagger on blog posts. Though desirable, such a manual verification of tagging accuracy was beyond the scope of the present research. A new MDA was conducted following guidelines provided in Biber and Gray (2013) and Egbert and Staples (*forthcoming*). The statistical software package R was used to perform Exploratory Factor Analysis. 61 lexico-grammatical features were selected by studying previous research on conversational registers (Biber, 2004; Titak & Roberson, 2013; Biber et al., 1999). After conducting multiple factor analyses with factor solutions from 2-7, a 3-factor solution with Principal Axis Factoring as factor extraction method and Promax as rotation method was deemed fit to describe the data. The details and descriptive statistics are provided in table 2.

Factor	+/-	Linguistic Features with Loadings
1	+	present tense 0.70, 2nd person pronouns 0.49, contractions 0.38, activity verbs 0.34, models of prediction 0.29, models of possibility 0.27 (1st person pronouns 0.34)
	-	prepositions -0.40, attributive adjectives -0.38, nominalisations -.35 (word length -0.29)
2	+	*that* deletion 0.57, mental verbs 0.49, *that* clauses controlled by verbs 0.48, *that* clauses controlled by communication verbs 0.45, communication verbs 0.43, *that* clauses controlled by factive verbs 0.40, *that* clauses controlled by likelihood verbs 0.40 (communication verbs in other contexts 0.30)
	-	(common nouns -0.38)
3	+	word length 0.53, common nouns 0.51, communication verbs in other contexts 0.40, process nouns 0.34, abstract nouns 0.27 (communication verbs 0.38)
	-	1st person pronouns -0.38, adverbs of place -0.33, general adverbs -0.32, past tense -0.29 (contractions -0.37), (nominalisations -0.25)
Other Descriptive Statistics		
Total Variance Explained		22%
Variables in final FA		25
Cut-off		+/- 0.25
Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy		0.54 (Classification: Miserable)

Table 2: Results of Factor Analysis and other descriptive statistics

The total variance explained and KMO values of the factor analysis are quite low. As Egbert and Staples (*forthcoming*) analysed in their study, total variance explained values have been generally lower for MDA studies. This is especially the case with internet-based registers. Similarly, they also reported KMO value less than .60 for one of their previous studies. A possible reason might be that the lexico-grammatical variables depend on other variables not included in present analysis. Heterogeneity of the data could be another possible reason.

Each factor in the solution has feature groups with positive and negative loadings, which are mutually less likely to co-occur (Biber, 1988). High factor loading indicates the feature is salient, and vice versa. The features within brackets overlapped with the ones in other factors, hence they were used in the interpretation of factors to dimensions, but not for dimension score calculation. The dimension scores were calculated for each text by summing z-scores of positive as well as negative features, and finally by subtracting negative total score from positive total score. The mean scores for

each register category were also calculated to compare the registers on each dimension. Parametric (One-way) or non-parametric (Kruskal-Wallis) ANOVA and respective post-hoc tests were used to check if corresponding registers had significant differences between the regional varieties.

3. Results and Discussion

3.1 Dimension 1: 'Interactive versus Descriptive Explanatory Discourse'

Dimension 1 has eleven features with 7 on positive and 4 on negative side. The features on positive side belong to an interactive discourse (Grieve et al., 2010). The less important features, such as possibility and prediction modals combine with human subjects and dynamic verbs to denote to intrinsic meaning (Biber et al., 1999). The texts with a high positive score are from FB groups. They discuss about present and future events, and are highly interactive.

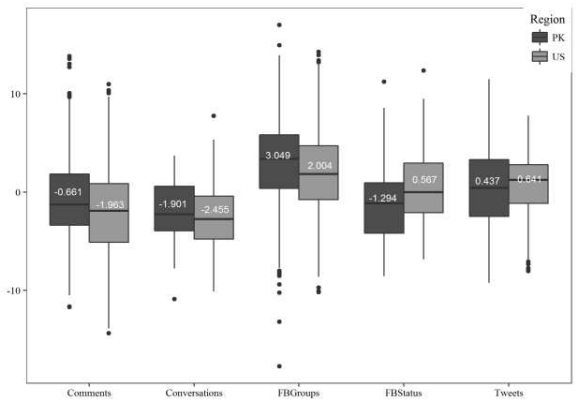


Figure 1: Conversational registers on Dimension 1: 'Interactive versus Descriptive Explanatory Discourse' (One-way ANOVA: $F(9, 3085) = 70.23$, $p < .001$; Post-hoc Tukey HSD significant groups between varieties: Comments, FB groups, FB status updates)

The features on negative side are prepositions, attributive adjectives, and nominalisations. Attributive adjectives indicate the presence of descriptive discourses. Prepositional phrases are "the most common type of post-modifiers" (Biber et al, 1999, p. 631). A look at high scoring texts from comments and FB groups show that the texts are descriptive and explanatory in general. Thus, combining positive and negative features the dimension 1 can be interpreted as 'Interactive versus Descriptive Explanatory Discourse'.

Figure 1 shows comparison of register categories on dimension 1. Comments, FB groups and FB status updates in Pakistani English are significantly different from their counterparts in U.S. English, while tweets and conversations do not have significant differences. FB groups is by far the most interactive register, while the category of conversations has the highest inclination towards descriptive and explanatory side of the dimension.

3.2 Dimension 2: 'Expression of Stance'

Dimension 2 has eight linguistic features on positive side, and only one feature on negative side. The positive features include communication verbs like *ask, shout, tell* etc., which show the activity of communication. Mental verbs like *think, know, love, want* etc. are used for cognitive meaning as well as to express attitudes of the speakers (Biber et al., 1999). *That* complement clauses controlled by communication and likelihood verbs are used to convey stance (Biber, 2006), or the presence of reported speech or activities (Titak and Roberson, 2013). An examination of texts with high positive scores from comments, conversations and FB groups show that they contain the elements of opinion or stance. Considering only positive features of this dimension, it can be interpreted as 'Expression of Stance'.

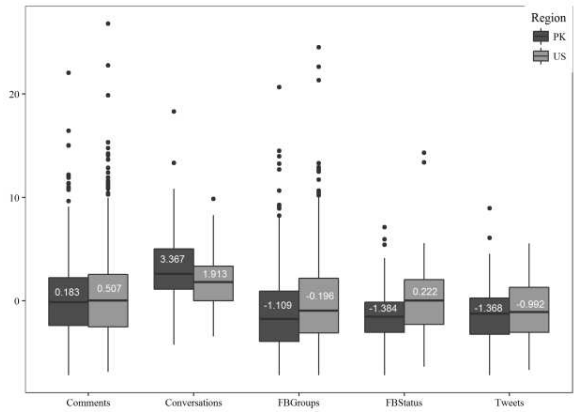


Figure 2: Conversational registers on Dimension 2: 'Expression of Stance' (Kruskal-Wallis ANOVA: $H = 206.0$, $p = 0$; Post-hoc Conover-Iman Test significant groups between varieties: FB groups, FB status updates)

Looking at the results in figure 2, expression of stance seems largely related to spoken conversations, which have the highest scores among all registers. Pakistani conversations have a higher score and a wider range as compared to the U.S. data, which is probably due to a wider variety of conversations (face-to-face, talk shows, interviews etc.). Comments do not have high mean scores, which indicates the possible presence of other stance marking devices like stance adverbs, nouns and

adjectives as observed by Biber (2006, p. 92). FB groups, status updates and tweets are less stance oriented, though both FB related registers show significant differences between Pakistani and U.S. English.

3.3 Dimension 3: 'Informational Focus versus 1st Person Narrative'

Dimension 3 contains twelve linguistic features with six features on either side. The positive features include various kinds of nouns and word length, which generally have a positive correlation with each other, i.e. a higher frequency of nouns indicates lengthier words. The majority of texts with high positive score on this dimension are from Pakistani FB groups, which either contain job ads followed by infrequent formulaic comments, or discussions related to study that include abstract and process nouns.

Among features on negative side, 1st person pronoun and contractions normally occur in informal texts with a personal focus. Past tense verbs have been found relevant to narrative texts (e.g. Biber, 1988). The texts with high negative score are from U.S. FB groups, which generally talk about events with the mention of places in 1st person using past tense. Combining both interpretations, dimension 3 can be labelled as 'Informational Focus versus 1st Person Narrative'.

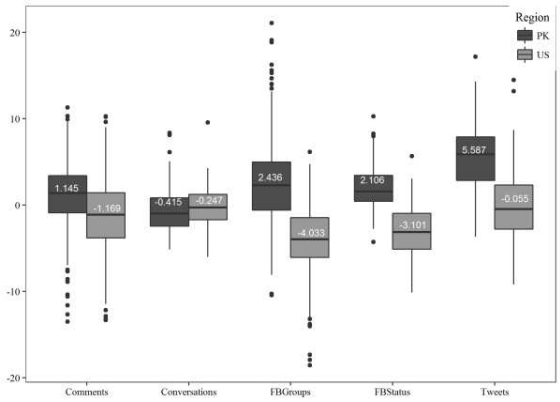


Figure 3: Conversational registers on Dimension 3: 'Informational Focus versus 1st Person Narrative' (Kruskal-Wallis ANOVA: H = 864.43, p = 0; Post-hoc Conover-Iman Test significant groups between varieties: Comments, FB groups, FB status updates, Tweets)

Figure 3 elaborates mean dimension scores of all registers on dimension 3. Though all CM conversational registers show significant differences between Pakistani and U.S. English, spoken conversations do not show much variation except a slight orientation towards the narrative side of the dimension. The most obvious differences are between FB groups, status updates and tweets, where all Pakistani registers have information focused orientation, while U.S. registers incline towards 1st person narration.

4. Conclusion

The differences between spoken and CM conversations mainly appear to be on dimension 1 and dimension 2. On dimension 1, spoken conversations incline towards descriptive and explanatory discourse. The reason for U.S. English seems to be the selection of spoken register for this category, i.e. broadcast discussions, which are different from spontaneous face-to-face conversations. For Pakistani English, though spoken conversations come from more than one registers, for example interviews, talk shows, and student face-to-face conversations, it appears that even face-to-face conversations are generally descriptive and explanatory instead of being involved and interactive. Dimension 2 'Expression of Stance' has been observed previously as well (Biber, 2004; 2006) for spoken conversations. The results apparently confirm that CM conversations are similar but quite different from spoken conversations (Titak & Roberson, 2013). Another possible reason could be the lesser representation of spoken registers in U.S. English and a smaller number of words in Pakistani English.

On the other hand, CM conversations show variation on all dimensions between registers as well as between regional varieties. Pakistani FB groups are generally related to study help, job, pet and game related talk, which makes them interactive as well as information oriented. However, U.S. FB groups are related to politics, community related issues, as well as pet and game related talk, so they are a little less interactive but highly inclined towards personal narration. Pakistani comments are slightly more interactive due to comments from "diary type blogs" (Grieve et al., 2010), while U.S. comments are descriptive in contrast due to an abundance of political "commentary type blogs" (ibid). FB status updates and tweets are highly informational in Pakistani English partially due to the use of local languages to talk about personal issues, while that is not the case with U.S. English. To conclude, Pakistani CM conversations differ from U.S. counterparts, though a more representative data of spoken conversations would help to better understand the relation between both types of conversations.

5. References

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press.

Biber, D. (2004). Conversation text types: A multi-

dimensional analysis. In *Le poids des mots: Proc. of the 7th International Conference on the Statistical Analysis of Textual Data*. Louvain: Presses universitaires de Louvain, pp. 15—34.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers, Vol. 23.* John Benjamins Publishing.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D. & Gray, B. (2013). Identifying multi-dimensional patterns of variation across registers. In M. Krug & J. Schlüter (Eds.), *Research Methods in Language Variation and Change*. Cambridge University Press, pp. 402—420.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Longman Publications Group. (ISBN: 0-582-23725-4.)

Bieswanger, M. (2016). Electronically-mediated Englishes: Synchronicity revisited. In L. Squires (Ed.), *English in Computer-Mediated Communication: Variation, Representation, and Change Vol. 93*. Walter de Gruyter GmbH & Co KG, pp. 281—300.

Coats, S. (2016). Grammatical feature frequencies of English on Twitter in Finland. In L. Squires (Ed.), *English in Computer-Mediated Communication: Variation, Representation, and Change Vol. 93*. Walter de Gruyter GmbH & Co KG, pp 179—209.

Collot, M. & Belmore, N. (1996). Electronic Language: A New Variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives Vol. 39*. John Benjamins Publishing, pp. 13—28.

Conversation. (n.d.). In *Oxford Living Dictionaries*. Retrieved May 15, 2017, from https://en.oxforddictionaries.com/definition/conversation.

Grieve, J., Biber, D., Friginal, E. & Nekrasova, T. (2010). Variation among blogs: A multi-dimensional analysis. In Alexander M., & S. Sharoff (Eds.), *Genres on the Web Computational Models and Empirical Studies*. Springer, pp. 303—322.

Herring, S. C. (2011). Computer-mediated conversation Part II: Introduction and overview. *Language@ internet* 8(2), pp. 1—12. Retrieved from http://www.languageatinternet.org/articles/2011/Herring..

Egbert, J. & Staples, S. (forthcoming). Doing multidimensional analysis in SPSS, SAS and R. In T. B. Sardinha & M. V. Pinto (Eds.), *Multidimensional Analysis*. London: Bloomsbury.

Jonsson, E. (2015). *Conversational Writing-A Multidimensional Study of Synchronous and Supersynchronous Computer-Mediated Communication.* PETER LANG LTD International Academic Publishers.

Kachru, B. B. (1992). *The other tongue: English across cultures.* University of Illinois Press.

Sardinha, T. B. (2014). 25 years later Comparing Internet and pre-Internet registers. In T. B. Sardinha & M. V. Pinto (Eds.),, *Multi-Dimensional Analysis, 25 years on A tribute to Douglas Biber Vol. 60*. John Benjamins Publishing Company, pp. 81—105.

Titak, A. & Roberson, A. (2013). Dimensions of web registers: an exploratory multi-dimensional comparison. *Corpora* 8(2), pp. 235—260.

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

Posters

Contents

MoCoDa 2: Creating a Database and Web Frontend for the Repeated Collection of Mobile Communication (WhatsApp, SMS & Co) 71

Michael Beißwenger, Marcel Fladrich, Wolfgang Imo and Evelyn Ziegler

Public Service News on Facebook: Exploring Journalistic Usage Patterns and Reaction Data 72

Daniel Pfurtscheller

The graphic realization of /l/-vocalization in Swiss German WhatsApp messages 73

Simone Ueberwasser

MoCoDa 2: Creating a Database and Web Frontend for the Repeated Collection of Mobile Communication (WhatsApp, SMS & Co.)

Michael Beißwenger[1], Marcel Fladrich[2], Wolfgang Imo[3], Evelyn Ziegler[1]
[1] University of Duisburg-Essen [2] University of Münster [3] University Halle-Saale
michael.beisswenger@uni-due.de, m.fladrich@uni-muenster.de,
wolfgang.imo@germanistik.uni-halle.de, evelyn.ziegler@uni-due.de

Abstract

The poster reports about intermediate results of *MoCoDa 2*, an ongoing project funded by the Ministry for Innovation, Science, Research and Technology of the German federal state North Rhine-Westphalia in which we are developing a database and web frontend for the repeated, donation-based collection of CMC interactions from smartphone messaging apps like WhatsApp. The database shall serve as a resource not only for quantitative but also for qualitative approaches in the analysis of CMC. MoCoDa 2 builds on experiences from the preceding project *MoCoDa*[1] which has collected a (relatively small) set of 2,198 interactions with 19,161 user posts or ~193,000 tokens since 2012. For MoCoDa 2 the database and web frontend will be re-implemented from the scratch and expanded with additional functions and features:

- A form for donating and editing the data, which involves the donators into the editing and anonymization process and assists them with capturing metadata on the context and topic of the donated sequences as well as on the interlocutors and their social relations. Anonymization will follow an anonymization guideline developed in the CLARIN-D curation project *ChatCorpus2CLARIN*.[2]
- Part-of-speech annotations which comply with the extended 'STTS 2.0' tagset for German CMC[3] and which will be created using a toolchain provided by the Language Technology Lab (LTL) at the University of Duisburg-Essen[4].
- A TEI export for the collected data on basis of the 'CLARIN-D TEI schema for CMC'[5].

Through adopting the STTS 2.0 tagset and a TEI-based export format the corpus data will be interoperable with corpora that are already part of the CLARIN-D corpus infrastructure at the Institute for the German Language (IDS) in Mannheim. To allow for comparative analyses of the MoCoDa 2 data with the discourse found in text corpora and in other CMC corpora, MoCoDa 2 will not only be made available as a standalone resource but also be integrated into the German Reference Corpus (DEREKO) at the IDS Mannheim.

Keywords: corpora, collection strategies, whatsapp, annotation

[1] The MoCoDa project (http://mocoda.spracheinteraktion.de/) is headed by Wolfgang Imo.
[2] Lüngen, Harald; Beißwenger, Michael; Herzberg, Laura; Pichler, Cathrin (2017): Anonymisation of the Dortmund Chat Corpus 2.1. In: Proceedings of the 5th Conference on CMC and Social Media corpora for the Humanities. Bolzano, Oct 02-03, 2017.
[3] The tagset and annotation guidelines can be found at https://sites.google.com/site/empirist2015/home/annotation-guidelines
[4] http://www.ltl.uni-due.de/ (Cooperation partner at the LTL: Torsten Zesch)
[5] http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema

Public Service News on Facebook: Exploring Journalistic Usage Patterns and Reaction Data

Daniel Pfurtscheller
University of Innsbruck
E-Mail: daniel.pfurtscheller@uibk.ac.at

Abstract

As social networking sites have become staples in everyday life an increasing number of people worldwide use social media as a source of news. To reach this audiences, news organizations and public service broadcasters have ventured on services such as Facebook, which in terms of news is by far the most important social networking site in many parts of Europe.

This poster presents an ongoing research project that explores the ways in which public service media from different European countries are delivering news on public Facebook Pages. In a first step, the Facebook Page of the Austrian news magazine "Zeit im Bild" is examined in a pilot study. The poster presents the project as work in progress and gives an overview of the planned corpus building process. The analysis is based on public data gathered from the public Facebook page operated by Austrian national broadcasting agency ORF (http://facebook.com/zeitimbild). The data are extracted using the public Facebook Graph API. The corpus contains all the posts and comments of the Facebook Pages as well as related metadata. No personally-identifiable information is collected.

The social media data are explored using the R software environment to identify and compare journalistic usage patterns and to visualize the interaction of Facebook users. This should provide an overview over the different forms of journalistic news content (i.e. types of posts) and the basic communicative practices that can be observed in the context of the Facebook Pages (i.e. number of comments, shares, likes and other "Reaction" types). To allow deeper insights an exploratory case-study approach is used. Drawing upon media linguistic research the focus is on the micro level of the media texts and their multimodal design. The in-depth analysis aims to characterize different forms of news reporting via Facebook and looks at the different usage of multimodal resources in the context of the Facebook posts and comments. This combination of qualitative and quantitative methods should allow a better understanding of how Facebook is used as a means of news distribution by public service media providers on a large scale and how technical affordances shape the design of news content and follow-up interactions. This knowledge is critical for the discussion of the emerging role of social media in the context of public opinion and political decision-making.

Keywords: Facebook, social media interaction, public service news, metadata, media linguistics

The graphic realization of /l/-vocalization in Swiss German WhatsApp messages

Simone Ueberwasser
University of Zurich

simone.ueberwasser@ds.uzh.ch

Abstract

This Poster represents a corpus based study into /l/-vocalization in the German speaking part of Switzerland. Its main scientific objective is to show that isolated occurrences of /l/-vocalization can be found in speakers who do not originate in the geographical area for which this phenomena has traditionally been described. This result is relevant for the discussion of /l/-vocalization because the study does not look at the realization of the sound in specific words as they are produces in an artificial setting but rather at real data as it is written by informants in everyday WhatsApp communications and as such shows an extended use of the phenomena by a group of speakers hitherto ignored.

/l/-vocalization, i.e. the replacement of a lateral approximant [l] by the vowel [u] is a well described feature of some Swiss German dialects (*GSW*). Phonological (e.g. Haas (1983), who mentions a implicational scale of phonological factors that favor /l/-vocalization) and sociolinguistic (e.g. Christen (2001)) restrictions and promoters are well described. The *Linguistic Atlas of German-Speaking Switzerland* (Sprachatlas der deutschen Schweiz, 1962 to 2003), is the most renowned documentation of the GSW dialects representing regional variation in linguistic features. It describes /l/-vocalization as a feature roughly to be found between Berne and Lucerne. A more recent study (Leemann et al., 2014) found /l/-vocalization to progress to "... southeasterly, southerly, and westerly directions, but with much less success to the north and northwest, where the equally influential dialectal areas of Basel and Zürich seem to exert opposing influences" (Leemann et al., 2014, 191).

A large scale multilingual corpus (617 chats, 5,5 Mio tokens) of authentic WhatsApp messages was compiled in 2014 at the University of Zurich (cf. Ueberwasser and Stark (in print) and www.whatsup-switzerland.ch). The 45 chats from this corpus that are in GSW and for which informants provided information about their home town today and when they were in 5th grade ($\pm$ 12 years old) were used for the study.

Focusing on the places where informants lived when in 5th grade, /l/-vocalization can, of course, be found in the expected area but also around Zurich and Basel in suburban places. However, in these areas, /l/-vocalization is not applied as consequently as in the core area but very sporadically. The tokens that are realized with /l/-vocalization follow different phonological patterns. Even though many of them come from the highest class in Haas' implicational scale, not all of them do. On the other hand, the lexemes to which /l/-vocalization is applied outside the expected area are mostly very frequent and thus salient in the GSW subcorpus and often interjections. These and more results will be presented in the poster.

Keywords: Swiss German dialects, /l/-vocalization, dialect use in WhatsApp messages

References

Christen, H. (2001). Ein Dialektmarker auf Erfolgskurs: Die /l/-Vokalisierung in der deutschsprachigen Schweiz. *Zeitschrift für Dialektologie und Linguistik*, (1):16–26.

Haas, W. (1983). Vokalisierung in deutschen dialekten. In Werner Besch, et al., editors, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Vol II*, pages 1111–1116. de Gruyter, Berlin.

Leemann, A., Kolly, M.-J., Werlen, I., Britain, D., and Studer-Joho, D. (2014). The diffusion of /l/-vocalization in Swiss German. *Language Variation and Change*, 26(02):191–218.

Sprachatlas der deutschen Schweiz. (1962 to 2003). Francke, Bern/Basel.

Ueberwasser, S. and Stark, E. (in print). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*.

Appendix

Contents

Author Index 75

Keyword Index 76

Author Index

A. Seza Doğruöz, 3
Abdulhafiz Alkhouli, 46
Aivars Glaznieks, 2
Angelika Storrer, 16, 52

Boris Borzic, 46

Céline Poudat, 52
Carole Etienne, 52
Cathrin Pichler, 21
Christophe Parisse, 52
Ciara Wigham, iii, 52
Claudia Marinica, 46

Dagmar Deuber, 65
Damjan Popič, 61
Daniel Pfurtscheller, 72
Darja Fišer, 4, 30, 52, 61

Egon Stemle, iii, 52
Erhard Hinrichs, 52
Evelyn Ziegler, 71

Harald Lüngen, 21, 52
Holger Grumt Suárez, 52

Julien Longhi, 46, 52

Laura Herzberg, 16, 21, 52

Lieke Verheijen, 6
Lisa Hilte, 11
Lothar Lemnitzer, 52
Lydia-Mai Ho-Dac, 52

Márton Petykó, 56
Marcel Fladrich, 71
Michael Beißwenger, 21, 39, 52, 71
Muhammad Shakir, 65

Nader Hassine, 46
Natali Karlova-Bourbonus, 52

Paola Leone, 44

Reinhild Vandekerckhove, 11

Simone Ueberwasser, 73
Stefania Spina, 25
Steven Coats, 35

Taja Kuzman, 30
Thomas Schmidt, 52
Tobias Horsmann, 39, 52
Torsten Zesch, 39, 52

Walter Daelemans, 11
Wilbert Spooren, 6
Wolfgang Imo, 71

Keyword Index

(linguistic) prestige, 61
/l/-vocalization, 73

annotation, 52, 71
anonymisation, 21, 52

bilingualism, 35
blog, 56

collection strategies, 71
conversations, 65
corpora, 21, 44, 52, 71
corpus linguistics, 35
cross-lingual CMC study, 16

data collection, 44
demonyms, 30
dialect writing, 2
discourse analysis, 30

emoticons, 25

Facebook, 2, 72

interaction signs, 16

language learning, 44
language modeling, 11
literacy, 6

machine learning, 3
media linguistics, 72
metadata, 44, 72
mixed-effects models, 25
motive attribution, 56
multidimensional analysis, 65
multilingualism, 3

nationalities, 30
NLP applications, 46
NLP for social media, 46
non-standardness, 11

orthography, 2, 61

part-of-speech, 39
protocol for data collection, 44
public service news, 72

quantitative methods, 35

rare phenomena, 39
register variation, 65
research infrastructures, 52

school writing, 6
Slovene, 30
small vs. large data sets, 3
social media, 6, 35, 39
social media data, 3
social media interaction, 72
Swiss German dialects, 73

teenage talk, 11
telecollaboration, 44
textometry, 46
troll(ing), 56
tweets mining, 46
Twitter, 25, 30, 35

WhatsApp, 6, 71
WhatsApp: dialect use in, 73
Wikipedia talk pages, 16
World Englishes, 65