Proceedings

of the 6^{th} Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)

17-18 September 2018, University of Antwerp

Proceedings

of the 6^{th} Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)

17-18 September 2018, University of Antwerp

Editors: Reinhild Vandekerckhove, Darja Fišer and Lisa Hilte

Antwerp, 2018

ISBN: 9789057285868

Conference website: https://www.uantwerpen.be/en/conferences/cmc-social-media-2018/

This publication is available from:
https://www.uantwerpen.be/en/conferences/cmc-social-media-2018/proceedings/
and is supported by:
CMC-corpora conference series and CLiPS research center

This publication was compiled using LuaLaTeX. The LaTeX template is based on KOMA script. The individual contributions are based on style files in LaTeX or MSWord format. These style files are modifications of the style files for the 2016 edition of the conference series, which are modifications of the Language Resources and Evaluation Conference (LREC) 2016 style files. The template and the style files are available online: https://github.com/cmc-corpora/.

Preface

This volume presents the proceedings of the 6^{th} edition of the annual conference series on 'Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)'. This conference series is dedicated to the collection, annotation, processing, and exploitation of corpora of computer-mediated communication and social media for research in the humanities and beyond. 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 6^{th} Conference on CMC and Social Media Corpora was hosted by the CLiPS research center at the University of Antwerp, Belgium, on September, 17^{th} and 18^{th}. This volume contains extended abstracts of the invited talks, short papers of oral presentations and extended abstracts of posters presented at the event. The conference attracted 29 valid submissions. Each submission was reviewed by members of the scientific committee. This committee decided to accept 14 papers and 12 posters of which 13 papers and 9 posters were presented at the conference. The program also includes two invited talks: a keynote talk by Beat Siebenhaar (University of Leipzig, Germany) and a keynote talk by Guy De Pauw (Textgain, Belgium). Finally, a tutorial on the online annotation platform WebAnno was given by Darja Fišer (University of Ljubljana and Jožef Stefan Institute, Slovenia).

The contributions in these proceedings cover a wide range of both topics and languages. No less than nine different languages are studied, including Arabic, Bengali, Dutch, English, French, German, Italian, Japanese and Romansh. Since different regional or national varieties are taken into account, geographic linguistic variation is one of the topics addressed in the contributions, and so are the code switching and borrowing patterns which often mark (informal) CMC utterances. Some contributions focus on the pragmatics of CMC by focusing on e.g. the use of discourse markers, on processes of linguistic accommodation and on community creation. Other papers link language use on social media to various aspects of social profiles or personality types, while covering a wide range of online media and forums (e.g. discussion forums, health forums). While some papers have a distinct linguistic research focus, others have a more practical orientation towards e.g. business applications. Finally, the creation of CMC corpora is a central topic of several submissions, with a focus on both the collection of the data itself and consequent editing, annotation and ethical issues.

We wish to thank all colleagues who 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), Ljubljana, Slovenia (2016), and Bolzano, Italy (2017), we hope that the Antwerp 2018 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 and beyond.

September 18, 2018
Antwerp

Reinhild Vandekerckhove, CLiPS, University of Antwerp (Belgium)
Darja Fišer, University of Ljubljana and Jožef Stefan Institute (Slovenia)

Chair of the Organizing Committee and Chair of the Coordinating Committee.

Table of Contents

Committees

Scientific Committee

Chairs	
Darja Fišer	University of Ljubljana and Jožef Stefan Institute (Slovenia)
Reinhild Vandekerckhove	CLiPS, University of Antwerp (Belgium)

Co-Chairs	
Michael Beißwenger	University of Duisburg-Essen (Germany)
Ciara R. Wigham	LRL, University of Clermont Auvergne (France)

Members	
Steven Coats	University of Oulu (Finland)
Helge Daniëls	KU Leuven (Belgium)
Daria Dayter	University of Basel (Switzerland)
Orphée De Clercq	Ghent University (Belgium)
Tomaž Erjavec	Jožef Stefan Institute (Slovenia)
Aivars Glaznieks	Eurac Research (Italy)
Axel Herold	Berlin-Brandenburgische Akademie der Wissenschaften (Germany)
Veronique Hoste	Ghent University (Belgium)
Gilles Jacobs	Ghent University (Belgium)
Mike Kestemont	University of Antwerp (Belgium)
Florian Kunneman	Radboud University Nijmegen (The Netherlands)
Els Lefever	Ghent University (Belgium)
Nikola Ljubešić	University of Zagreb (Croatia) and Jožef Stefan Institute (Slovenia)
Julien Longhi	University of Cergy-Pontoise (France)
Harald Lüngen	Institut für Deutsche Sprache (Germany)
Lieve Macken	Ghent University (Belgium)
Maja Miličević	University of Belgrade (Serbia)
Nelleke Oostdijk	Radboud University Nijmegen (The Netherlands)
Müge Satar	Newcastle University (United Kingdom)
Stefania Spina	University for Foreigners (Italy)
Egon W. Stemle	Eurac Research (Italy)
Angelika Storrer	University of Mannheim (Germany)
Simon Šuster	University of Antwerp (Belgium)
Hans van Halteren	Radboud University Nijmegen (The Netherlands)
Cynthia Van Hee	Ghent University (Belgium)

Coordinating Committee

Michael Beißwenger	University of Duisburg-Essen (Germany)
Darja Fišer	University of Ljubljana and Jožef Stefan Institute (Slovenia)
Ciara R. Wigham	LRL, University of Clermont Auvergne (France)

Organizing Committee

Walter Daelemans	CLiPS, University of Antwerp (Belgium)
Lisa Hilte	CLiPS, University of Antwerp (Belgium)
Reinhild Vandekerckhove	CLiPS, University of Antwerp (Belgium)

Invited Talks

Monitoring the Vox Populi:
Privacy, Free Speech and Other Opportunities

Guy De Pauw
Textgain
guy@textgain.com

Social media have allowed people to freely share their views on products, entertainment, politics, current events and… each other. This magnitude of opinions advances at an incredible pace, but managing the knowledge that is contained in this stream of unstructured data is no longer possible through mere human means. With automatic text analytics, however, we now have the technology to automatically collect and monitor opinions in order to turn language into insights, but also to convert it into money and strategic (political) advantage. Recent events have shown that this technology is indeed quite the double-edged sword.

In this talk, we will describe three case studies in which Textgain has leveraged their text analytics tools for societal gain:
1. Citizen participation: since its inception around the turn of the century, Web2.0 has reignited the dream of direct democracy, in which citizens can directly influence policy by voicing their opinions on online platforms. But this invariably involves the collection of a lot of (textual) data and the challenge of extracting insights that may guide policy makers. This case study shows how text analytics can help streamline this process through the use of data-driven modeling techniques.
2. Fake news: the post-fact era has had a huge impact on traditional media and social media alike. Opinion all too often trumps evidence and online filter bubbles tend to percolate stories that rarely serve the truth. In this case study, we will describe our work on the development of automatic techniques to detect the rhetoric of sensationalist and biased news media.
3. Online Extremism: venture just a little too far into social media and you will find yourself in a cesspool of polarization, hatred or even downright criminal rhetoric. ISIS is well known to use social media as a virtual battleground. Likewise, extreme-right communities incite each other inside deafening echo chambers of contempt for fellow human beings. This case study describes how machines can automatically be taught the rhetoric of such communities for the purpose of trend analyses and censorship tools.

We will zoom in on a very important and topical aspect of the last case study: the undeniable tension between privacy, censorship and free speech. The new privacy regulations of the European Union (GDPR) prohibit the collection of personally identifiable data. We are no longer allowed to uniquely identify extremists. Will we still be able to monitor extremism? Does ISIS have a right to privacy and free speech? And who gets to decide what constitutes (criminal) extremism? We hope to close this presentation with a lively discussion on these topics.

About Textgain

Textgain is a spin-off of the University of Antwerp and builds text analytics tools based on *machine learning*. This involves teaching computers how to recognize different types of texts by showing examples. Using this method, you allow machines to let the data speak for itself, rather than imposing a presupposed template of what one expects to find in the data.

Accommodation in WhatsApp Communication

Beat Siebenhaar
University of Leipzig
siebenhaar@uni-leipzig.de

Currently, WhatsApp is the most popular instant messaging application for smartphones. The huge amount of messages that are exchanged via WhatsApp on a daily basis opens the possibility for linguists to analyse a very dynamic form of written language. In order to do so, a corpus of WhatsApp chats was collected in Switzerland in summer 2014 (Ueberwasser & Stark 2017) and in Germany in winter 2014/15, containing a total number of about 1.2 million speech bubbles, some dating back to 2010. For this presentation, this dataset is used to answer questions about how individuals interact with communication partners and how they (non-)accommodate to their partners. On the one hand, I will look at emojis, which are of special interest for linguistic analysis because, even though they are broadly used in mobile communication, they are not yet part of a written standard. On the other hand, I will look at non-standard writing that is of a very different quality in the Swiss and the German data. Qualitative analyses of the data shed light on the functions in which individuals use different emojis and spellings in interaction, possibly being influenced by how their chat partners use these forms (cf. Functions of adjustments in the Communication Accommodation Theory, Dragojevic, Gasiorek & Giles 2016). How large this influence is, depends on the degree to which the individuals have developed habits in their use of spelling and emojis. Where individuals change their communicative patterns in the direction of their chat partners, this can be seen as instances of microsynchronisation in the sense of the linguistic dynamics approach (cf. Schmidt 2009). With quantitative approaches to specific chats I will show how in a process of mesosynchronisation sequences of microsynchronisations can result in a stabilisation on the level of two individuals. A further look at the use within the whole corpus may even point to a stabilisation in the communication community that could be a new orientation point for macrosynchronisation that retroacts on the individual use. With this focus on the use of individuals it will be possible to investigate how norms emerge in interaction and to analyse language dynamics and change. Moreover, in comparing these norms in the Swiss and in the German data it will be clear that despite of the globalisation processes fostered by CMC these norms show clear cultural borders within the German linguistic area.

References

Dragojevic, Marko, Jessica Gasiorek and Howard Giles (2016). Accommodative Strategies as Core of the Theory. In: Giles, Howard (ed.): *Communication Accommodation Theory. Negotiating Personal Relationships and Social Identities across Contexts*. Cambridge: CUP, pp. 36-59.

Schmidt, Jürgen Erich (2009). Language and space: The linguistic dynamics approach. In: Auer, Peter and Jürgen Erich Schmidt (eds.): *Language and Space: Theories and Methods. An International Handbook of Linguistic Variation*. Berlin & New York: De Gruyter, pp. 201-225.

Ueberwasser, Simone and Elisabeth Stark (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online,* 84, pp. 105-126.

Papers

BTAC: A Twitter Corpus for Arabic Dialect Identification

Mohammed Altamimi, Osama Alruwaili, William J. Teahan

University of Hail, Aljouf University, University of Bangor

mh.altamimi@uoh.edu.sa, osalruwaili@ju.edu.sa, w.j.teahan@bangor.ac.uk

Abstract

Arabic is spoken with different dialects throughout the Middle East and North Africa. However, dialectal corpora for Arabic are less prevalent compared to other languages. Recently, dialectal Arabic has witnessed growth over the web in the form of social media. The purpose of this paper is to present a new dataset for Arabic dialects collected from Twitter. Over 122K tweets have been collected. The Tweets have been annotated manually into five dialects in addition to Modern Standard Arabic and Classical Arabic. The Kappa result of the Inter Annotator Agreements is 0.864. The annotation and cleaning process is described in detail. Mixed dialects (where code-switching occurs) have also been tagged for further research.

Keywords: dialectal Arabic corpora, development of CMC corpora, Twitter, code-switching.

1. Background and motivation

Availability of resources is an important issue for NLP, and for Arabic in particular. There have been various efforts from researchers to create Arabic corpora (Al-Thubaity, Khan, Al-Mazrua, & Al-Mousa, 2013; Alrabiah, Al-Salman, & Atwell, 2013; Maamouri, Bies, Buckwalter, & Mekki, 2004; Smrž & Hajic, 2006). However, most of these contributions have targeted Modern Standard Arabic or Classical Arabic due to the accessibility of Arabic newspaper texts, books, and more recently weblogs and forums. Lately, social media has played a vital role in expanding dialectal resources for researchers. This notable increase in availability of dialectal Arabic has provided the motivation to produce this work.

The specific objective of this paper is to create a dialectal corpus for Arabic using Twitter text. Microblogging or Twitter messages is considered a unique style of writing compared to other corpora due to the short length of tweets, the type of language used, and the availability of data. This corpus would aid research in text analysis areas such as dialect identification and code-switching analysis.

Arabic is spoken by over 300 million people. It is the fifth most popular language in the world after Chinese, Spanish, English, and Hindi. It is widely used in most of the Middle East countries as a first or second language. There are three forms of Arabic: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). CA represents an older style of Arabic. It was used more among people in the sixth and seventh century (pre-Islam century) and continued beyond that (Holes, 2004). CA is used in religious books such as the Quran, Hadith (the speech of the prophet Muhammad), and also some traditional books such as poetry and history. MSA is more formally used by all the Arabic speaking people. It started to become popular with the increase of Arabic media in the second half of the 19th century and its popularity has continued until now. It is used more in modern life media sources such as newspapers, magazines, and formal TV programs. However, MSA is used much less frequently in daily conversation where people use dialects to communicate. Dialects are widespread and divided by geographical region and are used informally in daily life communication with people from the same location. However, it can often be difficult for Arabic speaking people to understand each other when using dialects from different regions.

There are many Arabic dialects that are used in the Arab world. However, there are five main dialects that are widely spread: the Gulf dialect, the Egyptian dialect, the Maghrebi dialect, the Levantine dialect, and the Iraqi dialect. The Gulf dialect gets used in countries such as Saudi Arabia, Kuwait, Bahrain, Qatar, Emirates, and Oman. The Egyptian dialect is widely spoken only in Egypt. The Maghrebi dialect includes dialectal variations from countries such as Morocco, Tunisia, Libya, and Algeria. The Levantine dialect includes countries such as Syria, Lebanon and Jordan, and in Palestine. Finally, the Iraqi dialect is spoken only in Iraq. Other dialects such as Sudanese, Somali, Yamani, and Mauritania are not included in our research due to the lack of the use of social media for these variations (Salem, 2017).

Our contributions for this paper can be summarized as follows:

- Over 200K tweets have been collected and used to build a new dialectal corpus for Arabic tweets.
- After cleaning, over 122K tweets were labelled into five dialects in addition to MSA and CA.
- The corpus includes other labels for each tweet such as gender, authorship, and topic to allow for researchers to perform other types of text analyses. Also, mixed dialects are tagged for further code-switching research.

The rest of the paper is organized as follows: section two provides the related work; section three describes the data collection process; section four gives an overview of the annotation process; section five mentions the annotation evaluation; section six discusses corpus statistics; section seven describes an initial research experiment; and finally, section eight provides the conclusion and future work.

2. Related work

Research that is specifically concerned with Dialectal Arabic is limited. Gadalla et al. (1997) created the first DA corpus called CALLHOME that focused on the Egyptian dialect, mainly collected from phone conversations. The Gumar corpus (Khalifa, Habash, Abdulrahim, & Hassan,

2016) was built from over 100M words of the Gulf dialect collected from novels. Diab et al. (2010) harvested weblogs with the focus on both Egyptian and Levantine dialects. Saad (2017) collected text in both Egyptian and MSA dialects from Wikipedia articles. Harrat et al. (2017) created an Arabic parallel corpus that is built for dialects: Maghrebi, Tunisia, Algerian, Palestine, and Syrian. It consists of 6400 sentences in each of the five dialects in addition to MSA. The Curras corpus (Jarrar, Habash, Alrimawi, Akra, & Zalmout, 2017) contains 56K tokens focused only on the Palestinian dialects. Omar and Callison (2011) produced a corpus mainly from newspaper commentary and Twitter data to perform dialect and genre categorization. Salama et al. (2014) built a corpus for dialectal Arabic collected mainly from YouTube commentaries, with multiple dialects such as Egyptian, Gulf, Iraqi, Maghrebi and Levantine.

Recently, Twitter has provided a rich resource for collecting dialectal text. Many researchers have taken advantage of this and used its API to collect texts in the form of tweets. Mubarak and Darwaish (2014) collected over 175M Arabic tweets. Those tweets are filtered according to the user location to perform dialect categorization on a subset of 6.5M tweets. The research by Alshutayri et al. (2016) also explores Twitter as a source of Arabic dialects. 8090 collected tweets were annotated according to the unique words of each dialect and user location.

3. Bangor Twitter Arabic Corpus (BTAC)

This section discusses the new corpus we have developed, the Bangor Twitter Arabic corpus (BTAC). It is believed that a smaller consistent corpus that represents a high-quality design is far more valuable than a larger corpus (Granger, 1993). Our goal is to create a corpus that contains high quality ground truth data. Other research presumes each tweet belongs to a dialect according to the geographical information of the tweet (latitude and longitude). However, we do not assume the tweets belong to a specific dialect by the tweet or username location. We checked the language used in the tweet to assess whether it belongs to a dialect. We also annotated the tweet manually according to the dialects used in the tweet to isolate dialectal text from Modern Standard Arabic and Classical Arabic text. In addition, we identified when mixed dialect (code-switching) occurs in some of the tweets for further research.

3.1 Collection process

We selected over 100 users from different locations. The selection process was based on the account location, profile picture, and bio information for users of the five main dialects: Egyptian, Gulf, Iraqi, Maghrebi, and Levantine. Twenty users were selected for each dialect in different topics such as religion, culture, politics, sport, and general. For the sake of balancing between both genders, the selection process involved both genders for performing gender classification (Altamimi & Teahan, 2017).

In order to create the training dataset, 2K tweets were collected from each account. In total, over 200K tweets were collected using the Tweepy library (Tweepy, 2009). Tweepy is a Python package that interacts with the Twitter API for collecting data. Also, certain hashtags were crawled separately and added to the training set afterward. The reason for this was to expand the size for the Iraqi, Levantine, Maghrebi dialects. These hashtags were chosen according to the geolocation from people speaking those dialects. Table 1 shows the list of hashtags we used.

Dialects	Hashtags
Iraqi	#شعر_عراقي #غزل_عراقي #اشعار_عراقيه #يوميات_مگرود #دارميات #شعر_شعبي_عراقي
Levantine	#لو_بتحبني #انك_لبناني_يعني #النقل_المشترك #بلد_الظلام
Maghrebi	#غرد_بالداريجه #غرد_بالامازيغيه #تونس_المزيانة #محرز_في_ليفربول

Table 1: List of hashtags.

3.2 Processing steps

In order to clean the text, the following processing steps were applied to all the tweets. A sample tweet before and after processing is shown in Table 2.

- Retweeted tweets are removed. This is to ensure that the tweets were collected for a specific username and were not tweeted by another person.
- Duplicate tweets were also removed. This is to eliminate redundant duplicate tweets.
- HTTP links, usernames, images, and non-Arabic tweets were also removed as the focus was only on Arabic text, and also to ensure that tweets do not contain spam and other non-relevant data that would not help when performing classification.

Label	Tweet
Before processing	@RyBookFair: قرايز تاقوأ باتكللي_ليلودلا_ضايرلا_ضرعم# https://t.co/A2lWzgBtq7
After processing	قرايز تاقوأ باتكللي_ليلودلا_ضايرلا_ضرعم#

Table 2: A sample tweet before and after processing.

- Hashtags, emoji's, stop words, and special characters such as underscores, and quotes were retained. We wish to keep this information as it can aid identification when we perform classification experiments in the future.

The test datasets were collected separately. This was to ensure there was no overlap between the training and testing sets and to ensure future evaluations using the dataset are consistent. The testing sets were collected at three different time periods in March, April, and July. The

File	Number of tweets before processing	Number of tweets after processing
Training	200.917	112.060
Testing	15.000	6.890
Hashtags	17.918	3.925

Table3: Size of the corpus before and after processing.

testing sets were processed the same way as we processed the training set. However, this time only the top 50 tweets were collected from the same usernames in our training set. The number of tweets including training, testing, and hashtags before and after processing is shown in Table 3.

4. Annotation process of the BTAC

As stated, the corpus contains five dialects, Egyptian, Gulf, Iraqi, Maghrebi and Levantine in addition to MSN and CA. Two Arabic native speakers (postgraduates with experience in NLP) have independently annotated the corpus manually. The goal of the annotation is to identify whether the tweet is written using one of the dialects, or MSN or CA. Tweets that could not be assigned to one of the dialects are marked as unknown so that they can be identified and excluded from both the training and testing sets if needed. Moreover, code-switching has also been identified for tweets that are written in mixed dialects (for example, a tweet that is written in the Egyptian dialect followed by MSA text or CA).

4.1 Annotation labelling

In order to assure the annotators followed the same annotation steps, we used the following annotation labels:

- Dialect: For tweets are written in one of the dialects, these should be annotated under the name of the dialect.
- Classical Arabic: This includes tweets that contain old writing styles such as the Quran, Hadith, Dao, or Poetry.
- Modern standard Arabic. This is for tweets written in a modern style of Arabic such as newspapers, magazines, or TV programs.
- Unknown: This is for tweets that are not meaningful or contain only symbols such as emojis or undetermined text.
- Mixed: This is for tweets that contain two dialects. The name of the second dialect is mentioned in this case.

5. Annotation evaluation

In order to evaluate the quality of the annotation, we used the *Kappa coefficient,* κ (Cohen, 1960) for measuring inter-annotator agreements (IAA) between the two annotators. We measured the *Kappa coefficient* for the total of 122K tweets that were annotated by our two annotators. Our results indicate that the obtained Kappa value was 0.864 for all the MSA, CA, and Dialects tweets as shown in Table 4. Our Kappa result is considered perfect according to Landis & Koch (1977). This reflects the correlation agreement of our annotators.

After checking the disagreement of the annotated tweets between the two annotators, we found that the cause for the difference was one of the following two reasons:

1. More than two dialects could be assigned to the tweets. To solve this disagreement, we added the annotation that indicated that code-switching had occurred to the other dialect.
2. Human error was the reason for disagreement. To overcome this, we modified the annotation label to reflect the accurate dialect after an agreement was reached between both annotators.

File	Agreement	Disagreement	Observed Agreement	Kappa
MSA	46.197	6.617	94.6	0.888
CA	33.000	15.323	87.5	0.723
DA	34.098	828	99.3	0.983

Table 4: Inter-annotator agreement, disagreement, and Kappa values.

6. Corpus statistics

Out of the 122K tweets collected, we found that the majority of the tweets were written in Modern Standard Arabic and Classical Arabic style. Most of the tweets that were written in Modern Standard Arabic style were generally news feed with a predominantly political influence due to the current situation in the Middle East. However, tweets that were written in Classical Arabic were religious, historical, and cultural tweets. The rest of the tweets were written in dialectal form ordered by the number of tweets as follows: Gulf (8%), Egyptian (8%), Levantine (7%) Maghrebi (3%), and Iraqi (1%). The Gulf and Egyptian dialects are highly used in social media especially on Twitter. The Maghrebi, Levantine and Iraqi dialects made the lowest number of tweets out of the collected tweets. This is due to Twitter not being very popular in those countries. In addition, countries such as Morocco, Tunisia, and Algeria also use other languages to communicate. Undetermined tweets were labeled as unknown. Those tweets contained emoji's, numbers and undecided dialectal text. Lastly, code-switching occurs in 2% of the entire corpus. Most of these tweets are mixed with either CA or MSA. Only three tweets were found to be mixed with other dialects as it was found to be extremely rare to have people speaking multiple Arabic dialects.

Unigram samples taken from the tweets show the obvious distinctions between all of the Arabic dialects. Table 5 shows the top 10 unigrams for BTAC. This table is generated by removing the stopwords using the list produced by Alrefaie (2016). Buckwalter transliteration is provided for illustrating the differences between the subset in Latin script using the same technique used by Alkhazi & Teahan (2017).

7. Initial text categorisation experiment

We performed an initial text categorisation experiment on our dataset using the WEKA toolkit (Hall et al., 2009). Our training set contained 105.59K labelled tweets divided unequally between the Arabic dialects: 9.15K from Gulf, 9.06K from Egyptian, 3.98K from Maghrebi, 7.86K from Levantine, 1.88K from Iraqi, 31.00K from CA, and 42.66K from MSA. We used the Multinomial Naive Bayes (MNB) algorithm to classify the tweets. Testing sets contained 6.587 labelled tweets: 615 for Gulf; 566 for Egyptian; 167 for Maghrebi; 360 for Levantine; 86 for Iraqi; 1463 for CA, and 3330 for MSA. We eliminated tweets that contained codes-switching and tweets that were labelled unknown from both training and testing sets. We achieved 0.723,

0.717, and 0.713 in Precision, Recall, and F-Measure.

8. Conclusion

In this paper, we present the BTAC, which contains over 122K annotated tweets for five Arabic dialects: Gulf, Egyptian, Levantine, Maghrebi, and Iraqi, in addition to Modern Standard Arabic and Classical Arabic. This corpus represents a valuable and rich resource for NLP applications targeting Arabic dialects. The annotation also highlights some tweets that contained code-switching. The evaluation of our annotators' performance is considered perfect according to the measurement of observer agreement for categorical data. We plan in the future to evaluate the corpus and compare it against other existing corpora. We will also explore various dialectal classification algorithms, perform automatic annotation, and implement dialect text segmentation by using some dialectal tweets from this corpus. For a copy of the corpus, please contact one of the authors.

9. References

Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). New language resources for arabic: corpus containing more than two million words and a corpus processing tool. In *Asian Language Processing (IALP), 2013 International Conference on* (pp. 67–70). IEEE.

Alkhazi, I. S., & Teahan, W. J. (2017). Classifying and Segmenting Classical and Modern Standard Arabic using Minimum Cross-Entropy. *International Journal of Adv, Comp. Science and Applications*, *8*(4), pp. 421–430.

Alrabiah, M., Al-Salman, A., & Atwell, E. S. (2013). The design and construction of the 50 million words KSUCCA. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics* (pp. 5–8). The University of Leeds.

Alrefaie, M. (2016). Arabic Stop Words. Retrieved April 20, 2018, from https://github.com/mohataher/arabic-stop-words/blob/master/list.txt

Alshutayri, A., Atwell, E., Alosaimy, A., Dickins, J., Ingleby, M., & Watson, J. (2016). Arabic language WEKA-based dialect classifier for Arabic automatic speech recognition transcripts. In Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (pp. 204-211).

Alshutayri, A. O. O., & Atwell, E. (2017). Exploring Twitter as a Source of an Arabic Dialect Corpus. *International Journal of Computational Linguistics* (IJCL), 8(2), pp. 37–44.

Altamimi, M., & Teahan, W. J. (2017). Gender And Authorship Categorisation Of Arabic Text From Twitter Using PPM. *IJCSIT*, 9(2), pp. 131-140.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37–46.

Diab, M., Habash, N., Rambow, O., Altantawy, M., & Benajiba, Y. (2010). COLABA: Arabic dialect annotation and processing. In *LREC workshop on semitic language processing* (pp. 66–74).

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., … Kingsbury, P. (1997). CALLHOME Egyptian Arabic Transcripts. *Linguistic Data Consortium, Philadelphia*.

Granger, S. (1993). English language corpora: Design, analysis and exploitation. *The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.)*, (pp. 57–69).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), pp. 10–18.

Harrat, S., Meftouh, K., & Smaili, K. (2017). Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., & Zalmout, N. (2017). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51(3), pp. 745–775.

Khalifa, S., Habash, N., Abdulrahim, D., & Hassan, S. (2016). A large scale corpus of Gulf Arabic. *arXiv Preprint arXiv:1609.02960*.

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

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic language resources and tools* (Vol. 27, pp. 466–467). Cairo.

Dialects	Classic		Modern		Gulf		Egyptian		Levantine		Maghrebi		Iraqi	
Top 10 unigrams	Allhm	اللهم	AlErAq	العراق	Ally	اللي	m$	مش	Em	عم	twns	تونس	mw	مو
	Al$yx	الشيخ	mSr	مصر	w$	وش	Ally	اللي	$w	شو	Ally	اللي	dArmy	ارمي
	AlnAs	الناس	swryA	سوريا	$y	شي	dh	ده	hyk	هيك	rby	ربي	byh	بيه
	Abn	ابن	AlmwSl	الموصل	mw	مو	dy	دي	Ally	اللي	bA$	باش	hAy	هاي
	AlHyAp	الحياة	IyrAn	إيران	ybArk	يبارك	Ayh	ايه	m$	مش	m$	مش	Any	اني
	AlIslAm	الإسلام	AlEAlm	العالم	tslm	تسلم	mSr	مصر	$y	شي	br$A	برشا	ErAqy	عراقي
	mHmd	محمد	Hlb	حلب	E$An	عشان	E$An	عشان	Anw	انو	$y	شي	rwHy	روحي
	$Ahd	شاهد	AlErbyp	العربية	AlnSr	النصر	AlzmAlk	الزمالك	knt	كنت	ly	لي	$nw	شنو
	AlImAm	الإمام	dAE$	داعش	bEmrk	بعمرك	kdh	كده	rH	رح	$kwn	شكون	ly$	ليش
	AlslAm	السلام	AlsEwdyp	السعودية	yArb	يارب	rbnA	ربن	HdA	حدا	ky	كي	$lwn	شلون

Table 5: The top 10 unigrams showing the clear distinction of each subset of BTAC.

Mubarak, H., & Darwish, K. (2014). Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 1–7).

Saad, M. (2017). Egyptian comparable Wikipedia corpus. Retrieved April 4, 2018, from: https://www.kaggle.com/mksaad/arb-egy-cmp-corpus

Salama, A., Bouamor, H., Mohit, B., & Oflazer, K. (2014). Youdacc: the youtube dialectal arabic comment corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Salem, F. (2017). *The Arab Social Media Report 2017: Social Media and the Internet of Things towards Data-Driven Policymaking in the Arab World.* (Vol. 7). Dubai. MBR School of Government.

Smrž, O., & Hajic, J. (2006). The other Arabic treebank: Prague dependencies and functions. *Arabic Computational Linguistics: Current Implementations. CSLI Publications*, *104*.

Tweepy. (2009). Tweepy. Retrieved March 5, 2016, from Tweepy.org

Zaidan, O. F., & Callison-Burch, C. (2011). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 37–41). Association for Computational Linguistics.

News from the *MoCoDa²* Corpus: A Design and Web-Based Editing Environment for Collecting and Refining Data from Private CMC Interactions

Michael Beißwenger[1], Marcel Fladrich[2], Wolfgang Imo[2], Evelyn Ziegler[1]
[1] Universität Duisburg-Essen [2] Universität Hamburg
michael.beisswenger@uni-due.de, marcel.fladrich@uni-hamburg.de,
wolfgang.imo@uni-hamburg.de, evelyn.ziegler@uni-due.de

Abstract

The paper reports on findings from the *MoCoDa²* project which is creating a corpus of private CMC interactions from smartphone apps based on donations by their users. Different from other projects in the field, the project involves users not only as donators but also as editors of their data: In a web-based editing environment which provides users with access to their raw data, they are supported in pseudonymising their data and enhancing them with rich metadata on the interactional context, the interlocutors and their relations, and on embedded media files. The resulting corpus will be a useful resource not only for quantitative but also for qualitative CMC research. For representation and annotation of the data the project builds on best practices from previous projects in the field and cooperates with a language technology partner.

Keywords: CMC, corpora, collection strategies, user involvement

1. Introduction

In recent years, there has been an increasing amount of work dedicated to the creation of corpora of computer-mediated communication (CMC) that shall be made available as resources for the scientific community through established corpus infrastructures (e.g. CLARIN) and through adapting to standards in the field of Digital Humanities (Beißwenger et al., 2017a; 2017b).

A desideratum in the CMC corpora landscape are resources that represent CMC discourse from the private sphere and that allow for research of discourses found in applications such as WhatsApp and similar mobile chat and messaging services which are frequently used by adolescents (cf. e.g. KIM, 2016; JIM, 2016). Data of that type as well as the metadata needed for an adequate interpretation (topic and context of the interaction; age, sex, languages, and socio-demographic background of interlocutors; social relations between interlocutors) can only be collected with the help of the users themselves.

We report on findings from the project *MoCoDa²* (*Mobile Communication Database*)[1], which is funded by the Ministry for Innovation, Science, Research and Technology of the German federal state North Rhine-Westphalia and in which a team of researchers from two universities has created a database and web front-end for the repeated collection of written CMC from mobile messaging services.

2. Related work

Collections of CMC data from the private sphere have been addressed in several previous projects: In the DiDi project (Frey et al., 2014), Facebook users were asked to give their permission to collect CMC data from their profile pages via a web application. However, the collection of data from private mobile phones can only be achieved when users actively donate their data by submitting them to a project API. Practices for donation-based collections of SMS and WhatsApp data have been developed in the *sms4science* project (Dürscheid & Stark, 2011)[2], in the projects "Whats up, Switzerland?" (WuS)[3] and "Whats Up, Deutschland?" (WuD)[4], and in the predecessor project *MoCoDa¹* (cf. Sect. 4 and Imo, 2015; 2017).

3. Data collection and editing design

MoCoDa² adopts a donation-based collection strategy. Different from the aforementioned projects, the CMC users are not only involved as donators but also as editors of their donated data: The data collection component allows users to donate selected parts of their private interactions via email and then to log into a web-based editing environment in which they can edit their donations, pseudonymise the data, and enhance them with relevant metadata to transform them into valuable contributions to a corpus that can be a useful resource both for quantitative and qualitative research on CMC. The Language Technology Lab at the university of Duisburg-Essen, the language technology partner in the project, provides an infrastructure which adds token and part-of-speech information to the data. The tokenisation task is performed as a preparatory task for the editing process, the part-of-speech task is performed after donators have finished editing their donated data.

As a matter of fact, the amount of data donated by a single user is expected to be smaller than the data collected in the WuS and WuD projects mentioned in Sect. 2, where the complete logfiles stored on the donators' mobile phones were submitted to the project API and integrated into the corpora. The goal of the MoCoDa² project is not to create a corpus which is intended as a competitor to the WuS and WuD corpora; instead, the goal is to create a corpus of interaction sequences which have been manually selected and edited by their donators to provide corpus users with all metadata needed to use the corpus for qualitative research. During the editing process users keep full control of their donations and may crop the donated log file to

[1] https://www.mocoda2.de

[2] http://www.sms4science.org/

[3] http://www.whatsup-switzerland.ch/

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

a certain selection. Additionally, and in contrast to the aforementioned projects, MoCoDa² does not perform a one-time collection, but the front-end is used repeatedly so that the size of the corpus will gradually grow over time and – as a long-term objective – allow for micro-diachronic research on language variation and change in mobile messaging discourse.

The figures given below illustrate how the online data-editing process is organised. The language of the interface is German.

Specification of metadata: The system automatically extracts the names of all participants of an interaction from the donated logfile and encourages the donator to add metadata for each of these individuals. Metadata for individuals include information on age, sex, place of residence (city, state, country), place of birth (city, state, country), educational level, profession, first language and other languages used in everyday life.

In a second step the donator is asked to define the social relations between the individuals that are detected as chat participants pair-by-pair. Relations can be specified according to several predefined dimensions of which more than one may apply and can be assigned to chat partners:

familial relations
- is married to
- is a parent of
- is a child of
- is a sibling of
- is a grandparent of
- is a grandchild of
- is an uncle/an aunt of
- is a nephew/a niece of
- is a cousin of
- is in a permanent relationship with
- is an ex-partner of
- is close friends with
- is a friend of
- is a close acquaintance of
- others

professional relations
- is his or her supervisor
- is his or her employee
- provides service for
- uses services of
- is a colleague of
- is a business partner of
- others

educational relations
- is a teacher of
- is a student of
- is his or her lecturer
- is a (university) student of
- is his or her instructor
- is a trainee of
- is a classmate of
- is a fellow student of
- is a fellow trainee of
- others

relations in the area of non-work and leisure activities
- is his or her group leader
- is a member of his or her team
- is his or her coach
- others

Fig. 1 illustrates how the values can be assigned to a pair of individuals (Stefan and Susanne) via selection from a dropdown menu. The predefined values can be enhanced with textual descriptions (e.g. as given in Fig. 1 "know each other from kindergarten" as an additional explanation of the predefined value 'is good friends with'). Fig. 2 shows an overview generated by the system on how many relations between the detected chat participants have been specified by a donator within a donated interaction. Fig. 3 shows how the metadata specified for one participant in a group chat (Alexa) is presented on the MoCoDa² user interface together with a donated sequence. Rows 1–12 contain individual metadata, row 13 includes information on the relations of the individual with the other participants (Olivia, Anna).

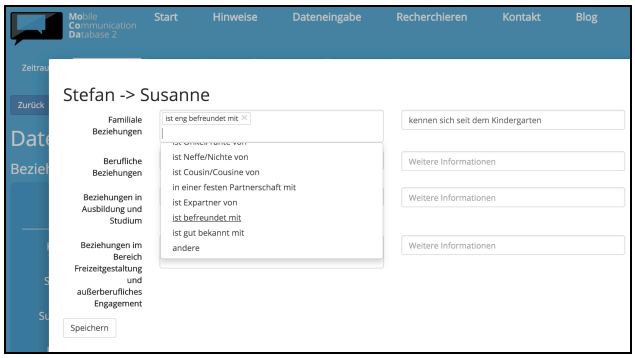

Fig. 1: Screenshot: Using the assistant for the specification of social relations between interlocutors.

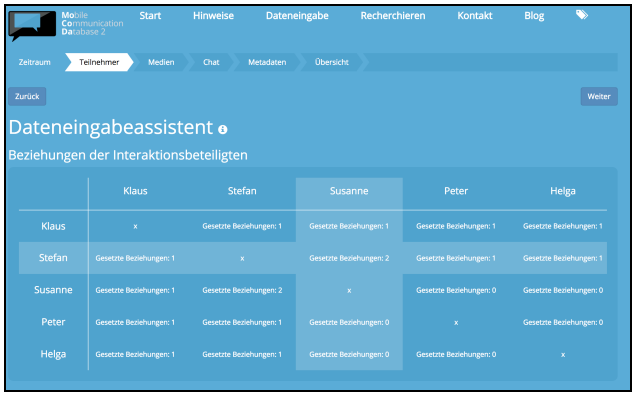

Fig. 2: Screenshot: Overview of relations specified for all couples of interlocutors in a donated sequence.

Pseudonymisation: When entering metadata for the interaction participants, donators are asked to assign a pseudonym to each of the participants following the following rule:

> "Please make sure to choose a realistic name as a pseudonym that resembles the gender and origin of the original name. A person's place of birth or residence only needs to be anonymised if it is a very small place."

All mentions of the participants' names as author names of posts in the logfile are then automatically replaced using the pseudonyms specified by the donator.

In a next step, the donators are asked to assign pseudonyms to person names mentioned in the body of the user posts. They can click on a token of their choice and replace it with a pseudonym. Once a certain character string (e.g. "Matthis") has been replaced by a pseudonym (e.g.

"Manuel"), for every other occurrence of the respective character string in all other posts of the sequence the system suggests to replace them by the same pseudonym accordingly. The donator can agree with the suggestion by clicking on the green "check" symbol that appears along with the automatically generated suggestion (Fig. 4: system suggests to replace the character string "Matthis" by "Manuel" because the donator has replaced a previous occurrence of "Matthis" by "Manuel"). During the process the system stores a temporary list of all pseudonyms previously used by the donator; this list is deleted from the database when the donator declares the editing process as finished. Besides the pseudonymisation of person and – if needed – city names, the donators are asked to anonymise URLs by replacing them with the category 'URL'.

Further editing steps include the formulation of textual description for media objects (images, videos) included in the original sequence (which, due to copyright restrictions, are not stored in the MoCoDa² database), a transcription of audio posts (which are not stored in the database due to IPR restrictions), a textual description of the interaction context and the specification of a brief title for the donated sequence.

Fig. 5 gives an example of how a donated and post-edited sequence is presented on the MoCoDa² user interface together with the title ("Skatergirls") and the textual description of the interaction context entered by the donator (left column). The sequence in the screenshot contains a textual description of an image that was embedded in the original data ("Zu sehen ist ein kleiner Pool, ..."). The pseudonyms of the participants given in the bottom of the left column under "Teilnehmer" link to representations of the participant metadata as given in Fig. 3.

Alexa

Alter:	26 - 30
Geschlecht:	weiblich
Wohnort Land	Deutschland
Wohnort Bundesland:	Nordrhein-Westfalen
Wohnort Stadt:	Essen
Geburtsort Land:	Deutschland
Geburtsort Bundesland:	Nordrhein-Westfalen
Muttersprache(n):	Polnisch
Alltagsprache(n):	Polnisch, Deutsch
Höchster Bildungsabschluss:	Abitur
Berufsgruppe:	Student/in
Weitere Informationen zum Beruf:	Arbeitet außerdem seit 3 Monaten Vollzeit als wissenschaftliche Mitarbeiterin an ihrer Universität.
Beziehungen:	*Olivia:* • Alexa ist befreundet mit Olivia • Alexa ist Kollege oder Kollegin von Olivia *Anna:* • Alexa ist eng befreundet mit Anna • Alexa ist Kollege oder Kollegin von Anna

Fig. 3: Presentation of metadata for one interlocutor on the MoCoDa² user interface (German).

Fig. 4: Screenshot: Using the pseudonymisation assistant.

4. Resources and technology

MoCoDa² builds on the expertise and resources from three preceding projects:

- *MoCoDa¹* (Imo, 2015; 2017) – a corpus project with a similar profile which had been initiated to collect a database for the qualitative analysis of CMC[5]. Since 2012, this project has collected a (relatively small) data set of 2,198 interactions with 19,161 user posts with ~193,000 tokens. For *MoCoDa²*, the database and web front-end have been re-implemented from scratch, and especially the editing environment was supplemented with a lot of additional functions and features.
- *ChatCorpus2CLARIN* (Lüngen et al., 2016) – a curation project in the context of the German CLARIN-D initiative in which the Dortmund Chat Corpus (Beißwenger, 2013), a well-established CMC corpus for German, has been remodelled following up-to-date standards for corpus resources in the digital humanities and integrated into the CLARIN language resources infrastructure[6]. The project has developed guidelines for anonymisation of CMC resources (Lüngen et al., 2017) and a schema for the representation of CMC corpora building on the TEI-P5 encoding guidelines of the Text Encoding Initiative (TEI)[7]. The schema and the anonymisation guidelines are adopted for representing and editing CMC data in *MoCoDa²*.

[5] http://mocoda.spracheinteraktion.de/

[6] The corpus can be retrieved via the repositories of the Institute for the German Language (IDS) at http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E and of the Berlin-Brandenburg Academy of Sciences (BBAW) at http://hdl.handle.net/11858/00-203Z-0000-002D-EC85-5. It can be queried online via the COSMAS II interface of the German Reference Corpus (DeReKo) and – after a free registration – via the text corpora component of http://www.dwds.de, the online information system on German language provided by the BBAW.

[7] The schema can be retrieved via http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema. A detailed description is given by Beißwenger (2018).

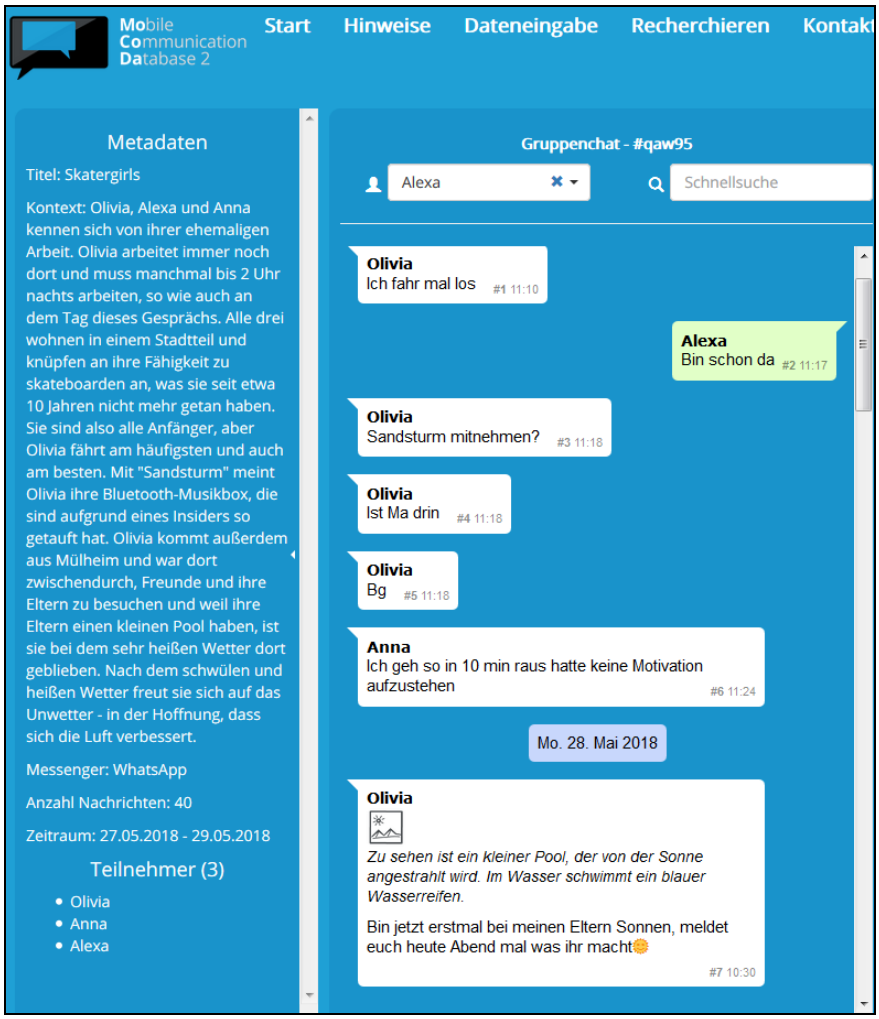

Fig. 5: Presentation of a donated sequence on the MoCoDa² user interface.

- *EmpiriST 2016* – a shared task on tokenisation and part-of-speech tagging of German CMC/social media data (Beißwenger et al., 2016) which developed guidelines for tokenisation and a part-of-speech tag set for German CMC ('STTS 2.0'[8]) and which resulted in a number of tokenisers and taggers which had been adapted or retrained to fit the linguistic peculiarities of CMC discourse. The tools and tag set used for the annotation of the *MoCoDa²* data build on the EmpiriST resources and results.

MoCoDa² is currently running on a development server provided by *ling•data*, a software company which is involved in the project as a partner. The technological backbone of the project is a *mongoDB* database which performs fast enough to execute processing operations relevant during the online editing process in real-time. In view of the large amount, data have to be processed in short time (raw text, tokenisation, annotations, metadata). The technology has completely been built on a *JavaScript* base with *Node* using the *Angular* framework. The system uses different microservices in order to handle certain operations like parsing, tokenising, annotating and indexing as different processes. The *docker* technology allows us to use load balancing and thus provide an optimal performance while importing larger amounts of data or handling several donation and editing processes simultaneously. Our beta tests have shown that the system can import even long logfiles quite fast and provide users with all information needed for editing their donations in (almost) real time.

5. Results so far

A beta version of the data collection and editing component has been subject of testing and optimisation in several university classes at three German universities during the summer term 2018. In this period, we collected 151 sequences from WhatsApp logs with 7,069 user posts and 69,734 tokens. It is planned to make the donation and editing component of the resource publicly accessible in late 2018 and then repeatedly encourage smartphone users (for example in, but not limited to, the context of university classes and projects with schools) to contribute to the further extension of the database. Donations entered into the database by users are checked for unethical content by the project leaders on a regular basis so that "borderline" content or sequences which were obviously not completely pseudonymised can be removed if necessary.

As a next step, the resulting corpus shall be made availa-

[8] The tag set and annotation guidelines can be retrieved via https://sites.google.com/site/empirist2015/home/annotation-guidelines.

ble for online querying via a project server. Furthermore, it is planned to integrate the resource into the German Reference Corpus DeReKo at the Institute for the German Language in Mannheim (Lüngen, 2017).

6. References

Beißwenger, Michael (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik,* 41(1), pp. 161–164.

Beißwenger, Michael (2018). Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI. In Lobin, Henning, Schneider, Roman & Witt, Andreas (Eds.), *Digitale Infrastrukturen für die germanistische Forschung*. Berlin/New York: deGryuter, pp. 307–349.

Beißwenger, Michael; Bartsch, Sabine; Evert, Stefan; Würzner, Kay-Michael (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.* Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-2606), pp. 44–56. http://aclweb.org/anthology/W/W16/W16-2606.pdf

Beißwenger, Michael; Chanier, Thierry; Erjavec, Tomaž; Fišer, Darja; Herold, Axel; Lubešic, Nikola; Lüngen, Harald; Poudat, Céline; Stemle, Egon; Storrer, Angelika; Wigham, Ciara (2017a). Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In Lars Borin (Ed.), *Selected papers from the CLARIN Annual Conference 2016*. Aix-en-Provence, 26–28 October 2016: CLARIN Common Language Resources and Technology Infrastructure (Linköping University Electronic Conference Proceedings 136), pp. 1–18. http://www.ep.liu.se/ecp/contents.asp?issue=136

Beißwenger, Michael; Wigham, Ciara et al. (2017b). Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? In Egon W. Stemle & Ciara R. Wigham (Eds.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities* (cmccorpora 2017). Bolzano, Italy, Oct 03-04, 2017, pp. 52--55. https://cmc-corpora2017.eurac.edu/proceedings/

Chanier, Thierry; Poudat, Céline; Sagot, Benoit; Antoniadis, Georges; Wigham, Ciara; Hriba, Linda; Longhi, Julien; Seddah, Djamé (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

Dürscheid, Christa; Stark, Elisabeth (2011). sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In Thurlow, Crispin & Mroczek, Kristine (Eds.), *Digital Discourse. Language in the New Media*. Oxford, UK: Oxford University Press, pp. 299–320.

Frey, Jennifer-Carmen; Stemle, Egon W.; Glaznieks, Aivars (2014). Collecting Language Data of Non-Public Social Media Profiles. In Faaß, Gertrud & Ruppenhofer, Josef (Eds.), *Workshop Proceedings of the 12th Edition of the KONVENS Conference*. Hildesheim, Germany: Universitätsverlag Hildesheim, pp. 11–15.

Imo, Wolfgang (2015). Vom Happen zum Häppchen... Die Präferenz für inkrementelle Äußerungsproduktion in internetbasierten Messengerdiensten. *Networx 69*, pp. 1–35. http://www.mediensprache.net/de/networx/networx-69.aspx

Imo, Wolfgang (2017). Interaktionale Linguistik und die qualitative Erforschung computervermittelter Kommunikation. In Beißwenger, Michael (Ed.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin/New York: de Gruyter (Empirische Linguistik), pp. 81–108.

[JIM 2016] Medienpädagogischer Forschungsverbund Südwest (Ed.). Jugend, Information, (Multi-)Media. Basisuntersuchung zum Medienumgang 12-19-Jähriger. http://www.mpfs.de/de/studien/jim-studie/2016/

[KIM 2016] Medienpädagogischer Forschungsverbund Südwest (Ed.). KIM-Studie. Kindheit, Internet, Medien. Basisuntersuchung zum Medienumgang 6-13-Jähriger. www.mpfs.de/de/studien/kim-studie/2016/

Lüngen, Harald (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, Harald; Beißwenger, Michael; Herold, Axel; Storrer, Angelika (2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (Eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 156--164. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf

Lüngen, Harald; Beißwenger, Michael; Herzberg, Laura; Pichler, Cathrin (2017). Anonymisation of the Dortmund Chat Corpus 2.1. In Egon W. Stemle & Ciara R. Wigham (Eds.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora 2017)*. Bolzano, Italy, Oct 03-04, 2017, pp. 21–24. https://cmc-corpora2017.eurac.edu/proceedings/

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

The Effects of "Populist" Style on Tweet Popularity

Fabio Carrella
University for Foreigners – Perugia
fabio.carrella@unistrapg.it

Abstract

This research analyses linguistic features present in populist discourse on Twitter in order to understand whether these are related to the popularity of the tweets. Studies regarding populism and its language highlighted particular discursive characteristics ascribed to populist politicians such as emotionalization, simplified rhetoric and intensified claims. Moreover, much of the current literature focuses on how social media are linked to the recent populist surge, underlining how the former benefitted the latter. However, the influence that linguistic elements related to populism may have on its spread has not been fully investigated. This contribution examined tweets from four populist leaders, namely Luigi Di Maio, Matteo Salvini, Marine Le Pen and Nigel Farage, and from three establishment politicians, specifically Matteo Renzi, François Hollande and David Cameron. Using linear mixed effect models, we found several correlations between discursive features and the number of "favourites" and "retweets", both in populist and control tweets.

Keywords: CMC corpora, social media, political discourse, populism, statistical analysis, linear mixed-effect models

1. Introduction

In recent years, the growth of populism has been indicated by several important events, especially in the Western world. In 2016, the United Kingdom negotiated its withdrawal from the European Union through the well-known "Brexit" referendum, which was strongly advocated by the UK Independence Party (UKIP) and its former leader Nigel Farage. In the same year, the republican businessman Donald Trump surprisingly won the presidential election in the United States, defeating his democratic opponent Hillary Clinton. These two events seemed to trigger a domino effect. In France, the populist leader Marine Le Pen was one of the two top candidates in the final ballot of the 2017 presidential election. In Italy, the constitutional referendum supported by the centre-left (and pro-Europe) Democratic Party was won by the "No" side, which instead was promoted by populist groups (mostly Northern League, Brothers of Italy and Five Star Movement). It is also worth noting that Five Star Movement and Northern League were respectively the first and the third most voted parties in 2018 Italian general election. Finally, the Hungarian right-wing populist alliance Fidesz-KDNP, leaded by Viktor Orbán, triumphed in 2018 parliamentary election with almost 50% of the popular votes. The causes of this populist surge may be mainly identified in the reaction of the people to various social issues such as immigration, racism, terrorism and economic crisis. However, it is also revealing that a considerable part of the populist propaganda has been spread through social media (Bartlett, 2014; Gerbaudo, 2014; Engesser et al., 2017).

Therefore, this study aims to investigate possible correlations between the language used by four European populist politicians on Twitter and the popularity of their messages (or "tweets"). The idea is to understand whether some of the features that are peculiar to the populist style, such as emotionalization, intensifications and simplistic rhetoric, may have favoured the spread of their discourse on social media.

Initially, we created a corpus of 10,365 tweets collected from the official accounts of Luigi Di Maio, Matteo Salvini, Nigel Farage and Marine Le Pen. The linguistic analysis of the discursive features was conducted with the Appraisal framework (Martin & White, 2005). This operation allowed to observe the presence of emotional, simplistic and intensified elements, which confirmed previous findings related to populism and populist style (Canovan, 1999; Heinisch, 2008; Bos et al., 2011). We also collected 8,209 tweets from the accounts of Matteo Renzi, David Cameron and François Hollande, in order to create a reference corpus and to observe whether "populist" discursive features were also present in establishment politicians. Finally, we used linear mixed effect models (Bates et al., 2015) in order to examine possible significant correlations between the presence of specific linguistic elements and the popularity of each tweet, namely the sum of "favourites" and "retweets" given by users to each message.[1]

2. Literature Review

Social media seem to have served an important function in the spread of populism, especially in the Western world. For example, Twitter had a decisive part during Trump's presidential campaign (McCormick, 2016), while users in favour of leaving the EU during the Brexit referendum in 2016 outnumbered Remain supporters and behaved more actively on different social networks (Polonski, 2016; Hänska & Bauchowitz, 2017). In addition, it is revealing that populist leaders openly praise the importance of social

[1] This study is part of a larger doctoral investigation regarding the relationship between populism and social media analysed with various techniques. This means that there could be similarities between this proposal and other contributions by the same authors, whether published or forthcoming. However, this examination is separated from the others as it exploits different methods and brings new results.

media, as in the following tweets by Marine Le Pen and Nigel Farage:

(1) "Les réseaux sociaux permettent de s'adresser directement au peuple. Ma campagne sera innovante en ce domaine." (Le Pen, 2017)[2]

(2) "Without the internet, the development and growth of UKIP in Britain would have been far tougher." (Farage, 2016)

The relationship between politics and social media is reflected in the considerable number of studies regarding their use by politicians from all over the world (see Grant et al., 2010; Hong & Nadler, 2011; Broersma & Graham, 2012; Larsson & Kalsnes, 2014). However, as suggested by Bartlett (2014, p. 100), it seems that "[…] populist parties in Europe have been quicker to spot the opportunities these new technologies present to reach out and mobilize an increasingly disenchanted electorate." On one hand, websites as Twitter or Facebook allow populists to bypass censorship and journalistic filters, often considered untrustworthy (Mazzoleni, 2008). On the other hand, their communication style, often consisting of emotionalization, simplified rhetoric and spectacular claims (Canovan, 1999; Bos et al., 2011; Kramer, 2014), seems to be successful in grabbing users' ephemeral attention (Shoemaker & Cohen, 2006). Examples of these characteristics can be observed in the following tweets collected from the study subjects:

(3) "I now fear every attempt will be made to block or delay triggering Article 50. They have no idea of the level of public anger they will provoke." (Farage, 2016)

(4) "Ecco chi sono i veri razzisti! Le tivù lo censurano, fai girare tu." (Salvini, 2016)[3]

(5) " "La bataille que nous allons mener est la plus belle, la plus grande : la bataille pour la France !" #Brachay" (Le Pen, 2016)[4]

Moreover, in-group favouritism and out-group discrimination are often promoted by populist leaders through social media, resulting in the amplification of negative social attitudes through online phenomena as the filter-bubble effect (Pariser, 2011), or the echo-chamber effect (Jamieson & Cappella, 2008).

Hence, the relationship between populism and social media has been well examined. However, previous research has not fully investigated whether the popularity received by populist leaders and their messages online also depended on the presence of specific linguistic elements. Thus, it is hoped that this research will offer new insights related to social media, populism and politics in general, possibly providing reliable results thanks to the application of statistical methods.

3. Methodology

The objects of interest in this paper are four European populist leaders, specifically Luigi Di Maio, Matteo Salvini, Marine Le Pen and Nigel Farage[5]. These politicians have been chosen for several reasons. Firstly, they all have obtained important political victories, an aspect that should guarantee a minimum of popularity received by their tweets. In addition, the parties they belong to seem to share similar views, tending towards the right-wing spectrum of populism when issues such as immigration, Euro and the EU are concerned. This is also suggested by the political alliances between the four groups: for example, Five Star Movement and UKIP are both members of the Europe of Freedom and Direct Democracy Group (EFDD), a Eurosceptic coalition of the European Parliament. The same can be said for National Front and the Northern League, which constitute the majority of the Europe of Nations and Freedom Group (ENF), a right-wing populist alliance. In addition, they often mention each other online, as showed below:

(6) "Nel gruppo EFD con Farage, potremo votare insieme tutti gli altri gruppi ogni volta che vorremo. Anche con i Verdi. La rete ha fatto una scelta di libertà." (Di Maio, 2014)[6]

(7) "Bravo à notre ami @matteosalvinimi pour cette victoire du NON ! MLP #referendumcostituzionale" (Le Pen, 2016)[7]

(8) "#Salvini: ho condiviso fin dall'inizio idee e percorso di Trump, come di Putin e della Le Pen. #LIntervista" (Salvini, 2016)[8]

We also included three establishment politicians in the research, namely Matteo Renzi, François Hollande and David Cameron, in order to create a control group. This choice was adopted to better evaluate possible similarities or discrepancies between populist and non-populist authors. In addition, the fact that the three control subjects were Prime Ministers should hypothetically guarantee the presence of a "standard" political language, diametrically opposed to the populist style. After having selected the sample of politicians for the analysis, we collected tweets

[2] Trans: "Social networks allow to speak directly to the people. My campaign will be innovative in this domain."
[3] Trans: "Here's who the real racists are! The TV censors it, spread it."
[4] Trans: "The battle we are going to fight is the most beautiful, the greatest: the battle for France! #Brachay"
[5] We are aware that Nigel Farage is no longer the leader of UKIP. However, he still seems to have influence on both the party and his supporters (McCrum, 2017; Lowles, 2018; Cohen, 2018), and his tweets are still more popular than any current member of UKIP.
[6] Trans: "In the EFD group with Farage, we will be able to vote all the other groups together whenever we want. Even with the Greens. The network has made a choice of freedom."
[7] Trans: "Congratulations to our friend @matteosalvinimi for this NO victory! MLP #referendumcostituzionale"
[8] Trans: "#Salvini: I agreed with Trump's ideas and path from the beginning, as I did with Putin and Le Pen. #LIntervista"

from their official accounts using FireAnt, an application that gathers and organises tweets (Anthony & Hardaker, 2016). All tweets were divided in two main groups, resulting in a total of 10,365 messages for the populist corpus, and 8,209 messages for the reference corpus. FireAnt also allowed us to exclude retweets from both corpora, in order not to spoil the data with external authors' texts. The remaining tweets in both corpora were manually annotated by the authors using UAM CorpusTool (O'Donnell, 2011). Further information regarding the tweets can be retrieved from Table 1.

Name	Tweets	From	To
Luigi Di Maio	2,117	11/06/2014	02/03/2018
Matteo Salvini	2.871	26/05/2016	16/02/2017
Marine Le Pen	3,056	02/12/2015	16/02/2017
Nigel Farage	2,321	04/04/2015	16/02/2017
Matteo Renzi	2,622	20/11/2012	11/01/2017
François Hollande	3,225	12/02/2012	07/01/2018
David Cameron	2,362	06/10/2012	18/01/2017

Table 1: Tweets Information for Populist and Control Politicians

The theoretical model chosen to analyse the tweets was the Appraisal framework (Martin & White, 2005), as it appeared to be the most suitable option for the observation of the discursive features that characterise the populist style. The framework, based on the Systemic Functional Theory (Halliday et al., 2004), is a system characterised by three main nodes: attitude, engagement and graduation. These three elements adequately match the principal characteristics of the populist discourse, being emotionalization, simple rhetoric and intensified claims (Canovan, 1999; Heinisch, 2008; Bos et al., 2011). In detail, attitude regards all instances of emotions, judgements of human behaviour and aesthetic evaluations; engagement illustrates how authors negotiate the arguability of their utterances through the inclusion/exclusion of others' stances; graduation is related to the force and the intensity of a statement.

The final part of the study consisted of a statistical analysis regarding the correlation between discursive elements contained in the tweets and the popularity of each message. The former are considered to be the independent variables, while the latter is the dependent variable. We used R version 3.4.2 (2017) to create linear mixed effect models with the "lmertest" package (Kuznetsova et al., 2017), setting the politicians as a random effect in order to include possible discrepancies due to their different personal styles. Statistical methodologies such as linear (mixed) models are becoming a popular tool in all linguistics sub-fields (Gries, 2013; p. 4): we chose to use mixed effect models in order to observe if there were significant correlations between the linguistic features and the tweet popularity and, if so, to what extent.

Considering that each of the three main framework features (attitude, engagement and graduation) comprises a substantial number of sub-nodes, we decided to keep the most discrete ones in order to reduce the number of independent variables. Therefore, the variables included in the study were the following:

- Affect: emotional language such as fear, joy, hope, displeasure;

- Judgement: praise or criticism of human behaviour;

- Appreciation: judgements regarding state of affairs, artefacts or human aesthetics;

- Positive: a trait referring to affect, judgement or appreciation;

- Negative: a trait referring to affect, judgement or appreciation;

- Contract: suppression of divergent positions by the authors;

- Expand: acceptance of the existence of alternative assertions by the authors;

- Hashtag: metadata tag used on Twitter to group tweets and create user affiliation;

- Mention: metadata tag used on Twitter to address one or more particular users;

- Vigree: blend category which accounts for "vigour" and "degree", respectively indicating assessments of degree of intensity over processes or qualities;

- Repetition: lists of terms composed by the same lexical items or by closely related words;

- Graphical: emoticons, exclamation points or capital letters;

- Focus: graduation regarding the prototypicality of non-scalable terms;

- Quantification: scaling with respect to amount of size, weight, number or extent of time and space.

Moreover, we also decided to observe how the correlation of two independent variables might affect the tweet popularity. However, in order to reduce the total number of variables, we only paired those belonging to different systems. Therefore, we excluded correlations such as "Contract:Hashtag" because the two elements are two (opposite) sub-nodes of the "Engagement" system. The only exception regarded "Positive" and "Negative", which were paired with intra-system elements as well, since they are considered traits of "Affect", "Judgement" and "Appreciation", and not elements on their own.

4. Results

Results are presented for both populist and control groups. Due to the high number of variables included in the two models, we decided to graphically delete from the summaries the predictors that showed a p-value > 0.05.

However, these were still included in the model when the outcomes were processed by R. Figure 1 shows the results regarding the populist sample.

```
summary (PopulistModel)

REML criterion at convergence: 9708.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.6159 -0.7125 -0.0681  0.6591  4.6124

Random effects:
 Groups   Name        Variance Std.Dev.
 USER     (Intercept) 0.04771  0.2184
 Residual             0.14405  0.3795
Number of obs: 10367, groups:  USER, 4

Fixed effects:
                          Estimate Std. Error   t value   Pr(>|t|)
(Intercept)              2.544e+00  1.097e-01    23.183   0.000159 ***
AFFECT                   1.486e-01  3.503e-02     4.241    2.25e-05 ***
JUDGEMENT                9.849e-02  2.499e-02     3.942    8.15e-05 ***
NEGATIVE                 5.586e-02  2.302e-02     2.427    0.015257 *
CONTRACT                 6.041e-02  1.173e-02     5.151    2.64e-07 ***
EXPAND                   5.167e-02  1.384e-02     3.735    0.000189 ***
HASHTAG                 -1.698e-02  6.723e-03    -2.526    0.011567 *
MENTION                 -3.112e-02  1.155e-02    -2.695    0.007045 **
GRAPHICAL                5.515e-02  1.386e-02     3.980    6.94e-05 ***
FOCUS                    1.045e-01  4.771e-02     2.191    0.028487 *
AFFECT:HASHTAG          -3.582e-02  1.289e-02    -2.779    0.005463 **
AFFECT:MENTION          -4.970e-02  2.240e-02    -2.219    0.026527 *
AFFECT:FOCUS            -1.681e-01  8.563e-02    -1.963    0.049657 *
JUDGEMENT:CONTRACT      -2.657e-02  1.181e-02    -2.249    0.024551 *
JUDGEMENT:EXPAND         4.033e-02  1.674e-02     2.409    0.016013 *
JUDGEMENT:HASHTAG       -2.957e-02  7.374e-03    -4.010    6.11e-05 ***
JUDGEMENT:VIGREE        -3.269e-02  1.551e-02    -2.107    0.035125 *
JUDGEMENT:REPETITION    -9.645e-02  3.481e-02    -2.771    0.005594 **
JUDGEMENT:GRAPHICAL     -2.141e-02  1.034e-02    -2.070    0.038480 *
CONTRACT:VIGREE         -5.441e-02  2.030e-02    -2.681    0.007358 **
HASHTAG:QUANTIFICATION  -2.169e-02  1.028e-02    -2.110    0.034848 *
MENTION:GRAPHICAL        4.857e-02  2.080e-02     2.335    0.019542 *
MENTION:FOCUS           -1.737e-01  6.574e-02    -2.643    0.008239 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Summary of Populist model

As it can be observed, all the main significant variables, with the exception of "Hashtag" and "Mention", show positive estimates, meaning that they positively affect the tweet popularity when they are included in the message. On the contrary, the majority of the correlations, apart from "Judgement:Expand" and "Mention:Graphical", negatively characterize the number of favourites and retweets obtained by the tweet. Focusing on the estimates, the largest positive value is related to "Affect" (0.148), followed by "Focus" (0.104) and "Judgement" (0.098). On the other hand, the largest negative estimates are showed by "Mention:Focus" (-0.173), "Affect:Focus" (-0.168) and "Judgement:Repetition" (-0.096).

Next, outcomes regarding the control group are presented in Figure 2. Although the number and the types of significant variables may appear similar between the two models, the most noticeable difference is the size of the estimates. Here, the largest positive estimate represented by "Hashtag:Graphical" amounts to 0.338. Then, we find "Judgement:Graphical" (0.321) and "Repetition" (0.21). Even larger estimates sizes are detected on the negative side, with "Mention" having a value of -0.504, followed by "Expand:Graphical" (-0.421) and "Focus" (-0.404).

```
summary(ControlGroupModel)

REML criterion at convergence: 15184.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9157 -0.6420 -0.1066  0.6267  5.1860

Random effects:
 Groups   Name        Variance Std.Dev.
 USER     (Intercept) 0.1211   0.3481
 Residual             0.3620   0.6016
Number of obs: 8185, groups:  USER, 3

Fixed effects:
                          Estimate Std. Error  t value   Pr(>|t|)
(Intercept)              2.344e+00  2.018e-01   11.612   0.006957 **
AFFECT                   1.543e-01  5.604e-02    2.753   0.005913 **
POSITIVE                 1.377e-01  4.128e-02    3.335   0.000856 ***
NEGATIVE                 2.058e-01  4.723e-02    4.357   1.34e-05 ***
EXPAND                  -4.243e-02  2.107e-02   -2.013   0.044108 *
HASHTAG                 -1.130e-01  1.282e-02   -8.816    < 2e-16 ***
MENTION                 -5.048e-01  1.752e-02  -28.812    < 2e-16 ***
VIGREE                   1.342e-01  3.578e-02    3.751   0.000177 ***
REPETITION               2.108e-01  9.198e-02    2.291   0.021971 *
FOCUS                   -4.043e-01  1.477e-01   -2.738   0.006186 **
AFFECT:HASHTAG          -1.383e-01  2.014e-02   -6.865   7.13e-12 ***
AFFECT:MENTION           6.877e-02  3.059e-02    2.248   0.024613 *
JUDGEMENT:HASHTAG       -4.187e-02  1.615e-02   -2.592   0.009559 **
JUDGEMENT:MENTION        1.035e-01  2.510e-02    4.123   3.78e-05 ***
JUDGEMENT:GRAPHICAL      3.217e-01  1.089e-01    2.955   0.003136 **
APPRECIATION:MENTION     1.118e-01  2.336e-02    4.785   1.74e-06 ***
CONTRACT:GRAPHICAL      -3.374e-01  9.207e-02   -3.665   0.000249 ***
EXPAND:GRAPHICAL        -4.216e-01  1.585e-01   -2.659   0.007844 **
HASHTAG:REPETITION      -1.798e-01  5.057e-02   -3.556   0.000379 ***
HASHTAG:GRAPHICAL        3.383e-01  8.625e-02    3.922   8.84e-05 ***
HASHTAG:FOCUS            2.022e-01  8.163e-02    2.478   0.013245 *
MENTION:GRAPHICAL        1.795e-01  7.313e-02    2.455   0.014117 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: Summary of Control Group model

5. Conclusion

Results for both models show a positive correlation between the presence of "Attitude" features (i.e. emotions, judgements) and the popularity of the messages, thus confirming similar previous findings (Zappavigna, 2011; Stieglitz & Dang-Xuan, 2014). Surprisingly, significant "Engagement" features negatively affect the tweet popularity, with the exception of "Expand" and "Contract" in the populist sample. In particular, "Hashtag" and "Mention" show high negative estimates in both groups. A possible explanation might be that hashtags are often included in tweets promoting political campaigns, while mentions usually address specific users on Twitter. Therefore, tweets containing these two elements may not raise enough interest in the general audience. With regard to "Graduation", results between and within the groups are inconsistent. In the populist sample, "Graphical" and "Focus" positively affect tweet popularity. However, when these two elements are related to other variables, as in "Affect:Focus" or "Judgement:Graphical", the popularity tend to decrease. In the control group, graduation elements are found on the opposite sides of the estimates spectrum: for example, "Graphical" helps to obtain more interactions when combined with "Judgement", "Hashtag" and "Mention", but behaves differently with "Contract" and "Expand". To conclude, the two groups show rather similar behaviours. More importantly, establishment politicians seem to be more affected by certain linguistic elements when tweet popularity is concerned.

6. References

Anthony, L., Hardaker, C. (2017). *Fireant (Version 1.1.4). Computer Software.* Tokyo: Waseda University. Retrieved from http://www.laurenceanthony.net/software

Bartlett, J. (2014). Populism, Social Media and Democratic Strain. In C. Sandelind (Ed.), *European Populism and Winning the Immigration Debate*. Stockholm: FORES, pp. 99--114.

Bates, D., Machler, M., Bolker, S., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), pp. 1--48. doi: http://dx.doi.org/10.18637/jss.v067.i01.

Bos, L., Van der Brug, W. & De Vreese, C. (2011). How The Media Shape Perceptions of Right-Wing Populist Leaders. *Political Communication,* 28(2): pp. 182--206. https://doi:10.1080/10584609.2011.564605.

Broersma, M., & Graham, T. (2012). Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice*, 6(3), pp. 403--419.

Canovan, M. (1999). Trust the People! Populism and the Two Faces of Democracy. *Political Studies,* 47(1), pp. 2--16.

Di Maio, L. [LuigiDiMaio]. (2014, June 12). Nel gruppo EFD con Farage, potremo votare insieme tutti gli altri gruppi ogni volta che vorremo. Anche con i Verdi. La rete ha fatto una scelta di libertà. [Facebook Status Update]. Retrieved from https://www.facebook.com/luigidimaio/posts/691590100877539

Engesser, S., Ernst, N., Esser, F. & Büchel, F. (2017). Populism and Social Media: How Politicians Spread a Fragmented Ideoogy. *Information, Communication & Society,* 20(8), pp. 1109--1126. https://doi:10.1080/1369118x.2016.1207697.

Farage, N. [Nigel_Farage]. (2016, April 06). Without the internet, the development and growth of UKIP in Britain would have been far tougher. [Tweet]. Retrieved from https://twitter.com/Nigel_Farage/status/717672818423566337

Farage, N. [Nigel_Farage]. (2016, November 3). I now fear every attempt will be made to block or delay triggering Article 50. They have no idea level of public anger they will provoke. [Tweet]. Retrieved from https://twitter.com/Nigel_Farage/status/794124251259432960

Gerbaudo, P. (2014). Populism 2.0: Social Media Activism, the Generic Internet User and Interactive Direct Democracy. In D. Trottier & C. Fuchs (Eds.), *Social Media, Politics and the State: Protests, Revolutions, Riots, Crime and Policing in the Age of Facebook, Twitter and Youtube*, New York, NY: Routledge, pp. 67--87.

Grant, W. J., Moon, B., & Busby Grant, J. (2010). Digital dialogue? Australian politicians' use of the social network tool Twitter. *Australian Journal of Political Science*, 45(4), pp. 579--604.

Gries, S. (2013). *Statistics for Linguistics with R*. Berlin: De Gruyter Mouton.

Halliday, M. A. K., Matthiessen, C. & Halliday, M. (2004). *An Introduction to Functional Grammar*, revised by Christian Matthiessen. London: Hodder Arnold.

Hänska, M., & Bauchowitz, S. (2017). Tweeting for Brexit: How Social Media Influenced the Referendum. In J. Mair, T. Clark, R. Snoddy & R. Tait (Eds.), *Brexit, Trump and the Media*. Suffolk: Abramis, pp. 31—35.

Heinisch, R. (2008). Austria: The Structure and Agency of Austrian Populism. In D. Albertazzi & D. McDonnell (Eds.), *Twenty-first Century Populism*. Basingstoke, UK: Palgrave Macmillan, pp. 67--83.

Hong, S., & Nadler, D. (2011). Does the early bird move the polls?: The use of the social media tool 'Twitter' by US politicians and its impact on public opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times.* New York, NY: ACM, pp. 182—186.

Jamieson, K. H., & Cappella, J. (2008). *Echo chamber: Rush Limbaugh and the Conservative Media Establishment*. New York, NY: Oxford University Press.

Krämer, B. (2014). Media Populism: A Conceptual Clarification and some Theses on its Effects. *Communication Theory,* 24, pp. 42-60. https://doi: 10.1111/comt.12029.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software,* 82(13), pp. 1-26. doi:10.18637/jss.v082.i13

Larsson, A. O., & Kalsnes, B. (2014). 'Of course we are on Facebook': Use and non-use of social media among Swedish and Norwegian politicians. *European Journal of Communication*, 29(6), pp. 653--667.

Le Pen, M. [MLP_officiel]. (2016, September 3). "La bataille que nous allons mener est la plus belle, la plus grande : la bataille pour la France !" #Brachay. [Tweet]. Retrieved from https://twitter.com/MLP_officiel/status/772021712913399808

Le Pen, M. [MLP_officiel]. (2016, December 04). Bravo à notre ami @matteosalvinimi pour cette victoire du NON ! MLP #referendumcostituzionale [Tweet]. Retrieved from https://twitter.com/mlp_officiel/status/805545998038761472

Le Pen, M. [MLP_officiel]. (2017, January 04). Les réseaux sociaux permettent de s'adresser directement au peuple. Ma campagne sera innovante en ce domaine. #VoeuxMLP. [Tweet]. Retrieved from https://twitter.com/mlp_officiel/status/816588935711379456

Martin, J. R., & White, P. R. (2005). *The Language of Evaluation*. Basingstoke, UK: Palgrave Macmillan.

Mazzoleni, G. (2008). Populism and the Media. In D. Albertazzi & D. McDonnell (Eds.), *Twenty-first Century Populism*. Basingstoke, UK: Palgrave Macmillan, pp. 49--64.

McCormick, R. (2016, November). Donald Trump says Facebook and Twitter 'helped him win.' *The Verge.* Retrieved from https://www.theverge.com/2016/11/13/13619148/trump-facebook-twitter-helped-win

O'Donnell, M. (2011). *UAM CorpusTool* (Version 2.8.7). Computer Software. Retrieved from http://www.corpustool.com/index.html

Pariser, E. (2011). *The Filter Bubble: What the Internet is hiding from you*. New York, NY: Penguin.

Polonski, V. (2006). Impact of Social Media on the Outcome of the EU Referendum. Accessed April 30, 2018. Retrieved from https://www.referendumanalysis.eu/eu-referendum-analysis-2016/section-7-social-media/impact-of-social-media-on-the-outcome-of-the-eu-referendum/

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Salvini, M. [matteosalvinimi]. (2016, June 25). Ecco chi sono i veri razzisti! Le tivù lo censurano, fai girare tu. [Tweet]. Retrieved from https://twitter.com/matteosalvinimi/status/746642514598367233

Salvini, M. [matteosalvinimi]. (2016, November 13). #Salvini: ho condiviso fin dall'inizio idee e percorso di Trump, come di Putin e della Le Pen. #LIntervista [Tweet]. Retrieved from https://twitter.com/matteosalvinimi/status/797753669278138368

Shoemaker, P., & Cohen, A. A. (2006). *News Around the World Content: Practitioners, and the Public*. New York, NY: Routledge.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems,* 29(4), pp. 217--248.

Zappavigna, M. (2011). Ambient Affiliation: A Linguistic Perspective on Twitter. *New Media & Society,* 13(5), pp. 788--806. https://doi.org/10.1177/1461444810385097.

Why Did Nobody Reply to Me? A Keyword Analysis of Initiating Posts and Lone Posts in Massive Open Online Courses (MOOCs) Discussions

Shi Min Chua
The Open University
shimin.chua@open.ac.uk

Abstract

It is a common phenomenon that online discussion spaces are overabundant with lone posts; in other words, few posts receive replies from others. Admittedly, circumstantial factors and content affect whether a post receives replies. Yet, linguistic features within a post might also play a role in inviting replies. To investigate this hypothesis, a keyword analysis comparing initiating posts, which receive replies, to lone posts, which do not receive replies, was conducted. The posts were from the discussion in massive open online courses (MOOCs). MOOC discussion is one type of computer-mediated communication (CMC), with an emphasis on learning and is typically monitored by course facilitators. The keyword analysis revealed that initiating posts were often constructed in a question format, with hedges and indefinite pronouns to open up a dialogue and invite others to pitch in, whereas lone posts tended to be reflective and monoglossic in nature, yet with positive sentiments.

Keywords: keyword analysis, online discussion, MOOC

1. Introduction

Online discussion spaces, such as Usenet groups (Burke et al., 2007; Himelboim, Gleave, & Smith, 2009), online news commenting spaces (Ziegele, Breiner, & Quiring, 2014) and distance learning online discussion (Dennen & Wieland, 2007), tend to have a huge number of lone posts. Lone posts are new posts that do not receive any replies, in comparison to initiating posts that do (Chua et al., 2017). Several reasons could account for the overabundance of lone posts in online discussion spaces. On one hand, internet users tend to create new posts, rather than replying to others, given that the online space is a levelling ground that allows users to say what they want to say instead of being obliged to respond to others as in a face-to-face conversation (Cavanagh, 2007). On the other hand, circumstantial factors such as timing of posting and design of the online discussion space, as well as the content of the posts may render a post less likely to be read, thus receiving no replies (Ziegele et al., 2014).

Besides these factors, it is possible that the discourse of the lone posts may be less dialogic than the initiating posts. Initiating posts could be constructed to create interaction with readers, thus inviting replies (Martin & White, 2005). To examine this hypothesis, linguistic features of lone posts and initiating posts were investigated in this study through keyword analysis. The posts comprising the corpus were taken from the discussions in Massive Open Online Courses (MOOCs) on FutureLearn[1]. On FutureLearn, learners can post their comments on almost every learning step/page, analogous to users' comments that appear below news articles published online. The discussion space in FutureLearn MOOCs is one kind of asynchronous computer-mediated communication (CMC). Yet, it differs from online forum or online news commenting spaces because it is set in a learning context and is often monitored by facilitators (Ferguson & Sharples, 2014).

2. Online Discussion as a Dialogic Space

Online discussion, whether in MOOCs or other settings, can be operationalized as a dialogic space, which can be shaped by technological affordances, learning activities, content, and language (Wegerif, 2010; Ziegele et al., 2014). Education researchers have proposed that a dialogic space is one that promotes reflection and thinking (Wegerif, 2010), exploratory talk (Mercer, 2004) and co-construction of meaning (Littleton & Whitelock, 2005). The present paper focuses on how one factor-linguistic resources, could shape such a space in MOOCs. According to Wegerif (2010) and White (2003), a dialogic space can be shaped by linguistic resources that create:

1. intersubjectivity such that subjectivity and stances of each user could be shared and negotiated (Chandrasegaran & Kong, 2007; Dennen & Wieland, 2007; Du Bois & Kärkkäinen, 2012);
2. heteroglossia such that multiple voices, whether anticipated views, alternative views or views that have been stated, are considered (Bakhtin, 1983);
3. intertextuality such that different sources of contents or others' utterances are referred to (Bakhtin, 1983);
4. politeness (Brown & Levinson, 1987) and interpersonal relationship in a community. (Lander, 2015);
5. personal agency for each participation (Al Zidjaly, 2009; Wagner & Herbel-Eisenmann, 2008)

Various linguistic features and grammatical structures can be used to open up and expand a dialogic space. For example, internet users could use linguistic features such as epistemic modality or hedges (e.g., *might, probably, I guess*) to qualify or mitigate their propositions by expressing their

[1] www.futurelearn.com

attitude, confidence, uncertainty or source of evidence (Hyland, 2005; Stubbs, 1986). Constructions with these linguistic features, in contrast to categorical or bare assertions, provide space for alternative voices, thus inviting others' contributions. Previous research has been fruitful in revealing the pragmatic and discourse functions of various lexical devices and grammatical constructions in relation to intersubjectivity, heteroglossia, intertextuality and politeness (e.g., Biber et al., 1999).

Nonetheless, it is generally agreed among researchers (e.g., Du Bois & Kärkkäinen, 2012) that a one-to-one mapping between word forms and functions is not possible because a linguistic feature can carry multiple functions, and the textual and social context can affect its interpretation. Therefore, in the present study, instead of comparing the word frequencies of a fixed list of linguistic features found in initiating posts and lone posts, a corpus- or data-driven approach (keyword analysis) is first utilized to reveal the linguistic features that are used significantly more often in initiating posts and lone posts respectively. Then, the keywords are subjected to discourse analysis and interpretation in the light of theories around dialogic space and MOOC learning.

3. Present Study

MOOCs are typically offered free to anyone around the world, thus attracting massive numbers of learners and discussion postings. This sheer massiveness may reduce the chance for learners to engage in repeated exchange with each other in discussions (Eynon et al., 2016), and may also lower the probability of a post being read and replied to. In MOOCs, learners may feel frustrated if their posts are seldom responded to (Hew & Cheung, 2014). In other online spaces, users were found to join the discussion for interactive purposes rather than cognitive gains (Springer, Engelmann, & Pfaffinger, 2015), and newcomers were more likely to continue their participation in the group if they received replies to their posts (Joyce & Kraut, 2006). It is therefore important to understand why only some posts receive replies. Nonetheless, MOOC discussion space may differ from other online discussion spaces in that it is not only an interactive space but also a channel for learners to reflect on the learning materials themselves (Laurillard, 2012). It is therefore important to understand the nature of the lone posts as well as the initiating posts in this particular context.

4. Methods

4.1 Corpus

The corpus consists of discussion posts from 12 MOOCs on the FutureLearn platform. Because the present study focuses only on the lone posts and initiating posts, the replies they receive are not included in the corpus. Furthermore, educators and facilitators' postings are also excluded because their language use might differ from learners' given their instructional role on the platform. The total number of lone posts and initiating posts in the corpus are 117,863 and 32,080 respectively, with 6,162,230 and 2,401,795 tokens each. In this corpus, the number of lone posts number almost four times as many as the initiating posts. As a reference, there are 54,172 replies, which is about half the number of the lone posts.

4.2 Keyword Analysis

Keyword analysis was conducted to compare lone posts with initiating posts to examine the difference in linguistic features between these two types of post. The statistical measure used for the keyword analysis was the log-likelihood ratio test, which has the benefit of not being biased by huge sample size differences between the two comparison (sub)corpora (Rayson & Garside, 2000). A word is considered a keyword when the *p*-value for the log-likelihood ratio test is < 0.000000000001 (Flowerdew, 2015). In addition, the effect size indicator Bayes factor must be > 10 (Wilson, 2013), and the normalized frequency must be 5 per 100,000 following McEnery (2016), in order to ensure the keyword is a common word in the corpus. Lastly, the dispersion measure, Gries' Deviation of Proportion (Lijffijt & Gries, 2012), of each word must be smaller than 0.30 to ensure that the keyword is evenly distributed across the 12 courses.

4.3 Analysis of Keywords

The keyword analysis revealed 70 keywords that were used significantly more often in the initiating posts than in the lone posts, while 77 keywords were used more frequently in the lone posts than in the initiating posts. These keywords were then labelled for their function by examining the collocations and concordance lines of the keywords. In cases where this distant reading did not provide insight into the function of the keyword, a randomly selected 100-150 posts containing the keywords were subjected to close reading. As mentioned earlier, a word can have more than one meaning or function, thus the label applied represents only the most salient function of the keyword in the corpus (McEnery, 2016). In other words, the labelling is based where possible on the function of the keyword in the MOOC discussion under examination. The labelling of the keyword functions was decided with reference to Biber et al. (1999) and Rayson (2008).

5. Findings

The keywords and their labels are shown in Table 1. It emerged that a major group of keywords were found to be used for stance expression, which according to Du Bois and Kärkkäinen (2012), was related to intersubjectivity, so they were labelled based on this discourse function. Discourse particles and meta-language were also labelled respectively because of their salient discourse function in the corpus. For example, although *question* could be used as a verb to realise a speech act, it was used mainly as a noun and meta-language in the corpus, as in *...the big question is...* Other keywords were labelled mainly according to their grammatical function because, while their use in the corpus

Categories	Initiating Posts	Lone Posts
Stance Expression		
Modals/Modal expression	might, would could	will, need, able
Hedging	perhaps, seems, sort[2]	
Quantifier	any	all, lot, much, every
Booster	surely, just, rather, else	really, very, definitely, always
Epistemic expression	wonder, wondering	aware, understanding, learned
Mental verbs		feel, feeling, think, agree, keen, hope, hoping, looking, forward, enjoy, enjoyed, love
Evaluative	wrong	difficult, easy, excellent, better, interesting, informative, great, important, good, new
Negation	cannot, ca[3], n't[4]	
Others		
Discourse particles	please, sorry	thanks, thank
Meta-language	question, article	information, course[5], knowledge
Pronouns	he	I, my, our, their
Indefinite pronouns	anybody, anyone	everyone
Connectors	if, or, then, example, e.g.	also, and
Comparative terms/relational	than, same	more
Grammatical	the, that, there, here, does, did, was, were, 's, on, by	am, 'm, have, for, about, with, to
Punctuation	,…-();?'":	!.
Speech act	mean, explain, tell, says, say, told, called	
Verbs in past tense/passive form	used, tried, came	joined
Verbs in present tense/infinite form		affects, helps, achieve, work , gain, meet, improve
Uncategorized	1, one, two[6], numbers, missing, following, why, whether	like, well, week, main, currently, working, opportunity, education, environment, mind

Table 1: Keywords in initiating posts and lone posts.

was taken into account, no one salient semantic meaning or function emerged. Among these keywords, there were three groups of lexical verbs, speech acts, present/infinite and past/passive verb forms. Because the communicative functions of the latter two groups of verbs could not be identified, they were labelled by their grammatical form, which is their shared characteristic.

There was one group of keywords labelled as grammatical, because they are grammatical or functional words involving in a wide range of communicative functions which cannot easily be categorised. Additionally, their high frequencies in the corpus also rendered an in-depth analysis of their function impossible. So they were conveniently grouped together. Admittedly, there could be different functions within this group, for example *does, did, was, were* are primary verbs whereas *here* and *there* could be deitic (Biber et al., 1999).

Lastly, a group of keywords were uncategorized because their most salient function could not be determined. Some carried multiple meanings and functions in the corpus. Examples include *well* in *female as well as male* and *feeling well*, and *like* in *I'd like to* and *it sounds like*. Other uncategorized keywords were labelled as such because they were the only keyword with a specific label, for example *why* was the only *wh*-question word as a keyword in the corpus, and *week* was the only referent to time.

In the next section, due to space constraints, only selected keywords that are relevant to dialogic spaces and MOOC learning are elaborated on.

5.1 Indefinite Pronouns

The indefinite pronouns *anybody* and *anyone*, which appeared as keywords in the initiating posts, were often used in questions to address other learners whose names are not known, or when there are simply too many people to address individually. For example *…So, does **anybody** have a good suggestion for a text book on…* and*…has **anyone** else come across this…* This usage of *anybody* and *anyone* is in contrast to the frequent usage of *you* in one-to-one text messaging in social contexts (Tagg, 2012) which may be more targeted and personalized. Yet, in the MOOC context, these indefinite pronouns open up space and provide agency to learners who would like to respond to the initiating posts. These two keywords also suggest that learners do not only orient towards facilitators but also other learners in their learning process.

In contrast, *everybody*, which was a keyword in the lone posts, was used in an all-inclusive way (Biber et al., 1999), as in *...we need **everyone** to control our daily waste…* and *...not **everyone** could afford them...* in order to take a strong stance. It was mainly used for greetings such as *Hi **everyone**…*, and for showing appreciation, as in *Thanks **everyone**…*

[2] 70% of the instances of *sort* collocated with *of*, forming the hedging expression *sort of*.

[3] *ca* is a token of *can* and resulted from the tokenization of *can't* into *ca* and *n't*. The tokenization was done by the treetagger (Schmid, 1994) used in the present study.

[4] *n't* resulted from the tokenization of *don't*, *can't*, *didn't*, *doesn't*, *isn't*, *couldn't*, *wouldn't*, *wasn't*, *haven't*, *won't*, *aren't*, *hadn't*, *hasn't*, *weren't*.

[5] *Course* was mainly used by learners to refer to the online course they were taking, as in *…looking forward to this course…,* although 8% of the instances were in *of course.*

[6] *1*, *one*, and *two* arguably function as quantifiers as well, but they differed from the other quantifiers in the sense that they are numerals that specify exact amount (Biber et al, 1999) and do not have the intensifying or down-toning function in stance expression.

5.2 Connectors

The connectors *if*, *example* and *e.g.* were all keywords in the initiating posts. All three could be said to qualify or elaborate on a proposition by specifying a condition, as in *...unless the development damages the land (**e.g.** excessive clearing...* and *...enhances your feeling of well being, **if** it is mutual but **if** it is unrequited...*, or by raising alternatives, such as *...Here I have an **example** of a vocabulary exercise which I came across earlier...* This qualification of a proposition provides details for others to understand or comment on and avoids sweeping generalizations that allow no space for discussion. Furthermore, *if* could also be used for politeness purposes to hedge an argument, as in *...**if** you think about it, this is far more...*

In the lone posts, *also* and *and*, which are normally used to connect similar ideas (Halliday & Matthiessen, 2014), were found to be keywords. This could be an indication that in lone posts, learners tended to pool ideas, without elaboration or specification (Dennen & Wieland, 2007). This is in contrast to initiating posts where *if*, *example* and *e.g.* were used to qualify proposition.

5.3 Stance Expression

As mentioned in the introduction, modals, hedges and boosters are typically used to intensify or minimize a speaker's or author's commitment to what they are saying in terms of the level of knowing, certainty, obligation, prediction or truth (Stubbs, 1986). The keywords found in the initiating posts, *seems*, *perhaps*, *might*, *would* and *could*, which are on the less certain end of the continuum (Biber et al., 1999), can serve to hedge one's ideas and invite others to fill the dialogic space with alternative opinions. For example, *...this is **perhaps** because we tend to...* and *...This **might** mean actually walking...* Furthermore, another two keywords, *wonder* and *wondering*, were also typically used in rhetoric questions where learners expressed uncertainty in their understanding, as in *...I **wonder** would the microbial diversity also mirror...* These linguistic features not only help express one's stance, but also invite the expression of others' stances, thus potentially facilitating intersubjectivity among learners.

Unexpectedly, keywords expressing a strong stance such as *just*[7], *surely*, *wrong*, *rather*, and negations, *n't* and *cannot* were also used frequently in the initiating posts, for example, *...I really do**n't** see the point of...* This is probably because strong negative views might be controversial and thus trigger responses from others (Chen & Chiu, 2008; Himelboim et al., 2009). In contrast, *think* which was mostly used in *I **think***, was more frequently used in the lone posts. The reading of concordance lines revealed that *I think* was commonly used in learners' responses towards discussion prompts or questions that were mentioned in the learning materials. Similarly, *agree*[8] was also used frequently to express agreement towards what had been mentioned in the learning materials, *...I **agree** with this definition regarding health...* or with what other learners had said in the discussions, *...I **agree** with many of the posts...* Both negation and agreement can be an intertextual acknowledgement of what has been discussed in the dialogic space (Dennen & Wieland, 2007).

Other keywords for the expression of stance found in the lone posts were those boosting a speaker's or author's stance through their semantic meaning of entirety (Rayson, 2008), such as *always*, *every*, and *all* in *...money taken in by a Company is not **all** down to their own effort, it relies on...*The semi-modal *need* which conveys obligation, was also used more often in the lone posts, *...We **need** to be more exact...* and *...I need to be ambidextrous...* The sweeping meaning of *all* and *we need* could prove face-threatening, thus inhibiting others from opposing and exploiting the space for other alternative voices. *I need* can be seen as an assertive personal resolution that is not intended to invite others to comment.

Lastly, the boosters, *really*, *very*, *definitely*, that were used more frequently in the lone posts, tended to collocate with expression of emotion, as explained in the next section.

5.4 Expression of Emotion, Appreciation and Reflection

In the lone posts, keywords for evaluation, *excellent, interesting*, *informative*, *great*, keywords for emotion expression, *keen*, *hope*, *hoping*, *looking forward*, *enjoy*, *enjoyed*, along with the boosters mentioned above, exclamation mark and discourse particles *thank(s)*, pointed primarily to the positive sentiments expressed by learners. Most of the positive sentiments constituted personal reflections on what the students want to learn, as in *...**Really looking forward** to learn...*, or on what they have learnt, *...I **enjoyed** this course and **definitely** learned a lot in...*, as prompted by the learning activities at the start and end of each course. These reflections as well as expressions of gratitude to the course educators, ***Excellent** range of resources, **thanks**!* may not be written with the intention of inviting responses, but serve as a public expression of stance and emotion. The first person pronouns *I*, *my*, *our* and the epistemic expressions, *understanding*, *aware* and *learned*, which were also used more often in the lone posts, also suggest the reflective nature of these posts. In the initiating posts, keywords with similar functions were not found.

5.5 Questions and Requests

In the initiating posts, the keyword *question*, the discourse particle *please*, question mark, and the indefinite pronouns *anybody*, *anyone*, as well as *wonder* and *wondering*, seemed to suggest that questions and requests were frequently constructed. The use of the keyword *question* may serve to attract others' attention, as in ***Question**: does **anybody** knows what kind ...**?**...* It was sometimes also used to refer to a concept under discussion, *...the **question** of sustainability needs...* or to refer to a specific question in the quiz, *...I noticed the Quiz **question** 3...* Intriguingly, among all the *wh*-words, only *why* is a keyword, perhaps because *why*-questions can trigger various speculations

[7] *just* could also be used as a hedging device as in *...I just want to say...*
[8] Only 2% of *agree* collocated with *n't* in the lone posts.

from others, thus creating a space for multiple voices. In lone posts, keywords with similar functions were not found.

6. Discussion and Conclusions

The keyword analysis indeed revealed a difference in the discourse between initiating posts and lone posts. This preliminary analysis showed that initiating posts were often constructed in a question format. In initiating posts, learners often used *anyone* or *anybody* to invite others to join the dialogic space. Their use of modals and hedges for mitigation also creates a less face-threatening space for others to join in (Lander, 2015; Martin & White, 2005). *If*-conditionals and *example* were also used to create a dialogic space through specifying an elaborated scenario in relation to their proposition (Dennen & Wieland, 2007).

Unexpectedly, the strong negation in the initiating posts also seemed to attract replies. This is in contrast to lone posts which expressed agreement or appreciation in a reflective way; that is, through personal pronouns, mental verbs, positive evaluative words, and *thank(s)*. Because reflective writing is often monologic and single-voice rather than heteroglossic, so this kind of post might not encourage replies. Yet, this reflective writing in the lone posts was in line with one of the education purposes in online discussion-reflection and thinking (Laurillard, 2012; Wegerif, 2010). Additionally, the positivity created by these posts may have helped create a positive learning environment (Lander, 2015; Walsh & Li, 2013), even though the number of such posts could sometimes be overwhelming.

Lastly, the occurrence of disagreement or agreement towards learning materials and other learners in lone posts and initiating posts suggests that learners engaged in intertextuality and heteroglossia even though they were not writing a reply towards a specific post (Dennen & Wieland, 2007). Perhaps the disagreement towards course content expressed in the initiating posts was raising another voice, so potentially opening up a dialogic space. In contrast, an agreement towards course content or other comments without targeting a specific learner, as expressed in the lone posts, could be deemed as an addition to a pool of similar ideas (Dennen & Wieland, 2007), similar to the cumulative talk that Mercer (2004) identified.

Admittedly, this keyword analysis is quantitative and exploratory in nature. The categorization of keywords provides only a broad picture of the typical linguistic features used in each type of posts. Additionally, it should be noted that keywords in one type of post were also used in the other type, but were used less often and could be for other functions that have not been explored. Given that the function and meaning of each word largely depends on the context it appears in, further in-depth discourse analysis of selected keywords, as well as full conversation threads including the replies that were not examined in the present study, should reveal how each linguistic feature opens up or closes down dialogic spaces.

7. Implications

The findings of this keyword analysis could inform MOOCs learners about how to construct their posts and what to expect in terms of responses to their posts. To engage with others, learners could try to construct their posts as questions, with hedges and indefinite pronouns. They could also be reminded that not receiving a reply to their reflective or appreciative posts should not be seen as a disappointment but reflects wider trends across this type of discussion forum.

8. References

Al Zidjaly, N. (2009). Agency as an interactive achievement. *Language in Society*, 38(2), pp. 177–200. https://doi.org/10.1017/S0047404509090320

Bakhtin, M. M. (1983). *The Dialogic Imagination: Four Essays by M. M. Bakhtin. Contemporary Sociology* (Vol. 12). https://doi.org/10.2307/2068977

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, United Kingdom: Pearson Education Limited.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Burke, M., Joyce, E., Kim, T., Anand, V., & Kraut, R. (2007). Introductions and requests: Rhetorical strategies that elicit response in online communities. In *Proceedings of the 3rd Communities and Technologies Conference, C and T 2007* (pp. 21–39). https://doi.org/10.1007/978-1-84628-905-7_2

Cavanagh, A. (2007). *Sociology in the Age of the Internet*. Maidenhead: Open University Press.

Chandrasegaran, A., & Kong, K. M. C. (2007). Stance-taking and stance-support in students' online forum discussion. *Linguistics and Education*, 17(4), pp. 374–390. https://doi.org/10.1016/j.linged.2007.01.003

Chen, G., & Chiu, M. M. (2008). Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Computers and Education*, 50(3), pp. 678–692. https://doi.org/10.1016/j.compedu.2006.07.007

Chua, S. M., Tagg, C., Sharples, M., & Rienties, B. (2017). Discussion Analytics: Identifying Conversations and Social Learners in FutureLearn MOOCs. In *LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference, FutureLearn Workshop* (pp. 36–62).

Dennen, V. P., & Wieland, K. (2007). From interaction to intersubjectivity: Facilitating online group discourse processes. *Distance Education*, 28(3), pp. 281–297. https://doi.org/10.1080/01587910701611328

Du Bois, J. W., & Kärkkäinen, E. (2012). Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction. *Text and Talk*, 32(4), pp. 433–451. https://doi.org/10.1515/text-2012-0021

Eynon, R., Hjorth, I., Yasseri, T., & Gillani, N. (2016). Understanding Communication Patterns in MOOCs: Combining Data Mining and qualitative methods. In S. ElAtia, D. Ipperciel, & O. Zaïane (Eds.), *Data Mining and Learning Analytics: Applications in Educational Research*. Wiley.

Ferguson, R., & Sharples, M. (2014). Innovative pedagogy

at massive scale: Teaching and learning in MOOCs. In C. Rensing, S. de Freitas, T. Ley, & P. J. M.-Merino (Eds.), *Open Learning and Teaching in Educational Communities, proceedings of 9th European Conference on Technology Enhanced Learning (EC-TEL 2014), Graz, Austria, September 16-19*. (pp. 98–111). Heidelberg: Springer.

Flowerdew, L. (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, 20, pp. 58–68. https://doi.org/10.1016/j.jeap.2015.06.001

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge. https://doi.org/10.4324/9780203431269

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review*, 12, pp. 45–58. https://doi.org/10.1016/j.edurev.2014.05.001

Himelboim, I., Gleave, E., & Smith, M. (2009). Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication*, 14(4), pp. 771–789. https://doi.org/10.1111/j.1083-6101.2009.01470.x

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7(2), pp. 173–192. https://doi.org/10.1177/1461445605050365

Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3), pp. 723–747. https://doi.org/10.1111/j.1083-6101.2006.00033.x

Lander, J. (2015). Building community in online discussion: A case study of moderator strategies. *Linguistics and Education*, 29, pp. 107–120. https://doi.org/10.1016/j.linged.2014.08.007

Laurillard, D. (2012). *Teaching as a design science: building pedagogical patterns for learning and technology*. Abingdon: Routledge.

Lijffijt, J., & Gries, S. T. (2012). Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics*, 17(1), pp. 147–149. https://doi.org/10.1075/ijcl.17.1.08lij

Littleton, K., & Whitelock, D. (2005). The negotiation and co-construction of meaning and understanding within a postgraduate online learning community. *Learning, Media and Technology*, 30(2), pp. 147–164. https://doi.org/10.1080/17439880500093612

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation : Appraisal in English*. Palgrave Macmillan.

McEnery, T. (2016). Keywords. In P. Baker & J. Egbert (Eds.), *Triangulating Methodological Approaches in Corpus Linguistic Research*. Routledge.

Mercer, N. (2004). Sociocultural discourse analysis: analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics*, 1(2), pp. 137–168. https://doi.org/10.1558/japl.2004.1.2.137

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 1(4), pp. 519–549.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora*, 9, pp. 1–6. https://doi.org/10.3115/1117729.1117730

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44–49). https://doi.org/10.1.1.28.1139

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: motives and inhibitors to write and read. *Information Communication and Society*, 18(7), pp. 798–815. https://doi.org/10.1080/1369118X.2014.997268

Stubbs, M. (1986). "A matter of prolonged field work": Notes towards a modal grammar of English. *Applied Linguistics*, 7(1), pp. 1–25. https://doi.org/10.1093/applin/7.1.1

Tagg, C. (2012). *The Discourse of Text Messaging: Analysis of Text Message Communication*. London: Bloomsbury.

Wagner, D., & Herbel-Eisenmann, B. (2008). "just don't": The suppression and invitation of dialogue in the mathematics classroom. *Educational Studies in Mathematics*, 67(2), pp. 143–157. https://doi.org/10.1007/s10649-007-9097-x

Walsh, S., & Li, L. (2013). Conversation as space for Learning. *International Journal of Applied Linguistics (United Kingdom)*, 23(2).

Wegerif, R. (2010). Dialogue and teaching thinking with technology: Opening, expanding and deepening the 'inter-face.' In K. Littleton & C. Howe (Eds.), *Educational dialogues: Understanding and promoting productive interaction*. Routledge.

White, P. R. R. (2003). Beyond modality and hedging: A dialogic view of the language of intersubjective stance. *Text*, 23(2), pp. 259–284. https://doi.org/10.1515/text.2003.011

Ziegele, M., Breiner, T., & Quiring, O. (2014). What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, 64(6), pp. 1111–1138. https://doi.org/10.1111/jcom.12123

Variation of New German Verbal Anglicisms in a Social Media Corpus

Steven Coats
University of Oulu, Finland
steven.coats@oulu.fi

Abstract

This study examines the morphological behavior of new German verbal Anglicisms by exploring the frequencies of non-finite verbal forms in a large and novel German-language social media corpus. In order to identify new Anglicisms, a list of potential words was created by building German word forms from English verbal stems and excluding words that exist in the standard German lexicon. Then, the frequencies of the new non-finite verbal forms were tabulated, including prefixed verbal forms. Although new German verbal Anglicisms are infrequent, many types are attested, some of which exhibit inflectional variation. The data suggest that assimilation of the past participle to German orthographical norms is influenced by phonological and phonotactic, semantic, and stylistic/pragmatic considerations, and is mediated by frequency effects. In addition, the derivational morpheme *-ier-* is shown to be only somewhat productive. By considering frequency patterns of verbal Anglicisms in an online medium in which multilingualism and non-standard language are prevalent, the analysis provides a snapshot of the process by which the verbal lexicon of German is undergoing change.

Keywords: German morphology, borrowing, Anglicisms, social media, corpus linguistics

1. Introduction

English lexical borrowings (Anglicisms) constitute a steadily growing component of the German lexicon. While the morphological behavior of Anglicisms (and other borrowings) in German is usually predictable, they can in some cases exhibit non-standard inflectional forms, a process that is affected by semantic and phonological as well as pragmatic considerations. Verbs constitute only approximately 5% of Anglicisms in German overall (Eisenberg 2013), but due to their inflectional and derivational richness, their behavior can shed light on the morphological integration of borrowed lexemes into a language. In this study, German Anglicisms as infinitives, participles, prefixed verbs, and verbs derived using the *-ier-* affix are considered.

1.1. Non-finite verbal forms

The German infinitive ends in *-(e)n*. Verbs borrowed into German or derived from borrowed lexical material typically assimilate to the weak inflectional paradigm, forming the past participle (*Partizip II*) via circumfixation of *ge-* and *-(e)t*.

Infinitive	Past part.
fühlen 'to feel'	*gefühlt* 'felt'
lieben 'to love'	*geliebt* 'loved'
jobben 'to work' (esp. temporary jobs)	*gejobbt* 'worked'
batteln 'to battle' (esp. rap battles)	*gebattelt* 'battled'

Table 1: Infinitive and *Partizip II* of weak verbs in German

In Table 1, *fühlen* and *lieben* belong to the core German lexicon, whereas *jobben* and *batteln* are Anglicisms. In *jobben*, the final stem consonant has undergone gemination after a short stressed vowel in a closed syllable (the so-called *Silbengelenk*). In *batteln*, metathesis of <*le*> has occurred in order to adhere to the German norm for phoneme-grapheme correspondence, and the schwa of the infinitive suffix *-en* has been elided after a liquid. Verbs formed from English words with the same phonological shape (e.g. *giggle*, *babble*, etc.) are usually subject to this process and their orthography adapted (Duden 2016: §38, §92–94; Eisenberg 2011: 242–244), although for recent borrowings, variation exists (e.g. *googeln* and *googlen* 'to google').

For some verbal Anglicisms, partially assimilated participial variants exist alongside forms that conform to German inflection. Examples (1) and (2) are tweets in which the past participle of *liken* ('to like', esp. social media) exhibits full (*gelikt*) or partial (*geliked*) assimilation to the German inflectional norm. The first example notes that an influential German language authority, the Duden publishing house, codified the assimilated form in its dictionary in 2017.[1]

(1) *@user Jetzt ist es offiziell: du hast gelikt, er/sie/es likt. #Duden* [Now it's official: you have liked, he/she/it liked. #Duden]

(2) *@user Grade erst gesehen :3 Das meist geliked Video auf mein Kanal mittlerweile, Dankeschön!!!* [Just saw it :3 The most liked video in my channel in the meantime, Thankyou!!!]

1.2. Verb Derivation via Affixation

Prefixation of a verbal stem with a separable or an inseparable particle has historically been a productive process in German verb formation. Separable prefixes (mostly) specify the semantic scope of the verb spatially or temporally, whereas inseparable prefixes can express a wide range of possible meanings (see Duden, §1054–1076). Examples (for a standard German verb) are shown in (3). Prefixed Anglicisms are relatively common in the data used in this study (see also Baeskow 2017).

(3) *laufen* 'to run' *auslaufen* 'to run out' *(sich) verlaufen* 'to get lost'

[1] Usernames have been anonymized.

The verbal infixes *-ier-* and *-isier-*, in verbs such as *studieren* ('to study') or *legalisieren* ('to legalize'), have historically been the most important morphemes for the integration of borrowed lexical material into the German verbal system, productive since at least the 12th century (Öhmann 1970). Older, codified *-ier-* derivations are in some cases in competition with verbal forms showing simple suffixation of *-en* (e.g. *attackieren* vs. *attacken*, both 'to attack').

In the following, a brief review of related work is provided, followed by a description of the methods used to collect and filter the data and identify new German Anglicisms. In Section 4, the semantic fields of the most frequent new Anglicisms are considered, and the frequencies of past participles (*Partizip II*) are analyzed with respect to their assimilation to German orthographical norms and their use as verbal elements or as adjectives. The frequencies of *-ier-* derivations are also considered.

2. Previous Work

English has long been a source of lexical material for other languages, and in the last hundred years, English words have been adopted into the vocabularies of languages worldwide (Görlach 2003). This is particularly true for German since 1945, a result of social, economic, and political factors (von Polenz 1999). Studies of English lexical borrowings in German have investigated their semantic and structural aspects, examined their pragmatic contexts of use, and estimated their overall prevalence in German, for example on the basis of corpora derived from printed material.

Carstensen (1965) described lexical, grammatical and syntactic influences of English on German on the basis of texts printed in West German newspapers and magazines from 1961–1964, primarily the weekly news magazine *Der Spiegel*, and introduced the distinction between *Bedürfnislehnwörter* ('necessary borrowings'), or words for which no lexeme exists in the receptor language, and *Luxuslehnwörter* ('luxury borrowings'), or words whose semantic content is covered by existing lexemes. Yang (1990), Onysko (2007), and Burmasowa (2010) utilized corpora of journalistic texts to show increased usage of Anglicisms over time. Onysko and Winter-Froemel (2011) utilized the terms *catachrestic* (representing a new concept) and *non-catachrestic* (expressing the same content as an existing lexeme) to take a closer look at the most frequent Anglicisms in the corpus of Onysko (2007), finding that for non-catachrestic borrowings, loanword age and usage pragmatics are important factors in the adoption of an item.

Eisenberg (2013) analyzed chronological trends in Anglicisms on the basis of corpora compiled from popular, scientific, journalistic, and literary texts published in the periods 1905–1914, 1948–1957, and 1995–2004, showing that some verbal Anglicisms (e.g. *flirten* 'to flirt' or *boykottieren* 'to boycott') were well attested in German already before 1914 (84). Winter-Froemel et al. (2015) regressed Anglicism frequency with several variables, finding that for words that replicate the semantic content of existing lexemes (*non-catachrestic* borrowings), shorter length and lexical field (technology and internet) positively influence the success of the borrowing. Baeskow (2017) discussed verbal Anglicisms with inseparable prefixes from the semantic field of information technology (e.g. *ergoogeln*), focusing specifically on the lexical aspect of inseparable prefixation.

While research into Anglicisms in German has been extensive, the status of inflectional variants of non-finite verb forms has not been a primary focus. Onysko suggested that participles derived from verbal borrowings are more likely to exhibit standard German weak participial inflection (e.g. *gecancelt* 'cancelled', *gechattet* 'chatted'), whereas forms borrowed as adjectives (i.e. not derived from a borrowed verb) are more likely to retain English or partially English orthography (e.g. *relaxed* or *gefaked*), especially if their phonological realization in English and German more or less coincide (2007: 235–237).

3. Data and Methods

653,457,659 tweets with "place" metadata were collected globally from the Twitter Streaming API from November 2016 until June 2017 using *Tweepy* (Roesslein 2015). From this "seed" data, 70,986 users who had authored at least one German-language tweet and with place metadata from Germany, Austria or Switzerland were identified and all of their tweets, or the most recent 3,250 tweets (whichever was larger), downloaded from Twitter's API during April 2018. The timelines of 60,683 users were downloadable (others presumably having been set to private, deleted, or banned by Twitter). Of the 61,118,733 tweets downloaded in this manner, 36,240,530 (59.3%) were in German, according to tweet metadata. Tweets were tokenized using the nltk tokenizer (Bird et al. 2009), resulting in a corpus of 534,211,366 tokens.[2]

To build a set of potential verbal borrowings, the 1,000 most frequent base verbal forms (corresponding to English infinitives without *to*) were accessed from the *British National Corpus*, the *Corpus of Contemporary American English*, and the *Wikipedia Corpus of English* (Davies 2004–, 2008–, 2015)[3], then combined with 1,413 forms from the *Pattern Dictionary of English Verbs* (Hanks 2013)[4]. From this list of 2,630 unique types, German infinitives and participles were created using regular expressions, taking into account German phonotactics and orthographic conventions. Forms with inseparable prefixes (*be-*, *er-*, *ent-*, *emp-*, *miss-*, *ver-*, *zer-*, *über-*) and separable prefixes (*ab-*, *an-*, *auf-*, *aus-*, *durch-*, *ein-*, *her-*, *herauf-*, *herum-*, *herunter-*, *hin-*, *hinzu-*, *mit-*, *voran-*, *los-*, *mit-*, *vor-*, *weg-*, *zurück-*, *zusammen-*) were created, as were infinitives of prefixed verbs with an infixed *-zu-* (e.g. *anzutwittern*). The same forms were generated from the stems for the *-ier-* and *-isier-* derivations, and adjectival inflections were accounted for (e.g. *das gelikte Foto* 'the liked photo'). English false positives were removed using an English word list of 236,736

[2] The corpus can be generated from the list of the tweet IDs available at https://github.com/stcoats/GermanAnglicisms

[3] http://corpus.byu.edu

[4] http://pdev.org.uk.

types from nltk (Bird et al. 2009).[5]

In order to exclude well-established Anglicisms that are considered part of the standard German lexicon, each of the forms generated from the procedure described above was matched against a list of 239,650 German word types (Kleuker 2016).[6] To account for forms not attested in the Kleuker (2016) list but which are nonetheless standard German words, Anglicisms were checked with SMOR, a finite-state transducer for morphological analysis of German words whose current lexicon contains approximately 6,000 verbal stem types (Schmid et al. 2004, Fitschen 2004). Only words not attested in standard German according to these two criteria were further considered.[7]

In total, the iterative procedure used to create new German verbal Anglicisms generated a large number of possible word forms.[8] While most of these forms were not present in the corpus, those attested exhibited significant variation.

4. Results and Analysis

4.1. Overall frequencies

New non-finite verbal Anglicisms in the corpus are attested from diverse semantic fields and exhibit variation in orthography. A total of 3,201 types in the corpus produced matches with the automatically-generated list, comprising 117,246 tokens. Table 2 shows the 20 most frequent types.

	Type	Freq		Type	Freq
1	*twittern*	28921	11	*adden*	1214
2	*streamen*	9248	12	*geupdated*	1188
3	*chillen*	8543	13	*haten*	1146
4	*getwittert*	6567	14	*rendern*	1054
5	*googlen*	2829[9]	15	*coden*	1000
6	*gestreamt*	2232	16	*followen*	831
7	*geliked*	1415	17	*gevotet*	810
8	*supporten*	1370	18	*cachen*	782
9	*gefixt*	1300	19	*tracken*	781
10	*geflasht*	1271	20	*sharen*	758

Table 2: Most frequent new Anglicisms

Many of the most frequent types clearly represent *Bedürfnislehnwörter*, or cultural borrowings that fill a gap in the receptor language lexicon: *twittern*, *streamen*, *googlen*, *liken*, *adden*, *updaten*, *rendern*, *coden*, *followen*, and *sharen*, and their past participles, are primarily used in the context of social media or information technology; their meanings correspond closely to the social-media- or IT-specific meanings of their English source words. In this data, *gefixt* is used in the sense of 'to repair/fix' (an online service or website): the older meaning of the denominal borrowing *fixen*, 'to inject drugs', is not attested.[10] Among the most frequent types, only three are used mainly in non-IT contexts: *supporten* 'to support' denotes support for a sports team, as in (4). *Geflasht* is used as a predicate adjective meaning 'excited' (*ich bin geflasht* 'I'm excited'), but also to denote rewriting the memory of an IT device. *Haten* is a stylistically marked equivalent to standard German *hassen* ('to hate') (5).

(4) *so kinder, jetzte jehts los. kurz vorm olympiastadion. supporten fuer hertha und die relegation. alle die daumen druecken!!!* [so children, now it begins. just in front of Olympic Stadium. supporting hertha and relegation. everyone cross your fingers!!!]

(5) *Ich bin ja ganz vorne mit dabei wenns darum geht den #EmojiFilm zu haten... aber den Trailer find ich gar nicht mal so scheiße.* 😶 [I'm among the first to agree when it comes to hating the #EmojiFilm... but the trailer is not even so shitty. 😶]

The frequency distribution of new Anglicisms exhibits a "long tail" – large number of types that occur only once in the corpus (i.e. *hapax legomena*). The semantic values of the 1,271 *hapax* types are diverse, and mostly unrelated to social media or information technology. A sample – the meanings of which are transparent from the verbal stem – is shown in (6).

(6) *annoyen*, *breathen*, *ercapturen*, *zurückcheaten*, *gehealed*, *mitgementioned*, *gelookt*, *killiert*, *encouragierend*, *failiert*

129 infinitive types with inseparable prefixes were found, the most frequent being *vertwittern* ('to twitter away/out'), *entfollowen* ('to stop following on social media'), and *entliken* ('to stop liking on social media'). For separable prefixes, 349 infinitive types were attested: *abfucken* ('to fuck up') was the most common, followed by *antwittern* ('to twitter to someone') and *abchillen* ('to chill out'). Other attested forms included *anbeefen* ('to start an argument/complain to someone'), *aufleveln* ('to level up in a computer game'), and *ansneaken* ('to sneak up on someone'). The prefixed infinitive form with infixed *-zu-* was attested by 70 types: *abzufucken*, *mitzutwittern*, and *anzutwittern* were the most frequent.

[5] Some Anglicisms generated by the procedure are actual English words – these (e.g. *driven*) are often present in longer codeswitched sequences rather than as single-word Anglicisms in German text.

[6] https://github.com/davidak/wortliste. The list aggregates data compiled by the Berlin-Brandenburg Academy of Sciences, the Leipzig Corpora Collection of the University of Leipzig, and the Institute for the German Language in Mannheim.

[7] An Anglicism wordlist comprising infinitives and past participles not matching standard German words is available at https://github.com/stcoats/GermanAnglicisms.

[8] For example, from the English verb *to wreck*, the non-finite German verbal forms *wrecken, wreckend, gewreckt, gewrecked, wreckieren, wreckierend, wreckiert, wreckisieren, wreckisierend*, and *wreckisiert* were generated; for each of these 28 prefixed forms were created.

[9] 6388 if *googeln*, whose stem is in the SMOR lexicon, is included.

[10] The prefixed form *angefixt* 'be hooked on', however, was well attested.

Some false positives were present in the frequency counts as the result of non-standard spellings. For example, *erfaren*, attested twice in the corpus, is a present participle in the match list derived from *to fare*. In the tweets in question, the type is a non-standard spelling of standard German *erfahren* ('to experience' or 'experienced'). Other non-standard spellings include *überagend*, from *to age* (*überragend* 'outstanding'), *forden* and *erforden*, from *to ford* (*fordern* 'demand' and *erfordern* 'require'), *gestatet*, from *to state* (*gestattet* 'allowed'), *ausgerut*, from *to rut* (*ausgeruht*, 'rested'), and *verwanten*, from *to want* (*verwandten* 'related' or 'relations'). Overall, the frequencies of these forms are low. Another false positive was the type *nabend*, created automatically as a present participle from *to nab*, but a common non-standard German word (a blend from ***guten Abend*** 'good evening').

4.2. Variation in the Past Participle

Variation between the assimilated and partially-assimilated forms of the past participle was attested for 219 past participle types: Table 3 shows the counts and an effect size measure, the logarithmic odds ratio, for the most frequent forms.[11] Figure 1 shows the log odds ratio versus the log of number of occurrences of the participle for forms for which both variants are attested at least once: More frequent participles are more likely to exhibit the standard inflectional ending *-t*, whereas less frequent participles are more likely to retain *-ed* endings.

	Type	Freq	Type	Freq	logOR
1	*getwittered*	4	*getwittert*	6567	-7.40
2	*gestreamed*	121	*gestreamt*	2232	-2.91
3	*geliked*	1415	*gelikt*	197	1.97
4	*geupdated*	1118	*geupdatet*	404	1.08
5	*geflashed*	309	*geflasht*	1271	-1.41
6	*gefixed*	223	*gefixt*	1300	-1.76
7	*geleaked*	375	*geleakt*	993	-0.97
8	*gevoted*	131	*gevotet*	810	-1.82
9	*gelaunched*	81	*gelauncht*	601	-2.00
10	*geadded*	98	*geaddet*	332	-1.22

Table 3: Variation in Past Participles

The partially assimilated forms *geliked* and*geupdated* are preferred to *gelikt* and *geupdatet*, but otherwise the more frequent variants have standard inflection. The degree to which English and German orthography overlap in the representation of vowel sounds appears to influence assimilation to German inflection. Retention of partially English orthography may help recognition of the diphthongs [aɪ] and [eɪ] in forms such as *geliked* or *geupdated*, whereas the German-inflected forms could be realized with [ɪ]/[i] and [a]. Forms more readily assimilated to German participial inflection (those with negative log-odds ratio values) have stem vowels whose realization is similar to that of the original English participles. The recentness of borrowing may also play a role — forms with negative log-odds

[11] The logarithmic odds ratio, $\log \frac{n_x}{n_y}$, is symmetrical about zero and results in positive values when x is more frequent and negative values when y is more frequent.

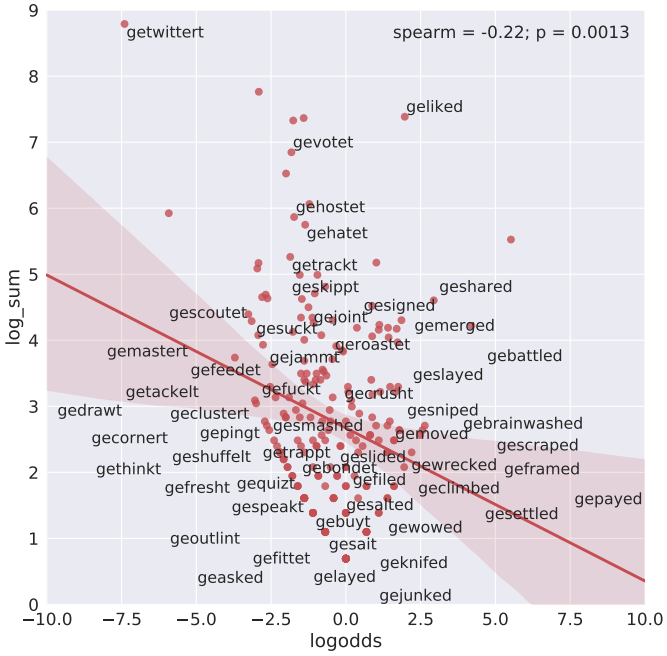

Figure 1: Assimilated and partially-assimilated past participle forms

ratios which are semantically not necessarily related to online behavior may be somewhat older borrowings and thus more advanced in terms of assimilation to the German inflectional pattern. The negative correlation between the log frequency and the log odds ratio shown in Figure 1 suggests that, as with other types of language change, frequency effects may mediate assimilation to standard orthography.

4.3. Past Participle as Attributive Adjective

In order to check use of participles as attributive and superlative adjectives, the frequencies of past participles with the inflectional suffixes *-e*, *-em*, *-en*, *-er*, *-es*, *-este*, *-estem*, *-esten*, *-ester*, and *-estes* were counted. Table 4 shows the ten most frequent fully assimilated past participles, their frequencies as participial or adjectival attributes, and the verbal to adjectival log odds ratio. While the tendency to be used as a verbal component or an adjectival attribute depends on the semantics of the verb, verbal use is more common — the verbal to adjectival log odds ratio for all fully-assimilated participles is 2.93, meaning the forms are almost 19 times more likely to be used as verbal elements.

For the partially-assimilated participles, only a handful are used as attributive adjectives (Table 5). The log odds ratio for all of these forms is 5.42. Adjectival usage is almost non-existent.

4.4. *-ier-* Derivations

83 types created via derivation with *-ier-* were attested. Many of these, however, are established dialect words (e.g. the Swiss German words *grillieren* 'to grill/barbecue' or *parkieren* 'to park a car') or non-standard spellings of established lexical items (e.g. *boycottieren* instead of

	Type	Freq_part	Freq_adj	logOR
1	*gebloggt*	8840	67	4.88
2	*getwittert*	6567	209	3.45
3	*geblockt*	5862	172	3.53
4	*gecheckt*	3111	7	6.10
5	*gerockt*	2433	2	7.10
6	*gegoogelt*	2276	28	4.40
7	*gestreamt*	2232	49	3.82
8	*gechillt*	1487	377	1.37
9	*geleakt*	993	411	0.88
10	*gefixt*	1300	20	4.17

Table 4: Variation in Past Participles

	Type	Freq_part	Freq_adj	logOR
1	*geliked*	1415	3	6.16
2	*geupdated*	1188	0	inf
3	*geleaked*	375	4	4.54
4	*geflashed*	309	0	inf
5	*gefeatured*	250	0	inf
6	*gefixed*	223	0	inf
7	*gehacked*	197	0	inf
8	*getagged*	164	0	inf
9	*gevoted*	131	0	inf
10	*gefollowed*	130	1	4.87

Table 5: Variation in Past Participles

boykottieren 'to boycott', *debatieren* instead of *debattieren* 'to debate'), and thus do not represent new Anglicisms. *-ier*-derived forms of the most common new verbal Anglicisms, those pertaining to social media and IT, are almost non-existent: *twitterieren* occurs once in the corpus, as does *updatieren*. A few lexemes appear to be new borrowings from English: *relatieren* ('to be relevant/similar/related') occurs 15 times. Verb formation from borrowed lexical items via the *-ier-* morpheme, although still somewhat productive in German, appears to be less common than suffixation of a borrowed stem with the *-en* infinitive suffix. Word length considerations and communicative economy may also play a role, especially considering the character limitation inherent to Twitter.

5. Conclusions and Future Outlook

Significant variation exists in the morphology of new verbal Anglicisms in a large German-language social media corpus from Twitter. The most frequent new Anglicisms denote entities from the domains of social media, computer-mediated communication, and information technology, and are typically used as infinitives or past participles. For past participles, variation in assimilation to German inflection may reflect phonological considerations as well as the recency of the borrowing, and is manifest in frequency counts. Partially-assimilated past participles are used almost exclusively as verbal elements, while fully assimilated past participles can be used as attributive adjectives.

Future work with the data can be organized along the following lines: First, a more thorough consideration of the phonological, semantic and pragmatic factors that prompt use of an Anglicism could be undertaken for widely-attested forms that have a high degree of semantic overlap with common verbs in the German core lexicon, such as *worken* ('to work'), *playen* ('to play'), *walken* ('to walk'), or *eaten* ('to eat'): In addition to being used for stylistic and pragmatic reasons, such lexemes may be undergoing semantic specialization as well. A quantitative approach using word embeddings could shed light on this process. Secondly, the productivity of both borrowed verbal stems and verbal affixes can be measured, for example by calculating vocabulary growth rates. Are borrowed stems more productive than stems from the core lexicon? Thirdly, sociolinguistic parameters of variation can be assessed by measuring correlations between Anglicism use and demographic features that can be gleaned from Twitter metadata such as user location, gender, or social network membership. Finally, by comparing aggregate measures of morphological variation in this data to similar measures in other large corpora drawn from social media and non-social-media sources, broader insight can be gained into the rate at which the lexicon of German is undergoing renewal.

6. References

Baeskow, H. (2017). "#Virtual Lexicality: The semantics of innovative prefixed verbal anglicisms in German". In: *Word Structure* 10.2, pp. 173–203.

Bird, S., E. Loper, and E. Klein (2009). *Natural Language Processing with Python*. Newton, MA: O'Reilly.

Burmasowa, S. (2010). *Empirische Untersuchung der Anglizismen im Deutschen am Material der Zeitung 'Die Welt'*. Bamberg: University of Bamberg Press.

Carstensen, B. (1965). *Englische Einflüsse auf die Deutsche Sprache nach 1945*. Heidelberg: Carl Winter Verlag.

Duden (2016). *Die Grammatik (9th ed.)* Berlin: Dudenverlag.

Eisenberg, P. (2011). *Das Fremdwort im Deutschen*. Berlin and New York: de Gruyter Mouton.

– (2013). "Anglizismen im Deutschen". In: *Reichtum und Armut der deutschen Sprache : Erster Bericht zur Lage der deutschen Sprache*. Ed. by Deutsche Akademie für Sprache und Dichtung, Union der deutschen Akademien der Wissenschaften. Berlin: de Gruyter, pp. 57–119.

Fitschen, A. (2004). "Ein Computerlinguistisches Lexikon als komplexes System". Ph.D. Thesis. Universität Stuttgart.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Kleuker, D. (2016). *Wortliste. https://github.com/davidak/wortliste*.

Onysko, A. (2007). *Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching*. Berlin: de Gruyter.

Onysko, A. and E. Winter-Froemel (2011). "Necessary loans – luxury loans? Exploring the pragmatic dimension of borrowing". In: *Journal of Pragmatics* 43.6, pp. 1550–1567.

Polenz, P. von (1999). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. Band III: 19. und 20. Jahrhundert*. Berlin: de Gruyter.

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

Schmid, H., A. Fitschen, and U. Heid (2004). "SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1263–1266.

Winter-Froemel, E., A. Onysko, and A. Calude (2011). "Why some non-catachrestic borrowings are more successful than others: a case study of English loans in German". In: *Language Contact Around the Globe*. Ed. by A. Koll-Stobbe and S. Knospe. Frankfurt am Main: Peter Lang, pp. 119–142.

Yang, W. (1990). *Anglizismen im Deutschen: Am Beispiel des Nachrichtenmagazins Der Spiegel*. Tübingen: Niemeyer Verlag.

Öhmann, E. (1970). "Suffixstudien VI: Das deutsche Verbalsuffix -ieren". In: *Neuphilologische Mitteilungen* 71.3, pp. 337–356.

The Polly Corpus: Online Political Debate in Germany

Tom De Smedt [1], **Sylvia Jaki** [2]
[1] University of Antwerp, [2] University of Hildesheim
tom.desmedt@uantwerpen.be, jakisy@uni-hildesheim.de

Abstract

Nowadays, a large part of political discourse takes place on social media like Twitter, from campaign rhetoric and fan talk to verbal mudslinging and aggressive / unethical hate speech. To encourage research on the language of political debate and especially hate speech, we present POLLY, a free multimodal corpus with about 125,000 German tweets posted before, during and after the 2017 German federal elections. It includes tweets about politicians, by politicians, by fans of politicians, and by far-right supporters.

Keywords: Twitter, political discourse, hate speech

1. Introduction

The 2017 German federal elections have experienced a considerable rise in right-wing populism, closely linked to the political party *Alternative für Deutschland* (AfD), which achieved a striking success with 12.6% of the votes. Social media such as Twitter are believed to have played an important role in the electoral debate, and in propelling the increasingly polarized rhetoric (Conover et al., 2011; Vaccari et al., 2014).

In the lead-up to the elections, Germany experienced a number of violent incidents such as the 2016 New Year's Eve sexual assaults by male refugee perpetrators, and an Islamist terrorist attack on the Berlin Christmas market in December 2016. This sharply polarized the German sentiment towards refugees (Dahlgreen, 2016).

More recently, the EU has pressed IT companies to increase their efforts to counter online hate speech. In Germany, the new NetzDG law now forces social media platforms to delete reported hateful content within 24 hours, with remarkable consequences such as one AfD politician being temporarily suspended from Twitter. A recent study shows a correlation between increased hate speech on German social media and increased physical violence towards refugees (Müller & Schwarz, 2017).

We present POLLY, a free multimodal study corpus of online political debate in Germany. It consists of about 125,000 German tweets and 4,000 linked images, posted between August 2017 and December 2017, with the election date in-between (September 24th).

2. Methods

The POLLY corpus was mined using the Pattern toolkit (De Smedt & Daelemans, 2012) and the Twitter API. We also added a number of tweets manually. The corpus is freely available as a Google Sheet[1] (Figure 1). The image set is available in Google Drive[2] (Figure 2).

[1] https://docs.google.com/spreadsheets/d/1c5peNMjt24U0FcEMSj8gD_JjzumqXTWbPWa_yb2nNt0

[2] https://drive.google.com/drive/folders/12VMjTlAUS2f0_5wgsYN4QI_R-Ad6J62L

Google Sheets can be viewed online, downloaded as an .xls or .csv file, and have automatic revision history. Registered contributors can add new rows of data and new columns with annotations, which we encourage.

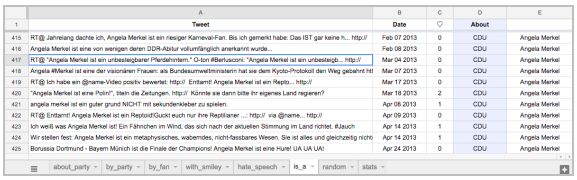

Figure 1: Dataset of tweets in Google Sheets

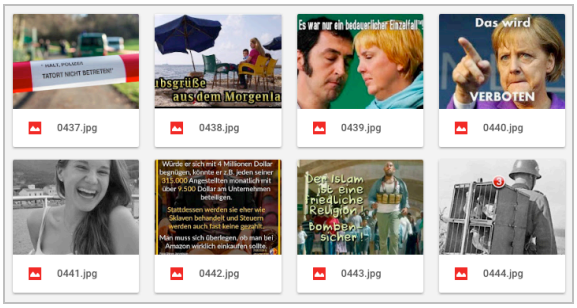

Figure 2: Dataset of images in Google Drive

The POLLY corpus is currently divided into 7 subsets (more may follow) that contain tweets *about* politicians, tweets *by* politicians, tweets *by fans* of politicians, tweets *with emojis*, tweets with *hate speech*, tweets with *is-a* statements, and *random* German tweets. Each tweet in the corpus has a timestamp, the number of likes, other metadata, and/or images, which have been labeled and transcribed by the Google Vision API.

Careful steps were taken to sample tweets from every month. The corpus covers political debate about elected political parties CDU + CSU, SPD, Linke, Grüne, FDP, and AfD. It contains tweets from 35 politicians (22 men, 13 women) selected for their number of votes, Twitter followers and Google News results. For example, Angela Merkel (CDU) has 25K followers and 5M news articles, while Alice Weidel (AfD) has 40K followers and 70K news articles. Both are well-known politicians.

No steps were taken to balance the number of tweets per politician, to represent how parties such as AfD are more prominent online than parties such as CDU, even though CDU has more voters (Kollanyi & Howard, 2017).

Party	Politician	Followers	Tweets
CDU	Angela Merkel	25K	0
CDU	Armin Laschet	20K	197
CDU	Julia Klöckner	50K	480
CDU	Peter Altmaier	220K	88
CSU	Horst Seehofer	<1K	17
SPD	Aydan Özoguz	15K	116
SPD	Frank-Walter Steinmeier	2K	37
SPD	Heiko Maas	240K	163
SPD	Hubertus Heil	60K	181
SPD	Manuela Schwesig	135K	298
SPD	Martin Schulz	**635K**	182
SPD	Olaf Scholz	30K	85
SPD	Ralf Stegner	40K	1,006
SPD	Sigmar Gabriel	225K	35
AfD	Alexander Gauland	1K	12
AfD	Alice Weidel	25K	183
AfD	Beatrix von Storch	35K	**1,106**
AfD	Frauke Petry	55K	247
AfD	Georg Pazderski	5K	151
AfD	Jörg Meuthen	15K	186
AfD	Kay Gottschalk	1K	51
FDP	Christian Lindner	245K	441
FDP	Hermann Otto Solms	5K	5
FDP	Katja Suding	5K	56
FDP	Marco Buschmann	5K	166
FDP	Nicola Beer	5K	625
FDP	Tobias Huch	10K	670
Linke	Bernd Riexinger	15K	291
Linke	Bodo Ramelow	25K	562
Linke	Gregor Gysi	250K	31
Linke	Katja Kipping	75K	341
Linke	Sahra Wagenknecht	285K	131
Grüne	Cem Özdemir	75K	284
Grüne	Katrin Göring-Eckardt	115K	403
Grüne	Simone Peter	30K	838

Table 1: Parties and politicians in the POLLY corpus

We have anonymized the dataset in line with the EU's new General Data Protection Regulation.[3] The names of politicians (which we consider to be public figures) are exposed, but usernames of private citizens have been anonymized to `@name`, unless they are mentioned in a tweet by a known politician. German media outlets were anonymized to `@news`, police to `@polizei`. URLs were removed, as were most images depicting private citizens.

Finally, we have taken careful steps to include diverse viewpoints. About 50% of the tweets in the corpus are from unique users. Excepting politicians, no more than a 100 tweets are from one single user.

[3] https://en.wikipedia.org/wiki/GDPR

3. Results

The POLLY corpus is divided into 7 different subsets:

3.1 Tweets about politicians

About 20,000 tweets that mention the name of a known politician or political party. For example: "Manchmal vergesse ich, dass Frauke Petry ein echter Mensch ist und nicht nur ein Meme" (posted August 30, 2017).

3.2 Tweets by politicians

About 15,000 tweets posted by politicians or political parties. For example: "Diese Zeilen stammen nicht aus einer Pressemitteilung des IS, sondern von Abdul (23) aus Berlin - leider kein Einzelfall! Der Antisemitismus hält unverhohlen Einzug - während die etablierte Politik schweigt. Gehört das etwa nun auch zu Deutschland?" (posted by Alice Weidel, December 12, 2017; Figure 3).

Figure 3: Tweet by AfD-politician Alice Weidel

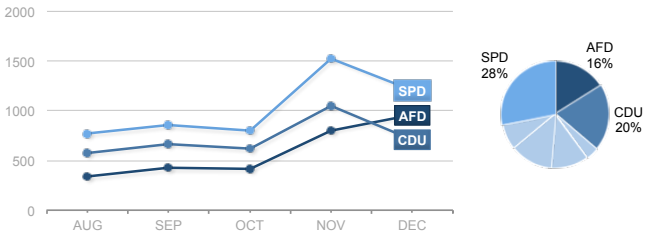

Figure 4: Distribution of tweets about politicians

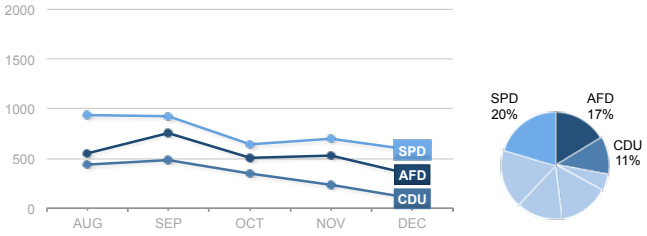

Figure 5: Distribution of tweets by politicians

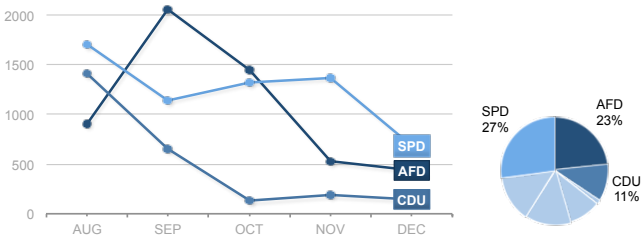

Figure 6: Distribution of tweets by fans

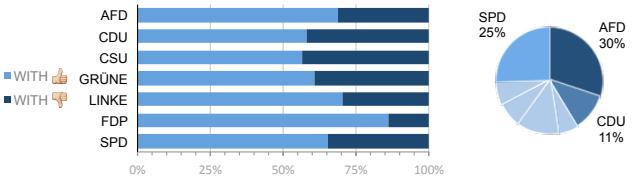

Figure 7: Distribution of tweets with 👍 or 👎

3.3 Tweets by fans

About 20,000 tweets by private citizens that liked tweets by a particular politician or political party. For example: "Ich glaub die SPD Hannover hat den Schuss nicht mehr gehört? Seit wann haben solche Kebabs uns zu drohen?" (posted by an AfD fan, October 5, 2017).

3.4 Tweets with emojis

About 20,000 tweets that contain 👍 (like), 👎 (dislike), ❤️ (love) or 😡 (hate). Some also contain the name of a politician or political party while others are random. For example: "AfD schockt Linksmedien! 😁😱👍 Ihr habt die Merkel Dämonin ('schaff euch alle') und die Misere ('lebt mit Terror') vergessen" (posted October 31, 2017).

3.5 Tweets with hate speech

About 20,000 tweets by 100+ far-right supporters (see Jaki & De Smedt, 2018), with instances of racist and/or violent rhetoric, e.g., "Schluss mit dem Religionsfreiheit! Islam gehört nicht zu Europa, Moscheen auch nicht! Wer Islam verharmlosen versucht, der ist entweder Dumm oder kennt nicht mal was Islam ist. Wenn wir Islam vernichten wollen, eben alle Religionen abschaffen!" (posted December 04, 2014).

3.6 Tweets with is-a statements

About 2,500 tweets with the name of a known politician followed by *ist ein(e)*, followed by a description to praise (*krasser Battle Rapper*) or mock the politician (*Spielzeug der Reptilienwesen*, *aufgeblasener Gartenzwerg*). As the larger part of the tweets can be classified as metaphors, this set could serve as a basis for research on German political metaphors both lexicalized (*mieses Arschloch*) and creative *(politischer Wünschelrutengänger)*.

3.7 Random tweets

About 20,000 random tweets in German, retrieved with search terms such as *der*, *die*, *das*, *Bahnhof*, etc.

4. Analysis

We compared different subsets to examine language use, by counting each word in each subset and then using the chi-square statistical test. This exposes keywords that are significantly biased ($p <= 0.05$). For example, AfD fans write more about Islam, refugees, crime, and left-wing voters (*Gutmenschen*, *Linksfaschisten*), while SPD fans write more about the EU, German railways, and trade.

CDU politicians write more about government, economy and electric vehicles, while Die Grünen politicians refer more often to climate, renewable energy, and the future. Politicians in general tend to write more about research, pensions, family, while private citizens write more about news articles and the German nation, often adding modal particles (*ach*, *naja*, *tja*, *vielleicht*). It is interesting to note that, after six months of government formation, the new "GroKo" coalition (CDU + CSU + SPD) included a new ministry of *Heimat* (homeland), possibly in response to voters' preoccupation with the German nation.[4]

Word	about	by	fans	hate	random
Forschung	19	**46**	27	24	17
Wirtschaft	141	**216**	147	111	68
Rente	67	**173**	88	70	17
Polizei	138	103	172	**726**	176
Terror	163	85	102	**513**	58
Flüchtlinge	122	84	162	**456**	38
Araber	6	3	11	**95**	5
Moschee	9	5	14	**103**	24
Volk	49	8	64	**226**	19
Krieg	134	76	171	**342**	154

Table 2: Sample keywords and their prevalence

Twitter is a relatively new phenomenon marked by the reciprocity of communication, the simultaneity of the private and public sphere, and the fast dissemination of ideas. It disrupts the unidirectional communication from policy makers to citizens, and offers greater possibilities for participatory debate about politics (Emmer, 2017). This participation influences the dissemination of news, which itself often entails a framed interpretation, since users connect events to personal experience and world views (Maiereder & Ausserhofer, 2014). Depending on a user's world view or communication habits, this may lead to hate speech. Defining hate speech thus is difficult (Warner & Hirschberg, 2012; Davidson et al., 2017) and a legal EU framework is in ongoing development.

Tweets are multimodal: they consist of text, but also and increasingly include visual information such as emojis and images (Schmidt & Wiegand, 2017). For example, hate tweets often contain emojis that display aggression, such as 😡 or 👊, and images that may glorify violence or constitute "visual racism" (van Leeuwen, 2000).

[4] https://www.reuters.com/article/us-germany-politics-heimat/home-is-where-the-heimat-is-germans-bemused-by-new-ministry-idUSKBN1FS2UD

Far-right supporters that post hate speech (i.e., 3.5) write more about security, terrorism, religion and immigration, topics that are also discussed by AfD fans (i.e., 3.3), but the far-right uses more racial slurs (e.g., *Muselstrümpfe, Negergesindel*, *Teppichknutscher*) and more references to violence, danger and death. Table 2 provides a sample of keywords and their prevalence in each subset.

The hate speech dataset can be used to train a system for hate speech detection (e.g., De Smedt, De Pauw & Van Ostaeyen, 2018). In our work, we used a neural network classifier with 84% accuracy (Jaki & De Smedt, 2018).

The tweets with emojis can be used to train a system for sentiment analysis on political tweets (e.g., Tumasjan et al., 2010). Following is an example Python script:

```
from grasp import download
from grasp import csv
from grasp import tmp
from grasp import Perceptron
from grasp import balanced
from grasp import chngrams
from grasp import kfoldcv

POLLY  = 'https://docs.google.com/spreadsheets/d/'
POLLY += '1c5peNMjt24U0FcEMSj8gD_JjzumqXTWbPWa_yb2nNt0'
POLLY += '/gviz/tq?tqx=out:csv&sheet='

polarity = {
    u'\u2764\ufe0f' : +1, # Red Heart
    u'\U0001f44d'   : +1, # Thumbs Up
    u'\U0001f44e'   : -1, # Thumbs Down
    u'\U0001f621'   : -1  # Pouting Face
}

def v(s):
    for k in polarity:
        s = s.replace(k, '')
    v = set()
    v.update(chngrams(s, n=1))
    v.update(chngrams(s, n=3))
    return v

s = download(POLLY + 'with_emoji', cached=True)
f = tmp(s)# file-like

data = []
for tweet, date, likes, about, k in csv(f.name):
    tweet = tweet.replace(about, '@name')
    data.append((v(tweet), polarity.get(k, +1)))

data = list(balanced(data, n=7500))

P, R = kfoldcv(Perceptron, data, n=5, k=3)
print(P)# precision
print(R)# recall

model = Perceptron(data, n=5)
print(model.predict(v('Islamgeile Propaganda!')))
```

The Python script uses grasp.py[5] to download the POLLY tweets with emojis as a CSV and then creates a training example from each tweet, in the form of a vector with character trigrams as features. The Perceptron machine learning algorithm (a single-layer neural network) is then trained using five iterations (n=5) and tested using 3-fold cross-validation, with precision and recall of 83%.

In a test with sentiment analysis on the POLLY corpus, we find that tweets posted by politicians are more positive (67%), while tweets by fans are more evenly distributed between positive (56%) and negative. Unsurprisingly, tweets with hate speech are the least positive (47%).

[5] http://github.com/textgain/grasp

5. Discussion

Whether or not it is desirable to use the POLLY corpus to train AI systems that, for example, can detect political preferences or rhetoric that is seen as unethical is open for discussion. Nonetheless, we hope that POLLY can help in the study of political discourse. New contributions to the corpus are welcomed, as are standardization efforts (e.g., Fišer & Beißwenger, 2017).

6. References

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F. and Flammini, A. (2011). Political polarization on twitter. *ICWSM*, 133, pp. 89-96.

Dahlgreen, W. (2016). German attitudes to immigration harden following attacks. http://goo.gl/RvWm5f

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language.

De Smedt, T. and Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13, pp. 2063-2067.

De Smedt, T., De Pauw, G. and Van Ostaeyen, P. (2018). Automatic Detection of Online Jihadist Hate Speech. *CLiPS Technical Report Series*, CTRS-007.

Emmer, M. (2017). Soziale Medien in der politischen Kommunikation. In J.-H. Schmidt (Ed.), *Handbuch Soziale Medien*. Wiesbaden: Springer, pp. 81-99.

Fišer, D. and Beißwenger, M. (2017). *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World.* Ljubljana: Ljubljana University Press.

Jaki, S. and De Smedt, T. (2018). Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection. Manuscript submitted.

Kollanyi, B. and Howard, P. N. (2017). Junk News and Bots during the German Federal Presidency Election: What Were German Voters Sharing Over Twitter?

Maiereder, A. and Ausserhofer, J. (2014). Political Discourses on Twitter. In K. Weller (Ed.), *Twitter and Society*. New York: Peter Lang, pp. 305-318.

Müller, K. and Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime.

Schmidt, A. and Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proc. of SocialNLP*, pp. 1-10.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1), pp. 178-185.

Vaccari, C., Valeriani, A., Barberá, P., Jost, J. T., Nagler, J. and Tucker, J. A. (2016). Of echo chambers and contrarian clubs. *Social Media + Society*, 2(3).

van Leeuwen, T. (2000). Visual racism. In M. Reisigl (Ed.), *The Semiotics of Racism. Approaches to Critical Discourse Analysis*. Vienna: Passagen, pp. 333--350.

Warner, W. and Hirschberg, Julia (2012). Detecting Hate Speech on the World Wide Web. *LSM*, pp. 19-26.

"That spelling tho": A Sociolinguistic Study of the Nonstandard Form of *Though* in a Corpus of Reddit Comments

Marie Flesch
Université de Lorraine, CNRS, ATILF - Nancy, France
marie.flesch@univ-lorraine.fr

Abstract

Tho, the nonstandard spelling of *though* which was proposed by American spelling reformers in the 19th century, is making a comeback. In 2013, internet memes such as *that backflip tho* gave a boost to the shortened form. This sociolinguistic study investigates the use of *tho* in a 17 million-word corpus of comments posted by 1042 Reddit users. Results show that *tho* is rarely used in the meme construction that contributed to popularize it, and that it appears more often as an adverb than as a conjunction. They also seem to indicate that the use of *tho* is correlated with age and ethnicity, with the youngest Redditors and Redditors identifying as Hispanics and Blacks using it the most. This suggests that the shortened spelling is not simply a way to save time when typing, but is a marker of affiliation with a social group and of familiarity with internet subcultures.

Keywords: sociolinguistics, internet slang, Reddit, corpus study, nonstandard spelling, memes

1. Introduction

CMC nonstandard spellings do not seem to have drawn the attention of sociolinguists as much as other Netspeak features such as emoji, emoticons or acronyms. For instance, so far, no large-scale study has focused on the spelling *tho*, or investigated it from a sociolinguistic perspective. This short paper sets out to look at the way the shortened form of *though* is used on Reddit, a popular community website. It presents a quantitative study of a 17 million-word corpus of Reddit comments, drawing on demographic data manually gathered from the content posted by 1042 Redditors.

2. A short history of *tho*

In the late 19th century, American spelling reformers advocated for the use of shortened forms such as *tho*, *thru*, *catalog*, *gard*, *giv*, or *liv* (Marshall, 2011). The shortened spellings of *though* and *through* were again proposed by the Simplified Spelling Board in 1906 (Ranow, 1954) but they seem to have never really caught on ("Tho", n.d.). Today, however, *tho* is making a remarkable comeback online as one of the nonstandard spellings which, together with acronyms, emoticons and abbreviations, make up "Internet slang".

Tho is one of the CMC forms Crystal found in his corpus of tweets (2012), and it was also one of the most frequent nonstandard elements in Kemp's corpus of "textisms" (2010) and Tagliamonte's corpus of email, instant messaging and texting (2016). The Reddit Ngram viewer (King & Olson, 2015), which allows to search Reddit comments from late 2007 to July 2017, shows a steep and steady increase in the use of *tho* on the American community website (Figure 1).

The shortened form seems to have picked up momentum around late 2013, at a time when several *tho* memes circulated on the web. They often followed the construction *that [noun] tho*, sometimes adopting the alternate spelling *dat [noun] doe*. In this construction, *tho* is used "to place a positive emphasis on a particular aspect or feature within a story, image or video that has been shared online" according to the website *Know Your Meme* ("Dat Tho", n.d.).

This same site tried to retrace the history of the meme; it suggests that the slang expression *dat ass*, which was posted on 4chan around 2009, was a precursor of *that [noun] tho*. The meme appears to have spread with the video "Dat Dagger Tho" by gaming YouTuber TSirDiesAlot, which was posted in April 2013, and most notably with a video posted by KingBach in June 2013 on the defunct video service Vine, which received 620,000 likes in a year. Captioned "#ButThatBackflipTho", it shows a young man doing a backflip instead of chasing the thief who has just stolen a woman's purse (Figure 2). The man then proudly says to the camera "Yeah but that backflip though!", pointing out that even though he did not help the woman, his backflip was still impressive.

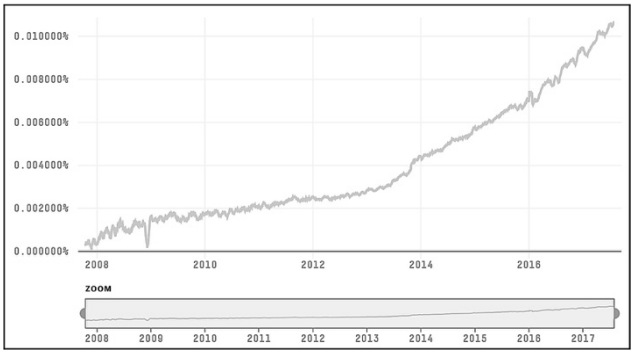

Figure 1: Result of a search for *tho* on the Reddit Ngram viewer

The *y tho* meme, which was posted on Imgur in December 2014, also probably contributed to the spread of *tho* ("Y tho", n.d.). Associated with a painting of Pope Leon X by Botero (Figure 3), it is, according to the *Know Your Meme* site, "a popular slang phrase usually asked in a trolling manner in response to a senseless action or statement".

3. The Reddit corpus

The Reddit corpus was built by the author of the study as part of her PhD research about CMC and gender, which is ongoing. The corpus is not available for consultation. It is made up of comments written by 1042 Redditors, and

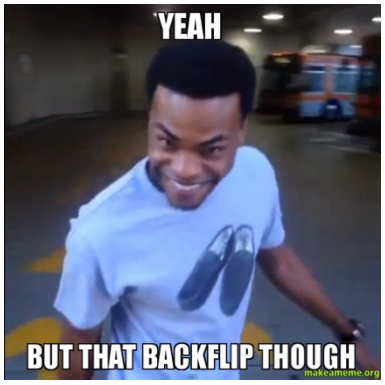

Figure 2: *That blackflip tho* meme[1]

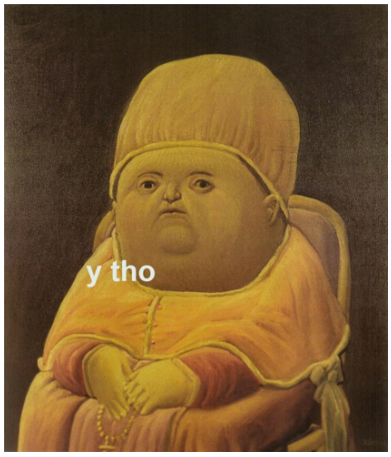

Figure 3: *y tho* meme

contains around 17 million words.

The corpus was designed with the purpose of conducting a sociolinguistic study of CMC; random sampling was thus excluded, Reddit being an anonymous platform. Convenience sampling was used instead. Redditors were selected based on several criteria. First, the number of comments posted by each user needed to be large and balanced, so as to have enough content to examine relatively rare CMC features and to conduct qualitative analyses at a further stage of research. Furthermore, only users who clearly indicated demographic information about themselves were chosen. The data was either searched for in the users' comments with the Ctrl + F search functionality using keywords like "I'm", or was found in their "flairs", little boxes containing information available on certain subreddits.

Gender was the main variable needed, but in many cases age, sexual orientation, occupation, ethnicity and country were also collected. Since Reddit is predominantly white, heterosexual, and male (Barthel et al., 2016), it was decided to over-represent certain categories of the site user base, such as female, LGBTQ, Hispanic, Asian and Black users, in order to be able to study the interaction of gender with other variables.

The comments were collected between March 2017 and July 2017, but often date back to several weeks to two years prior. Of the 1042 Redditors included in the corpus, 78 % are American; 371 are male, 371 are female, 100 are trans-women, 100 are trans-men and 100 are non-binary. It was not possible to collect data about ethnicity for all Redditors. Only 460 users gave information about their ethnic background, of which 92 are Black, 68 are Hispanic, 202 are White, 69 are Asian, and 49 are "Other" (Arab, Native American, mixed ethnicity, etc.).

Age groups are based on Finlay's study of age and gender in Reddit commenting (2014); they correspond to certain stages of educational progress in the United States. Older groups are larger than younger groups in terms of lifespan, because Reddit user base is young: almost two-thirds of users are aged 18 to 29 (Barthel et al., 2016).

4. Results

4.1 Use of *tho* in the corpus

Tho appears 1123 times in the corpus, while the conventional spelling *though* occurs 17,725 times. As a comparison, *thru*, the shortened form of *through*, appears only 167 times, with *through* having a frequency of 11,280. The standard spelling of *though* is only 15.7 times more common than *tho*, while *through* is 67.5 times more frequent than *thru*. *Doe*, a variant associated with African-American Vernacular English (McCulloch, 2015) was much rarer than *tho*; it appeared only 40 times.

Analysis of the concordance lines showed that *tho* does not often occur in the meme construction *that [noun] tho*. Even when all the variants of the meme were taken into account (Table 1), the structure appeared only 53 times. The meme structure also occurred twice with the traditional spelling of *though*. The *y tho* meme was a lot less frequent than *that [noun] tho*: it was used only 3 times.

Further inspection of the concordance lines was conducted in order to see if the shortened spelling had spread to the conjunction *though*, or if it was only used as an adverb, as in the meme. The *Oxford Living Dictionaries tho* entry certainly suggests that it is possible; 9 of its 16 example sentences use *tho* as a conjunction ("tho", n.d.), even though it is unclear where these examples come from. In the corpus, however, *tho* is overwhelmingly used as an adverb, with a frequency of 1016. Only 81 *tho* tokens were conjunctions. By contrast, in a random sample of 1000 standard spellings of *though*, almost a third (306) were used as conjunctions. *Doe* was always used as an adverb.

4.2 Sociolinguistic analysis

Five demographic variables were analyzed: age, gender, sexual orientation, ethnicity, and "Reddit age", meaning the number of years a Reddit user has been registered on the site. It was thought that familiarity with the platform could lead to a greater use of internet slang. Kruskal-Wallis tests were performed to see if there was any significant difference within each of the categories. The results were significant for ethnicity ($p=0.0002$) and age ($p<0.0000001$). A series of Mann-Whitney tests were then conducted. It showed that Hispanics used significantly more *tho* than the Asian ($p=0.0003$) and White groups ($p=0.001$). The results were significant with both the full-size sample of White Redditors and a smaller random sample of White Redditors, designed to compare same-size samples. There was no

[1] The caption was added on a screenshot by an anonymous internet user. KingBach captioned his video #ButThatBackFlipTho

Variant	Examples from the corpus	Raw frequencies
dat [blank] tho	DAT ASS THO But dat cute suit tho.	15
that [blank] tho	That guitar riff tho, and the ending is so cool : (That last pic tho!! Kill it girl	13
the [blank] tho	B-b-but the castration tho But the glow on the bride tho! Mixed babies ftw.	5
Others	Damn son, those eyes tho! Cats tho < 3	14
dat [blank] doe and variants	Dem arms tho. But dat beat tho	6
that [blank] though	Wow, that cast though. But that achievement though	2

Table 1. Meme constructions of *tho* used in the corpus

significant difference between the Hispanic and the Black groups. The difference between the Black group and the White (p=0.004) and the Hispanic (p=0.002) groups was also significant. On average, Black and Hispanic Redditors produced respectively 0.13 and 0.17 nonstandard spellings of *though* per 1000 words, while Asian and White Redditors used *tho* only 0.03 and 0.05 time per 1000 words (Table 2). Thus, even though Redditors identified as Blacks and Hispanics only make up 15% of the sample, they produced 38% of all the nonstandard spellings of *though* (N=428).

	Number of Redditors who used *tho* at least once	Frequency of *tho*, per 1000 words
All Redditors (N = 1042)	304	0.06
Black Redditors (N = 92)	57	0.13
Hispanic Redditors (N = 68)	32	0.17
White Redditors (N = 202)	57	0.05
Asian Redditors (N = 69)	15	0.03

Table 2: Use of *tho* in the corpus, per ethnic group

Significance tests also showed a correlation between age and use of *tho*. The frequency of the nonstandard form decreased with age, with the youngest Redditors using it more than the older groups (Table 3).

Mann-Whitney tests showed that, for instance, the age group 17-18 used *tho* significantly more than all the oldest groups, with a p-value of 0.002 when compared with the age group 23-27, and $p < 0.00001$ when compared with Redditors age 28-35. Significance tests revealed no significant differences for sexual orientation or Reddit age. As for gender, the results of the Mann-Whitney tests only showed significant differences between the female, the male and non-binary groups, females tending to use *tho* less than males (p=0.03) and non-binary Redditors (p=0.03).

Age groups	Reditors who used *tho* at least once	Frequency of *tho*, per 1000 words
12-16 (N = 24)	16	0.15
17-18 (N = 44)	23	0.17
19-22 (N = 154)	54	0.07
23-27 (N = 246)	77	0.06
28-35 (N = 288)	70	0.05
36-45 (N = 132)	22	0.03
46+ (N = 54)	6	0.02

Table 3: Use of *tho* in the corpus, per age group

5. Discussion of results

Given the significant differences of usage observed in the corpus, it is unlikely that *tho* is just a shorter form of *though*, which would have the sole purpose of typing faster. It seems to have connotations that *though* does not have, through its connection to the *that [noun] tho* meme. Since age has been shown to be correlated with the use of CMC forms such as nonstandard spellings or emoticons (Sánchez-Moya & Cruz-Moya, 2015; Oleszkiewicz et al., 2017), it is not surprising that younger Redditors tend to use *tho* a lot more than older Redditors as a result of age grading. The findings about ethnicity are maybe more unexpected. Hispanics and Blacks appeared to have adopted *tho* more readily than other Redditors. It could be that CMC nonstandard spellings are some of the linguistic strategies they use to differentiate themselves from the overwhelmingly white Reddit user base. *Tho* would then be a marker of affiliation with a social group as well as a sign of familiarity with memes and internet subcultures.

6. Limitations

The greatest limitation of this study is perhaps its lack of generalizability. The convenience sample is not representative of Reddit, or of the American population. Furthermore, the demographic data collected may not be fully accurate. It was assumed that the demographic information available in Redditors' comments is true, but it may not always be reliable.

The make-up of ethnic categories can be problematic, especially for the Hispanic category or the Asian groups, which may contain users of different ethnic backgrounds.

7. Future work

Further examination of the corpus would allow to know what other CMC elements Black and Hispanic Reddit users have adopted more readily than other ethnicities. Factorial analyses performed with all the demographical variables collected during the corpus building process, including country of origin or occupation, could also reveal trends that this study did not. It would be interesting to see how *tho* is used on other platforms, such as Twitter, Instagram or instant messaging services. Studies about ethnicity and CMC would allow to describe different "internet dialects", or shine a light on the influence of AAVE on internet slang.

8. References

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, February 25). Reddit news users more likely to be male, young and digital in their news preferences. Retrieved October 28, 2017, from http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Crystal, D. (2011). *Internet linguistics: a student guide*. Milton Park, Abingdon; New York, NY: Routledge.

Dat Tho | Know Your Meme. (n.d.). Retrieved April 28, 2018, from http://knowyourmeme.com/memes/dat-tho

Finlay, S. C. (2014). Age and Gender in Reddit Commenting and Success. *Journal of Information Science Theory and Practice, 2*(3), pp. 18–28. https://doi.org/10.1633/JISTaP.2014.2.3.2

Kemp, N. (2010). Texting versus txtng: reading and writing text messages, and links with other linguistic skills. *Writing Systems Research,* 2(1), pp. 53–71.

King, R., & Olson, R. (2015, November 18). How The Internet* Talks. Retrieved April 28, 2018, from https://projects.fivethirtyeight.com/reddit-ngram/

Marshall, D.F. (2011). The Reforming of English spelling. In J. Fishman & O. Garcia (Eds.), *Handbook of Language and Ethnic Identity* (vol 2). New York, NY: Oxford University Press, pp. 113-125.

McCulloch, G. (2015, May 27). The Evolution of "That [Noun] Though." Retrieved April 19, 2018, from http://mentalfloss.com/article/64323/evolution-noun-though

Oleszkiewicz, A., Karwowski, M., Pisanski, K., Sorokowski, P., Sobrado, B., & Sorokowska, A. (2017). Who uses emoticons? Data from 86702 Facebook users. *Personality and Individual Differences,* 119 (Supplement C), pp. 289–295.

Ranow, G. R. (1954). Simplified Spelling in Government Publications. *American Speech,* 29(1), pp. 36–44. https://doi.org/10.2307/453594

Sánchez-Moya, A., & Cruz-Moya, O. (2015). Whatsapp, Textese, and Moral Panics: Discourse Features and Habits Across Two Generations. *Procedia - Social and Behavioral Sciences*, 173, pp. 300–306.

Tagliamonte, S. A., & In collaboration with Dylan Uscher, Lawrence Kwok, and students from HUM199Y 2009 and 2010. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, 45(1), pp. 1–32.

tho' | Definition of tho' in English by Oxford Dictionaries. (n.d.). Retrieved April 28, 2018, from https://en.oxforddictionaries.com/definition/tho'

Tho | Definition of Tho by Merriam-Webster. (n.d.). Retrieved April 28, 2018, from https://www.merriam-webster.com/dictionary/tho

Y Tho | Know Your Meme. (n.d.). Retrieved April 28, 2018, from http://knowyourmeme.com/memes/y-tho

The Myth of the Digital Native? Analysing Language Use of Different Generations on Facebook

Jennifer-Carmen Frey, Aivars Glaznieks
Eurac Research, Institute for Applied Linguistics, Bolzano, Italy
JenniferCarmen.Frey@eurac.edu, Aivars.Glaznieks@eurac.edu

Abstract

Digital Natives, i.e. people who grew up in a digital world, are said to be different to their counterparts, digital immigrants, regarding their communication habits and use of digital services. In this paper, we investigate the linguistic behavior of digital natives compared to digital immigrants in a sociolinguistically annotated corpus of personal Facebook texts using methods from corpus linguistics, computational sociolinguistics and data mining. The texts are data donations from the profiles of 133 users of various ages from the northern Italian province of South Tyrol. In order to investigate if and how digital natives differ from older generations with respect to language choice, variety choice and the use of style markers, we use three analysis methods: (1) we disclose and compare central tendencies of the two groups in a quantitative analysis, (2) we train text classifiers to distinguish both groups automatically and compare prediction results, and (3) we investigate a ranking of features. The two groups differ in particular in their use of language varieties. However, taking into account the user's first language, their choice of language and use of CMC-specific style markers also differ significantly.

Keywords: Facebook, CMC, youth language, sociolinguistics

1. Introduction

In 2001, Prensky published an essay on the distinctiveness of post- and pre-digitalization generations (Prensky, 2001), which he named *digital natives* and *digital immigrants*, respectively. The digital natives, i.e. people who were born in an already digital era and hence grew up with computers and other digital devices, were said to be different to their older counterparts, the digital immigrants, with regard to communication habits and their use of digital services, for example. Since then, several studies from domains like sociology and pedagogy have investigated his claim, trying to figure out if and how both generations differ (e.g. Palfrey and Gasser 2013, Kennedy et al. 2008, Bennett et al. 2008, Helsper and Eynon 2010). However, there is a lack of empirical linguistic investigations of such "generational" differences due to the unavailability of socio-linguistically annotated data that could represent such differences. Without doubt, age is a relevant category in computer-mediated communication (CMC) and its impact on writing has been further acknowledged in recent studies (Hilte et al., 2016; Glaznieks and Glück, Forthc; Peersman et al., 2016; Verheijen, 2017). However, we are not aware of any linguistic study investigating Prensky's note on post- and pre-digitalization generations. In this paper, we used the DiDi Corpus of South Tyrolean CMC (Frey et al., 2016) to investigate linguistic differences in the writings of digital natives and digital immigrants. We will focus our analysis on three characteristics of the investigated texts: (a) the writer's choice of language, (b) his/her choice of language variety and (c) the use of style markers that are specific to CMC.

We will start with a brief overview of the data used for this analysis (section 2.) followed by a detailed description of our approach and the methodology used (section 3.). In section 4., we report on the results obtained with regard to the two groups and summarize them in section 5.

2. Data: The DiDi Corpus

The data we used for our investigation is a corpus of Facebook texts published on the personal Facebook accounts of 133 voluntary data donors from South Tyrol. The so-called DiDi Corpus (Frey et al., 2016) is a multilingual corpus that contains in total around 40,000 texts (~11,000 status updates, ~6,500 comments and ~23,000 chat messages) from German and Italian native speakers and provides socio-demographic metadata such as gender, first language, education and age (collected via a questionnaire that was filled in by the data donors) for each text. The data donors were recruited via a Facebook application following the necessary privacy restraints and obligations (cf. Frey et al. 2014).

For the analysis described in this paper, we used three types of information on language use provided in the corpus:

Languages: The corpus provides language labels for each text that are based on a semi-automatic annotation[1]. The labels state the predominant language of the text, ignoring any kind of code-switching. The main languages in the corpus are German (58.7%), Italian (20.9%) and English (9.5%). Texts exclusively composed of non-language elements such as emoticons or hyperlinks are labeled as "non-language" texts.

Varieties: The corpus provides variety labels for all German-tagged texts. The variety labels are: dialect (contains dialect-specific lexical items and/or a high ratio of non-standard spellings), non-dialect (no dialect-specific items, a very low amount of non-standard spellings) or an undefinable variety (text too short to classify or contains mixed spellings).

CMC style markers: The corpus provides labels for style markers frequently named in the literature on CMC (Crystal, 2001; Vandergriff, 2013; Darics, 2013; Androutsopoulos, 2011), namely acronyms, emoticons, emojis, hashtags, hyperlinks, @mentions and iterations of graphemes. As CMC style markers are provided on token level, we will use the total number of style markers (and the number per subcategory) normalized for text length for our investigation.

With reference to Palfrey and Gasser (2013) and Bennet

[1] For further details see: Frey et al. (2016).

(2008), we split our data donors into two groups: people born from 1980 onwards (i.e. digital natives) and people born before 1980 (i.e. digital immigrants). Accordingly, 42% of the writers were classified as digital natives and 58% as digital immigrants. While digital natives and immigrants are almost equally represented in terms of writers, immigrants produced significantly more texts (66% of all texts compared to 34% written by digital natives). Table 1 gives an overview of available profiles and texts for both groups.

	profiles	texts	mean	sd
Digital Natives	56	13,529	242	439.2
Digital Immigrants	77	26,296	342	516.0

Table 1: Overview of profiles and texts in the DiDi Corpus

3. Methodology

We explored three strategies for our analysis of the use of languages, varieties and CMC style markers by digital natives vs. immigrants.

First, we conducted a manual statistical analysis and compared measures of central tendencies for the investigated features for both groups. We used the Mann-Whitney U test and Student's t-test to check the statistical significance (.95 confidence level) of the averaged differences.

Secondly, we applied a data mining approach comparing prediction performances of different text classifiers using machine learning. In particular, we based our research on other studies in author profiling, computational sociolinguistics (Nguyen et al., 2016) and age prediction, in which machine learning is used to predict author characteristics on the basis of their texts (Rosenthal and McKeown, 2011; Nguyen et al., 2013; Schler et al., 2006). We trained a number of text classifiers to distinguish digital natives and digital immigrants on the basis of our selected features. Then we evaluated accuracy and F-measures using 10-fold cross validation (CV) in order to validate the classifier's ability to learn underlying relations in the data. Although more sophisticated methods like neural networks would probably provide better prediction results, we used a decision tree algorithm (J48 implementation of WEKA (Witten et al., 2016)) to build our classifiers, because we were rather interested in the interpretation of the models than in reaching high accuracies.

Finally, we used a feature ranking method to check for the most informative features as it is frequently carried out in computational sociolinguistics (e.g. Simaki et al.2016, Vajjala 2017).

4. Results

In the following section we report the results of the three approaches described above.

4.1. Comparing central tendencies

Since the majority of the users in the DiDi Corpus stated German as their L1, we only used texts from L1 German users for our statistical analyses to remove potential interactions (e.g. regarding L1-dependent language choice). Furthermore, we excluded all users who wrote less than 10 texts in order to account for data skewness. The analysed subset thus contained 29,808 texts from 90 users. Table 2 shows the calculated measures of central tendencies for both groups for each feature and the corresponding p-values of the significance tests.[2]

Feature	Natives	Immigrants	p
German	70.83%	83.33%	0.1
Italian	1.09%	5.66%	0.003
English	9.01%	2.08%	1e-04
non-lang.	13.71%	5.72%	3e-05
dialect	41.94%	10.91%	5e-06
non-dialect	15.38%	43.07%	1e-05
CMC (tokens per text)	1.205	0.762	9e-05

Table 2: Comparison of central tendencies

Languages: After calculating the proportion of each language per user, we used median values to aggregate over both groups and performed a two-tailed Mann-Whitney U test ($\alpha = 0.05$) to test if the differences are statistically significant. The results show significant differences for the use of English, Italian and non-language texts between digital natives and digital immigrants (see Table 2). While there is no significant difference with regard to the use of German, the natives use significantly more English, produce more non-language texts and use less Italian than the digital immigrants.

Varieties: Per user, we compared the percentages of dialect and non-dialect texts of all German-tagged texts (in total 20,337 of 29,808 texts of the subset) averaged for both groups. As averages were not distributed normally, we used the median to average the percentages for the two groups. The results show a significant difference in the use of varieties of German between digital natives and digital immigrants. Digital natives wrote significantly more dialectal texts than immigrants when writing in German (Table 2).

CMC style markers: We calculated the average number of CMC style markers per text for each user and compared mean values for digital natives and immigrants (as the values were normally distributed). A two-tailed Student's t-test showed a significant difference between digital natives and digital immigrants. As can be seen in Table 2, natives used more CMC style markers (1.21 per text) on average than immigrants (0.76 per text).

4.2. Comparing prediction results

In our second approach, we trained a number of decision tree classifiers to label texts automatically on the basis of the provided features, instead of meticulously sampling our data and analysing aspect per aspect individually. We compared the results for classifiers with different feature combinations and controlled the effects of class imbalance and first language as a confounding factor using both the whole data set as well as the subset for training.

[2]Percentage values are median proportions per user of the group, CMC style markers represent the users' average amount of CMC-specific tokens per text, aggregated for the group.

The classification performance of our classifier, trained with all three feature categories (language, variety and number of CMC style markers) on the whole data set, proved to be significantly above the baseline (71.2% accuracy compared to 66.03%, which would be achieved when always assigning the majority class).
Table 3 shows the classification results for the different feature categories and combinations of categories.[3] When investigating each feature category individually, we found that only variety choice gave prediction results that were significantly above the baseline. However, when accounting for the interaction between a users' first language and his/her language choice by using only the L1 German subset, we could also achieve performances above the baseline with the language feature category. The number of CMC style markers, when used exclusively, did not achieve any performance improvement to the baseline. However, in combination with other features, CMC style markers contribute to the overall classification result.

Feature	Whole corpus		L1 German subset	
	Acc.	F-Score	Acc.	F-Score
CMC	0.661	0.53	0.572	0.42
Language	0.660	0.53	0.592*	0.51
Variety	0.704*	0.67	0.675*	0.68
CMC + Lang.	0.667*	0.55	0.598*	0.53
CMC + Variety	0.706*	0.68	0.674*	0.67
Lang. + Variety	0.703*	0.67	0.695*	0.70
All	0.712*	0.69	0.700*	0.70
Baseline	0.660	0.53	0.572*	0.42

Table 3: CV results for different feature combinations

4.3. Feature ranking

Table 4 shows a feature ranking based on the information gain metric. According to the ranking, the use of Italian, the use of the South Tyrolean dialect and the use of the non-dialect variety in German texts are the highest ranked and thus the most informative features to distinguish digital natives from digital immigrants in the DiDi Corpus.

Rank	Feature	InfoGain
1	Lang_IT	0.077
2	Var_dialect	0.052
3	Var_non-dialect	0.026
4	Lang_DE	0.022
5	Lang_non-lang.	0.008
6	CMC	0.003
7	Lang_EN	0.0004

Table 4: Information Gain ranking

5. Conclusion

In this paper, we approached the distinction of digital natives and digital immigrants using three different methods, a) calculating central tendencies for both groups and testing for statistical significance, b) training a text classifier to apply a data mining strategy based on machine learning and c) calculating the most informative features by applying a feature ranking method.
The results of this study show that the investigated features of language choice, variety choice and the use of CMC style markers have proven informative for the distinction of texts written by digital natives and digital immigrants in the DiDi Corpus.
The compared measures of central tendencies showed statistically significant differences between digital natives and digital immigrants for all investigated features. The digital natives used more English as well as more dialectal writings. They also used significantly more CMC style markers, but less Italian.
The data mining approach based on text classification with decision trees similarly showed relations between the choice of both language and variety, the use of CMC style markers and the categorization of the writer as digital native or digital immigrant.[4]
In the manual investigation, all features were analysed individually using a well-defined subset. The machine learning approach provided further possibilities to test feature combinations as well as to test and rank more fine-grained features. However, the data mining approach was also sensitive to the interaction between users' first language and their language choice. When using individual feature categories for training on the whole data set, language features could not achieve performance above the baseline. This shows us that, for this approach too, methods should not be used without critical reflection, especially when relatively small data sets are used.
Furthermore, we saw that variety choice was the most important feature for the automatic text classification to discriminate between both groups. However, investigating the features individually, the use of Italian as an L1 German speaker, the use of the South Tyrolean dialect in German texts and the use of a non-dialect variety were the most important features for text classification.
The relevance of these features is also reflected in the results of the information gain calculation which ranked the use of Italian as most informative feature, followed by the use of the dialect and non-dialect variety in German texts.
The results support the general impression that South Tyrolean writers from the younger generation are more open to using different global and local varieties in CMC. In addition, they are more open to various writing styles, comprising non-language texts and texts with a high amount of CMC style markers. However, whether this originates from being a digital native or from belonging to different social groups with different communication habits cannot be answered with our data. The fact that older generations composed more texts in Italian than the younger generation (their second language with a high local and national value)

[3] Values are weighted averages for 10-fold CV. Values with asterisk are significantly higher than a baseline accuracy achieved when always assigning the majority class.

[4] Although the performance of the trained text classifiers was not particularly high (around 71%), we still accept this result as an indication to answer our linguistic research question, as we were interested in the inherent structure of the data and not in the prediction of age groups.

might also hint at societal changes (in the region or in general) in which younger people are internationally connected and English becomes more and more important.

6. Outlook

In future work, we plan to extend this research in two directions. First, by questioning the split of the age groups at the year 1980. For this, we want to compare different splits of age groups based on the numerical age, as well as taking into consideration alternative age concepts based on digital media experience (cf. Glaznieks and Stemle 2014). Second, methodologically, by using more sophisticated models for the statistical analysis (mixed-effects models to consider random effects) and extended feature sets for the classification approach (e.g. phenomena of multilingualism, shallow features like word or character n-grams).

7. References

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. *Standard Languages and Language Standards in a Changing Europe*, pages 145–159.

Bennett, S., Maton, K., and Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British journal of educational technology*, 39(5):775–786.

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

Darics, E. (2013). Non-verbal signalling in digital discourse: The case of letter repetition. *Discourse, Context and Media*, 2(3):141–148.

Frey, J.-C., Stemle, E. W., and Glaznieks, A. (2014). Collecting language data of non-public social media profiles. In Gertrud Faaß et al., editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference (2014)*, pages 11–15, Hildesheim, Germany, oct. Universitatsverlag Hildesheim, Germany.

Frey, J.-C., Glaznieks, A., and Stemle, E. W. (2016). The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In Anna Corazza, et al., editors, *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), 5-6 December 2016, Napoli*, pages 157–161, Torino. Academia University Press.

Glaznieks, A. and Glück, A. (Forthc.). From the Valleys to the World Wide Web: Non-Standard Spellings on Social Network Sites. In Egon W. Stemle et al., editors, *Postvolume Monograph of the 5th CMC-Corpora Conference*. Clermont Auvergne University Publishing House, Clermont Auvergne.

Glaznieks, A. and Stemle, E. W. (2014). Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *Journal for Language Technology and Computational Linguistics (JLCL)*, 29(2):31–57, dec.

Helsper, E. J. and Eynon, R. (2010). Digital natives: where is the evidence? *British educational research journal*, 36(3):503–520.

Hilte, L., Vandekerckhove, R., and Daelemans, W. (2016). Expressiveness in Flemish Online Teenage Talk: A Corpus-Based Analysis of Social and Medium-Related Linguistic Variation. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia*, pages 30–33.

Kennedy, G. E., Judd, T. S., Churchward, A., Gray, K., and Krause, K.-L. (2008). First year students' experiences with technology: Are they really digital natives? *Australasian journal of educational technology*, 24(1):108–122.

Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). 'How old do you think I am?': A study of language and age in Twitter. In *Proceedings of the seventh international AAAI conference on weblogs and social media, 8-11 July 2013, Cambridge, Massachusetts, USA*, pages 439–448.

Nguyen, D., Dogruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

Palfrey, J. G. and Gasser, U. (2013). *Born digital: Understanding the first generation of digital natives*. Basic Books.

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. *arXiv preprint arXiv:1601.02431*.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon*, 9(5):1–6.

Rosenthal, S. and McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Simaki, V., Mporas, I., and Megalooikonomou, V. (2016). Evaluation and sociolinguistic analysis of text features for gender and age identification. *American Journal of Engineering and Applied Sciences*, 9(4):868–876.

Vajjala, S. (2017). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 18:1–27.

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

Verheijen, L. (2017). WhatsApp with social media slang?: Youth language use in Dutch written computer-mediated communication. In Darja Fišer et al., editors, *Investigating Computer-Mediated Communication. Corpus-Based Approaches to Language in the Digital World*, pages 72–101. Ljubljana University Press, Ljubljana.

Witten, I. H., Frank, E., Hall, M., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

The Flow of Ideologies Between a Political Figure and a Militant Community: A CMC Corpora Analysis

Julien Longhi[1], Claudia Marinica[2], Zakarya Després[3], Clément Plancq[4]
[1] Paris Seine University, EA7392 AGORA, IDHN
[2] Paris Seine University, ETIS (ENSEA/UCP/CNRS UMR 8051), IDHN
[3] Paris Seine University, IDHN
[4] Paris 3 University, Lattice (ENS/ CNRS UMR 8094)
julien.longhi@u-cergy.fr, claudia.marinica@u-cergy.fr, zakarya.despres@gmail.com, clement.plancq@ens.fr

Abstract

This paper presents an exploratory work on the valorisation of political ideologies in CMC corpora, and more precisely their flow between a political figure and a militant community. Our main goal is to put forward a methodology for analysis and to present our preliminary results. We consider that the CMC corpora we used are traces that allow us to answer to political and social interrogations: first we present the context of the study, the corpora, and the addressed themes. Then, we present our methodology, and the intermediary results obtained with two softwares (Iramuteq and TXM), that led us to build an original analysis method. Finally, we compare the results from both corpora to provide hypothetical answers to the political question of the construction and circulation of political ideas from a movement with a charismatic leader and an important militant community.

Keywords: CMC corpora, textometry, digital spaces, tweets, forums.

1. Introduction

This article is the continuation of Djemili et al. (2014), Longhi (2017), Longhi et al. (2017), Plancq et al. (2018) and Marinica et al. (2018), and it aims to complete the analyses of the digital political discourse on Twitter through other CMC communication mediums, like the forums. To this end, we chose to study the tweets and forum threads during the 2017 French presidential election. We collected all tweets from the account @JLMelenchon, which belongs to the leader of the movement and candidate to the election, as well as posts from the "Blabla 18-25" forum on the website "Jeuxvideo.com", and messages from the *"Discord des Insoumis"*.

We consider that the CMC corpora we use are traces that allow us to answer to political and social interrogations; by doing so, we deal with complex questions (defining ideologies, identifying their possible flow, creating a timeline for the emergence of ideas and political assertions) by computer processing, on a basis of structured, comparable social data.

More precisely, this paper presents exploratory work on the valorisation of political ideologies in CMC corpora, especially their potential flow between a political figure and a militant community. Our main goal is to put forward methods and analyses, supported by a set of results that open up new questions.

This paper is structured as follows: first, we present the context of the study, the corpora, and the addressed issues. Then, we present our methodology, and the intermediary results obtained with two softwares, that led us to build an original analysis method. Finally, we compare the results from both corpora to answer the political question of the construction and flow of political ideas from a movement with a charismatic leader and an important militant community.

2. Context, data and goals

2.1 Context

In this work, we focus on the candidate Jean-Luc Mélenchon and on the *France Insoumise* militant community, who named their members the *Insoumis* (the insubordinates), as they expressed online and offline.

Mélenchon is different from the other candidates to the 2017 French presidential elections, first, because he placed his digital strategy at the centre of his campaign: he was active on social media and he had a YouTube channel. Second, he was supported by self-organised militant relays with the *Discord* server created by the *Insoumis*, an idea which started in the *Jeuxvideo.com* forums. These two distinct digital spaces, an official one and a participative one, fit our objectives of analysing the circulation of *ideologies*.

According to Knight (2006), "specific *ideologies* crystallize and communicate the many beliefs, opinions and values of an identifiable group". The use of a statistical method will allow us to compare corpora, and to infer their ideology, without having to subjectively decide on the polysemous status of terms, or their possible interpretation.

2.2 Data

We used two corpora for our study:

- 3,036 tweets from the Twitter account @JLMelenchon, representing 51,552 words;
- messages from Insoumis militants on different platforms.

This second corpus comes from two sources: a first part comes from a series of threads dedicated to Jean-Luc Mélenchon and France Insoumise on the *Blabla 18-25* forum of Jeuxvideo.com, a generalist forum on a gaming website. Then, we extracted the messages that militants

sent on the Discord server, an instant messaging platform which is also originally gaming-related. In total, the militant corpus represents 383,403 messages with 6,850,823 words.

2.3 Goals of the paper

The objective of the study is to create a methodology that, according to the temporality of the messages, would allow us to see how the themes, terms and ideologies from the candidate Mélenchon are spreading in the militant community, and/or how the discussions of the militants can bring substance to the candidate's digital discourses.

3. Methodology and results

In order to assess the flow of terms and to quantify it, we used a method bringing together two tools and two different perspectives.

In the first place, a textometric study (statistical text analysis) was performed with the software Iramuteq (http://iramuteq.org), allowing us to detect a set of lexical classes. These classes, along with the terms composing them, are used in the second software, TXM (http://textometrie.ens-lyon.fr/), in order to understand the temporality and the relations between the terms.

The Iramuteq software offers a set of analysis procedures for the description of a textual corpus. One of its principal methods is Alceste. This allows a user to segment a corpus into *context units*, to make comparisons and groupings of the segmented corpus according to the lexemes contained within it, and then to seek *stable distributions* (Reinert, 1998).

In addition to the Alceste method, Iramuteq provides other analysis tools including prototypical analysis, similarities analysis, and word clouds analysis. All of these methods allow the users of this tool to map out the dynamics of the discourses of the different subjects engaged in interaction (Reinert, 1999).

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

More precisely, as explained by Camargo & Justo (2016), text segments (TS) are clustered according to their vocabularies and distributed according to the reduced forms frequencies. The descending hierarchical analysis uses matrices that cross reduced forms with TS (in repeated texts of X2 type). This method allows users to obtain a definitive classification. We obtain TS clusters with similar vocabulary within, but different from other segments. The software computes descriptive results of each cluster conforming to its main vocabulary and words with asterisk (variables). This analysis gives another way of presenting data, derived from a correspondence factor analysis. Based on the chosen clusters, the software calculates and provides the most typical TS of each cluster, giving context to them. These word clusters and TS integrate several segments according to the vocabulary distribution.

The authors explain that on the interpretative level, it depends on the theoretical scope of the research. For example, Reinert (1990), when studying French literature, considered each cluster as a "world", a cognitive-perceptive framework with a certain temporal stability related to a complex environment. From another point of view, research in linguistics considers these clusters as lexical fields or semantic contexts.

Figure 1: The result of the Hierarchical Descending Classification

This analysis highlights four disjoint classes presented in Figure 1, which can be interpreted as four major themes / areas of Mélenchon's tweets: Class 1 (nearly 29.9% of the vocabulary) concerns the economy; Class 2 (29.2%) concerns Europe and related defense issues; Class 3 (15.6%) concerns institutional issues; and Class 4 (25.2%) mainly includes SEO or communication terms. Each class is described by a set of words that will be used further in the analysis; one word belongs to only one class. In the following analyses we focus on the first three classes.

The following analyses were processed with the textometry software TXM, with which we partitioned the corpora according to the dates of each tweet or message. This allowed us to get the daily frequency of each word extracted from the classes found with Iramuteq. In order to visualise the evolution in time of the utilisation of these terms, we used RAWGraphs (https://rawgraphs.io) to produce the following horizon graphs, readable as a heatmap-like timeline.

4. Results and interpretations of the results

Following this method, we obtained several types of results: general results about the presence of classes during our time period (without differentiating specific words), and specific class by class analyses with a visualisation of the repartition of each word in the class. For this paper, we will focus on the two last weeks of February 2017, which provided the most relevant results. In this period, we collected 659 tweets from Mélenchon, and 50954 messages from the Insoumis.

4.1 Global analysis of the repartition of classes

The global analysis that we carried out is presented in Figure 2. Over the month, we found 147 occurrences of words from Class 1 in Mélenchon's tweets (JLM) and 1700 in Insoumis' messages, 24 words of Class 2 from JLM and 1033 from Insoumis, and 85 words of Class 3 from JLM and 705 from Insoumis. In order to better represent the frequency peaks in each corpus despite their size discrepancy, the following figures use different scales for JLM tweets and Insoumis messages.

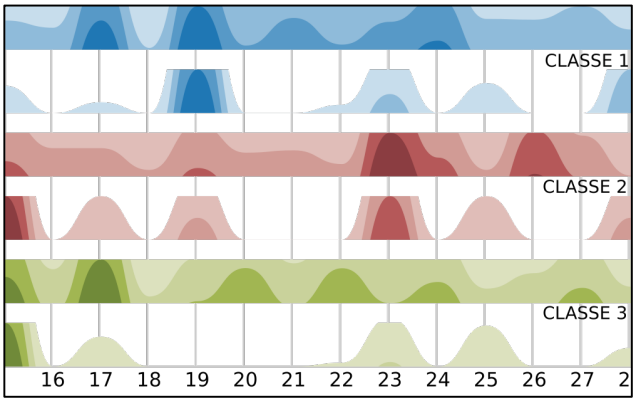

Figure 2: Frequency of the classes in February 2017. Top: Insoumis, bottom: Mélenchon.

Over the period, we can observe the temporal repartition of the apparition of classes, that allows us to evaluate if a dependence between the candidate Mélenchon and the Insoumis community exists with the use of lexicons that can assess the presence or absence of some themes.
If we take a closer look at Class 1 in Figure 2, dealing with economical issues with words like *impôt* (taxes), *euro*, *payer* (to pay), etc., we can see that terms which are frequent in the candidate's tweets (on the 19th of February) are used later in the community (on the 19th, 23rd and 25th of February). But, this class was also used on the 17th of February in the community, thus, we cannot conclude on the flow of Class 1 without having a more specific analysis of the class word by word. We can just admit that the dynamic is not unilateral, and the discursive relationships between Mélenchon and the Insoumis are more complex than they appear. To verify this, we propose to focus on the terms of Class 1.

4.2 Specific analysis of the economy-themed words

By zooming in on Class 1, we find the following terms: *coûter* (to cost), *euro, impôt* (tax), *milliard* (billion), *payer* (to pay), *retraite* (retirement), *salaire* (salary), *santé* (health), *SMIC* (minimum salary), *social*, *sécurité 'sociale'* (social security system), *travail* (work/job).
By projecting the frequency of several of these terms (the more relevant/interesting ones) on a temporal axis, we obtain the visualisation in Figure 3.

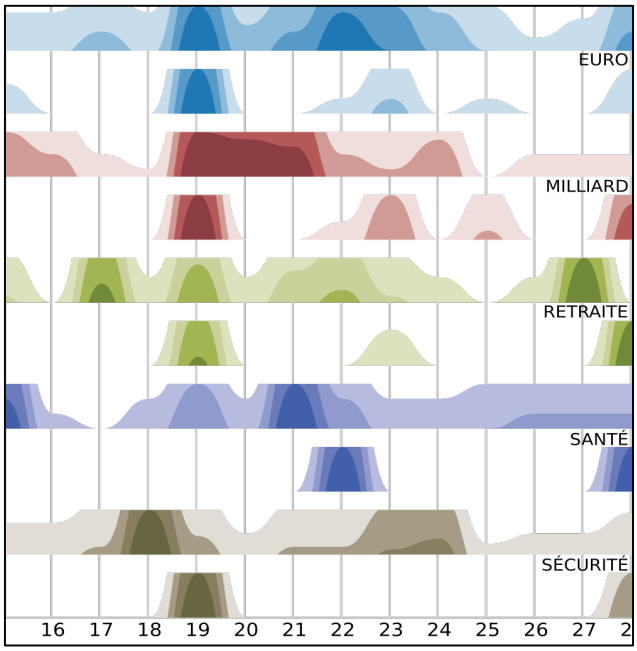

Figure 3: Frequency of the words from Class 1 in February 2017. Top: Insoumis, bottom: Mélenchon.

Economy-themed words are used a lot on the 19th of February, and continue to make echo in the community, like *milliards* or *euros*. Other terms, like *retraite*, *santé* and *sécurité 'sociale'*, seem to come from the community first, and are then reused by the candidate. In Figure 4 we can see an example of a tweet by Mélenchon, posted on February 19th, which contains the words "milliard" and "euros", saying "We are going to invest 7 billion euros in public service".

Figure 4: Tweet of @JLMelenchon on 19/02/17.

Three days after, we find discussions on the same themes using the same terms on Discord, such as the one presented in Figure 5 saying "If you do the math, for less than 10 billion, you can raise the RSA to 1100 euros per month".

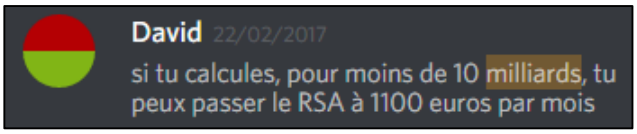

Figure 5: Message on Discord on 22/02/17.

For comparison purposes, the terms from Class 3, that refers to institutional issues, are presented in a different way, as shown in Figure 6. The temporality of these terms, but also their intensity, show that they concern broader discussions for the campaign, and that the candidate uses them occasionally.

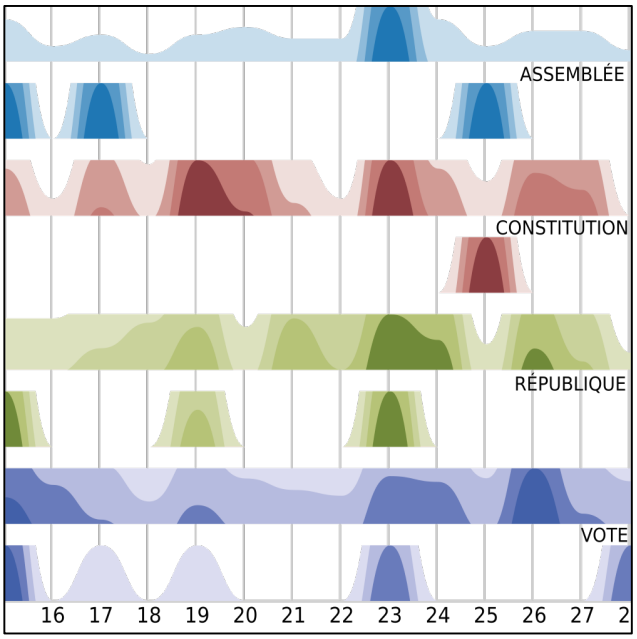

Figure 6: Frequency of words from Class 3 in February 2017. Top: Insoumis; bottom: Mélenchon.

5. Conclusions

This paper, still being exploratory, on the ideological circulation between political discourse and militants, allowed to build a method that highlights political themes, through the Iramuteq software, then showcases their quantitative and temporal values thanks to the TXM and RAW Graphs softwares. These preliminary results provide important perspectives on the work and ongoing analyses will bring us precisions on the link between the visualisation lines, and locate these phenomena accurately in the corpora. The current results provide already interesting insights on how to grasp this complex phenomenon, and to measure the porosity between two types of CMC corpora.

6. References

Camargo, B., Justo, A. (2016). Iramuteq Tutorial http://iramuteq.org/documentation/fichiers/IRaMuTeQ%20Tutorial%20translated%20to%20English_17.03.2016.pdf

Djemili, S., Longhi, J., Marinica, C., Kotzinos, D., Sarfati G.-E. (2014). What does Twitter have to say about ideology? In Gertrud Faaß & Josef Ruppenhofer (Eds.). *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media - Pre-conference workshop at Konvens Oct 2014, Germany*. Universitätsverlag Hildesheim, 1, pp. 16--25. <halshs- 01058867v2>

Kathaleen, K. (2006). Transformations of the concept of ideology in the twentieth century. *American Political Science Review*, pp. 619--626.

Longhi, J., Marinica, C., Haddioui, N. (2016). Extraction automatique de phénomènes linguistiques dans un corpus de tweets politiques : quelques éléments méthodologiques et applicatifs à propos de la négation. In *Res per nomen 5, Négation et référence*, Reims: EPURE.

Longhi, J., Saigh, D. (2016). A textometrical analysis of French arts workers "fr.Intermittents" on Twitter. In *Proceedings of the 4th Conference CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenia, pp. 44--47.

Longhi, J., Marinica, C., Hassine, N., Alkhouli, A. Borzic. B. (2017). The #Idéo2017 Platform. *Conference on Computer-Mediated Communication and Social Media Corpora for the Humanities*, Oct 2017, Bolzano, Italy. <hal-01619236>

Marinica, C., Longhi, J. Hassine, N., Alkhouli, A. Borzic, B. (2018). #Idéo2017: une plateforme citoyenne dédiée à l'analyse des tweets lors des événements politiques. In *Extraction et Gestion des Connaissances (EGC)*, 2018.

Plancq, C., Després, Z., Longhi, J. (2018). " L'avenir en commun" des Insoumis. Analyse des forums de discussion des militants de la France Insoumise. In *Atelier Fouille de Données Complexes*, EGC 2018, Jan 2018, Paris, France. <halshs-01719374>

Reinert, M. (1990). ALCESTE, une méthodologie d'analyse des données textuelles et une application: *Aurélia* de G. de Nerval. *Bulletin de méthodologie sociologique*, 28, pp. 24--54.

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

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

Iramuteq: www.iramuteq.org

TXM: http://textometrie.ens-lyon.fr

Reply Relations in CMC: Types and Annotation

Harald Lüngen[1], Laura Herzberg[2]
[1] Institut für Deutsche Sprache, R5 6-13, D-68161 Mannheim
[2] Universität Mannheim, Germanistische Linguistik, Schloss, D-68131 Mannheim
luengen@ids-mannheim.de, herzberg@uni-mannheim.de

Abstract

This paper analyses reply relations in computer-mediated communication (CMC), which occur between post units in CMC interactions and which describe references between posts. We take a look at existing practices in the description and annotation of such relations in chat, wiki talk, and blog corpora. We distinguish technical reply structures, indentation structures, and interpretative reply relations, which include reply relations induced by linguistic markers. We sort out the different levels of description and annotation that are involved and propose a solution for their combined representation within the TEI annotation framework.

Keywords: reply relations, computer-mediated communication, CMC, corpus annotation, TEI

1. Introduction

In this paper, we examine the nature of various types of "reply", "addressing", or "reference" relations that exist between post units in computer-mediated communication (CMC) and which describe a reference from one given post to a previous post. We classify three types of reply relations in CMC interactions: *technical replies*, *indentations*, and *interpretative reply relations*. Our goal is to sort out the different levels of description and annotation that are involved, and to propose a solution for their combined representation within the TEI annotation framework.

The paper is structured as follows: The following section gives an overview of the reply relations as well as the existing practices in the description and annotation of such relations in chat, wiki talk, and blog corpora. In Section 3 we describe default and overriding reply relations. Section 4 presents our proposal for a CMC annotation scheme, which provides strategies for annotating the reply relations. In Section 5 we discuss perspectives for future work.

2. Reply Relations in written CMC

a. *Technical reply*

The most obvious and unambiguous type of reply relation can be observed in CMC genres where the client software, which is used to send a message (post) to the CMC platform, offers the possibility to reply directly to a previous post by clicking a button that is associated with the post and labelled 'reply'.

It can be activated to start the process of composing and eventually sending the reply, and it represents the standard reply action available in CMC genres such as email, Usenet news, Youtube, or blog comments. Generally, the reply relation (which message replies to which) will also be documented in the message metadata, for example the "References" field in the NNTP header (Schröck & Lüngen, 2015). We dub this type of reply relation a *technical reply*. Technical reply relations frequently form reply 'chains', and since several replies can be directed to the same previous message, the characteristic thread structures of such interactions arise. A post that is sent to the server without invoking a technical reply simply starts a new thread. CMC clients frequently display threads as indented list structures based on the reply ("references") information in the message protocol, (e.g. in the email client Thunderbird or in web browsers) via an HTML representation using nested lists or divisions (Fig. 1).

b. *Indentation*

A second type of reply relations is represented by the indentation structures found on wiki talk pages. Talk (or discussion) pages serve as a platform where wiki authors coordinate their work and share ideas about edits and improvements to the associated wiki article. From a technical point of view, talk pages are ordinary wiki pages, just like the articles. Traditional wiki software does not offer message or comment posting using a technical reply action as sketched under a., but instead, users are instructed to insert their contributions in the existing talk page and to indent and sign them properly using the wiki markup language.[1]

The sending action then always involves the sending of the whole, updated wiki page to the server. Clearly, the indentation policy serves to imitate a threaded reply structure as known from the layout of CMC with technical reply, and as a result, the collaborative dialogues look like discussion threads on the web page. With respect to reply relations, a talk contribution (likewise called a 'post') is by default assumed to indicate a reply to the post that is one level higher in the indentation hierarchy (Laniado et al., 2011; Margaretha

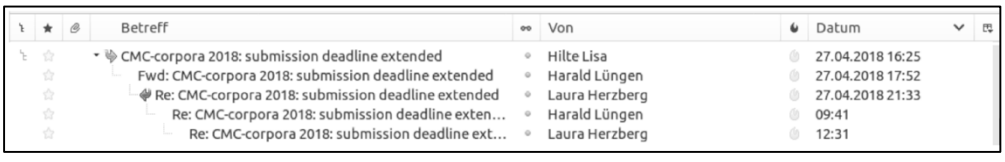

Figure 1: Display of technical reply relations in an email.

[1] https://en.wikipedia.org/wiki/Help:Talk_pages#Indentation.

& Lüngen, 2014; Poudat et al., 2014, Ho-Dac et al., 2016).

c. *Linguistic markers of address*
Besides technical replies and indentations, we observe that relations between posts in CMC, which researchers have identified as referencing or replying (e.g. Holmer, 2008), can also be signalled by other structural or linguistics means. A good example of CMC with such alternative signalling is chat, since neither in the composition of chat messages nor in the display of a chat log are technical replies or indentation structures applied. Hence, in chat, other indicators of the users' replying or addressing intentions are used, such as:

- a user name in combination with the address marker '@' as in *@James* (default reading: this post is a reply to the most recent post by James)
- a name in combination with a greeting (*Hi Harry*, *Hello Laura*)
- simply a name (*Laura*)
- citation: explicitly quoting a piece of a previous post (common in forum or email communication, often supported by the respective client software) (Schröck & Lüngen, 2015; Grumt Suárez et al., 2016)
- Q-A structures: giving an answer to a question raised in a previous post.

Reference	Order #	User	Message
	27	Maira	test farbe gewechselt *test colour changed*
	28	Mausi	akela: ihr franken seid ja eh so ein völkchen *grins* nicht bös gemein, mein freund ist ja auch einer *akela: you Franconians are a race apart *smile* mean no harm, my boyfriend is one as well*
	29	Clara	@mausi *gg* *@mausi *gg**
27	30	akela	so schön bunt hier *so colourful here*
28	31	akela	looool @mausi mein mann ist niederbayer es ist immer wieder zu schön *looool @mausi my husband is Lower Bavarian it's too good always*
30	32	Philina	ich versuch es mal in rot *I'm trying red*
31	33	Clara	das wird ja richtig multikulturell... badner, franken, bayern.. *this is becoming really multicultural... people from Baden, Franconia, Bavaria..*

Example 1: Excerpt of the chat *2223001_Nagetier-Chat_18-03-2003*[2].

[2] Chat names have been pseudonymised. All examples are in German. For understanding purposes, an English translation is added in italics.

In Example 1, the column *Reference* displays the number of the referred post, i.e. post (30) refers to post (27). The column *Order #* describes the position of the post within the thread. This example of the Dortmund Chat Corpus shows an unmoderated leisure chat with several users. The participants use different markers to signal reply relations, for example the address marker @, as in (29) and (30) or stating the username they are referring to at the beginning of their post, as in (28).
There are more potential indicators, but these tend to get more implicit and ambiguous, e.g., when taking a closer look at Example 1, we can infer, because of the topic continuation, that (30) refers to (27), and similarly (32) to (30). Also, a use of the pronoun *Du* (*you*) (not in the example) would signal a direct address to another user and consequently to one of her previous posts, but based on its form alone it cannot be decided who the addressee actually is. Humans no doubt often infer reply relations by understanding and interpreting that the content of a message forms a reply, to some extent, to the content of a specific previous message, even without overt indicators.

3. Default and overriding reply relations

Obviously, linguistic reply markers are also used in CMC genres that already offer a formal reply strategy (technical reply or indentation). In these cases, a linguistic reply marker may a.) indicate the same reply relation already marked by the formal reply and hence enforce it, b.) introduce an additional reply relation from the same post, or c.) introduce a reply relation that *overrides* the one originally introduced by the (apparently erroneously applied) formal reply strategy. Especially in wiki talk, the latter case frequently occurs. Even though there are generally accepted conventions on how to reply to previous postings, not all users follow them and, for example, stick to the given level of indentation (Laniado et al., 2011). In wiki discussions, the indentation level does not always correspond to the interpretative addressing cues within a post, as illustrated in Example 2. Normally, a non-indented post can be interpreted as another initial post which relates to the overall topic that is stated in the thread title (Laniado et al., 2011). Analysing linguistic cues of reference shows that this concept cannot be taken for granted in all wiki discussion threads, as in Example 2.

Fehler bei Sprachgruppen
Error in language groups
73,80% sind deutscher Muttersprache!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! (*nicht signierter Beitrag von* 71.231.556.63 (Diskussion) 17:32, 24. Aug. 2015 (CEST))
73,80% are German native speakers!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! —Preceding unsigned comment added 71.231.556.63 (talk) 17:32, 24. Aug. 2015 (CEST)
Nein, das stimmt schon so: http://www.gemeinde.bozen.it/servizi_context02.jsp?area=154&ID_LINK=3980 Weston (Diskussion) 09:52, 5. Sep. 2016 (CEST)

No, it's correct as it is: http://www.gemeinde.bozen.it/servizi_context02.jsp?area=154&ID_LINK=3980 Weston (talk) 09:52, 5. Sep. 2016 (CEST)

> Hi IP, wenn du mal wieder vorbeikommst: warst du 100 Jahre im Eis oder hast du Bozen mit Südtirol verwechselt?--Ophorlan ⁉ 15:00, 7. Nov. 2016 (CET)
>
> *Hi IP, just in case you stop by: did you spend 100 years in the ice or did you confuse Bolzano with South Tyrol?--Ophorlan ⁉ 15:00, 7. Nov. 2016 (CET)*

Example 2: Excerpt of the Wikipedia talk page to the article *Bozen*[3].

The second post by Weston is not indented but still easily interpreted as a reaction to the previous post. The user reacts to the initial message posted by the IP address 71.231.556.63. As a direct reply it should be indented like the post by Ophorlan. Not only does the indentation indicate that their post is a response to the initial post, but the addressing term *Hi IP* does this as well. This example shows that there are indeed differences between the formal indentation and the linguistic cues given in the post.

Apart from the user, the technical nature of a certain platform, for example, a blog platform, can set limitations. Even though it is possible that an initial comment beneath a blog post can trigger a large follow-up discussion, there might be a limit on (displayed) indentation levels so that users resort to linguistic markers of address to signal the reply status of a message (Grumt Suárez et al., 2016).

4. Annotation proposal

We propose that a CMC annotation scheme provides different annotation strategies for annotating a.) the technical reply references as sketched under section 2.a above and documented in the protocols of email, Usenet, or blog comments, b.) the indentation structure as represented in wiki text markup or HTML as known from wiki talk (section 2.b), and c.) the more interpretative reply structures induced by linguistic markers as sketched under 2.c. We also propose a separate annotation layer to represent the interpretative, final reply structure at a more abstract level, which would combine reply relations of all three kinds.

Our proposal adheres to the TEI Special Interest Group (SIG) on CMC, in which solutions for representing CMC corpus documents within the TEI framework have been developed either by defining good practices for using elements from the regular TEI, or by customising CMC-specific new elements and attributes (Beißwenger et al., 2016). We use the @replyTo and @indentLevel attributes as customised by the TEI CMC SIG for the <post> element, as well as grouped <link> elements from the regular TEI.

Remember that a reply relation instance always occurs between a post and a previous post within one CMC interaction. We propose to encode technical reply relations using the attribute @replyTo at the <post> element as customised in the CLARIN-D TEI schema for CMC (Listing 1).

```
<post synch="#t046" who="#u012_waschke"
xml:id="p007" replyTo="#p004">
```

Listing 1: Attributes of a post from a blog comment thread.

We propose to use the attribute @indentLevel at the <post> element as customised in the CLARIN-D TEI schema to represent all indentation structures in wiki talk (Beißwenger et al., 2016), regardless of whether they are to be interpreted as reply relations or not (Listing 3).

Finally, we newly propose to encode and collect all *interpreted* reply relations (whether based on technical reply, indentation, or linguistic markers) in the TEI header of the CMC document as a set of <link> elements gathered within a <linkGrp>. According to the TEI Guidelines, a <link> element quite generally "define[s] an association or hypertextual link among elements or passages"[4]. A <link> implies a set of targets in its @target attribute, i.e. pointers to those elements in the text that are to be linked (always a pair of post IDs in our application) (Listing 2).

```
<correspDesc type="replyRelations">
  <linkGrp>
    <link target="#p001 #h001" type="initial"/>
    <link target="#p002 #p001" type="implied"/>
    <link target="#p003 #p001" type="addressing"/>
  </linkGrp>
</correspDesc>
```

Listing 2: Interpreted reply relations in the TEI header.

```
<div type="thread">
  <head xml:id="h001">Fehler bei
  Sprachgruppen</head>
  <post xml:id="p001" who="#u001" synch="#t001"
  indentLevel="0" decls="#cd001">
     <p>73,80% sind deutscher
     Muttersprache!!!!!!!!!!!!!!!!!!!!!!!!!!!!
[...]</p></post>
  <post xml:id="p002" who="#u002" synch="#t002"
  indentLevel="0">
     <p>Nein, das stimmt schon so [...]</p></post>
  <post xml:id="p003" who="#u003" synch="#t003"
  indentLevel="1">
     <p><ref type="addressingTerm"
     target="#u001">Hi IP</ref>, wenn du mal
[...]</p></post>
</div>
```

Listing 3: Part of the TEI document body for Example 2.

We argue that the right place for the links is a link group in the TEI header of the CMC document because firstly, it is nice to have all of them collected in one place, so that they can be easily evaluated. Secondly, we can also *type* the abstract reply links using regular TEI means, i.e. the

[3] https://de.wikipedia.org/wiki/Diskussion:Bozen.

[4] http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-link.html.

@type attribute at the <link> element, such as to capture information about the source or reason of an interpreted reply relation; according to our examples, we suggest the possible values “technical”, “indentation”, “addressing”, “QA-relation”, “quoting”, and “implied” for the time being. Thirdly, the encoding via <link> references offers the possibility to encode multiple reply relations originating from one post if desired (e.g. Grumt Suárez et al., 2016), thus it has the potential to go beyond the proper tree structure of threads. Lastly, <link> references can even be applied to represent reply relations that occur between other parts of the CMC documents than posts, such as paragraphs within posts.

We propose that the <linkGrp> could go in the <corrispDesc> element of the file description of the TEI header, a section originally introduced to include information about the addressing, sending and receiving actions concerning one epistolary document; though we are open to alternative suggestions for its placing.

5. Conclusion and Prospects

Our next aim is to implement a routine that will automatically derive interpretative reply relations in Wikipedia talk pages. We think that this is a prerequisite for annotations for higher levels of interaction analysis such as dialogue acts (Ferschke et al., 2012), or discussion trees (Laniado et al., 2011).

6. References

Beißwenger, M., Ehrhardt, E., Herold, A., Lüngen, H., Storrer, A. (2016). (Best) Practices for Annotating and Representing CMC and Social Media Corpora in CLARIN-D. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, pp. 7--11.

Ferschke, O., Gurevych, I., Chebotar, Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 777--786.

Grumt Suárez, H., Karlova-Bourbonus, N., Lobin, H. (2016). A Discourse-structured Blog Corpus for German: Challenges of Compilation and Annotation. In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pp. 1--5.

Ho-Dac, L.-M., Laippala, V., Poudat, C., Tanguy, L. (2016). French Wikipedia Talk Pages: Profiling and Conflict Detection. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, pp. 34--38.

Holmer, T. (2008). Discourse structure analysis of chat communication. *Language@Internet* 5, article 9.

Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A. (2011). When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. *ICWSM-11*, The AAAI Press, pp. 177--184.

Margaretha, E., Lüngen, H. (2014). Building Linguistic Corpora from Wikipedia Articles and Discussions. In M. Beißwenger, N. Oostdijk, A. Storrer & H. van den Heuvel (Eds.), *Building and Annotating Corpora of Computer-mediated Communication: Issues and Challenges at the Interface between Computational and Corpus Linguistics*. pp. 59--82.

Poudat, C., Jin, K., Chanier, T. (2014). Manuel du corpus wikiconflits (cmr-wikiconflits-tei-v1- manuel.pdf). In *XX., Corpus Wikiconflits extraits de Wikipedia*. Dans banque de corpus CoMeRe.org. Ortolang.fr: Nancy. [cmr-wikiconflits-tei-v1].

Schröck, J., Lüngen, H. (2015). Building and Annotating a Corpus of German-Language Newsgroups. In M. Beißwenger & T. Zesch (Eds.), *Processing for Computer-Mediated Communication / Social Media*, pp. 17--2

What's Up, Switzerland?
Challenges of Twofold Non-Canonical Texts for Normalization

Simone Ueberwasser[1], Anne Göhring[1,2], Massimo Lusetti[1], Tanja Samardžić[3], Elisabeth Stark[1]

[1]Institute of Romance Studies, University of Zurich, Switzerland
[2]Institute of Computational Linguistics, University of Zurich, Switzerland
[3]URPP Language and Space, University of Zurich, Switzerland
simone.ueberwasser@ds.uzh.ch

Abstract

In 2014, Elisabeth Stark and her team from the project "What's up, Switzerland?" (www.whatsup-switzerland.ch) collected more than 600 WhatsApp chats (around 5 million tokens) in all four national languages from the Swiss population. Since 2016, six linguistic doctoral and postdoctoral students use the data as a basis for their theses while at the same time the computational linguistics team is enriching the data by means of different annotations. For more information about the project cf. Ueberwasser and Stark (2017).

The presentation will cover three topics. After a short introduction to the project as a whole, we will focus on how we processed the multilingual data, characterized by non-canonical spelling and dialectal language, to identify languages and varieties, how we cleaned up the data in terms of duplicated records, anonymization etc. and how we applied part-of-speech (PoS) tagging to some of the languages. The third topic will be a presentation of an innovative approach towards normalization of data written in Swiss German dialects by means of a recurrent neural network architecture consisting of an encoder-decoder model enhanced with a language model trained on different units (characters and words).

Keywords: CMC, WhatsApp, dialect, Switzerland, German, multilingual data, computational linguistics, text normalization

1. The Corpus

In 2014, a group of researchers invited the Swiss population via calls published in journals, radio shows etc., to send in their original WhatsApp chats. The goal was to study language use in this written yet dialogical communication form.

The corpus *What's up, Switzerland?* was collected in 2014 from e-mail attachments of WhatsApp protocols sent by the population. Next to sending the text messages, many informants also filled in an anonymous questionnaire and thus provided demographic information about themselves. Since the chats were always sent by one person but also contained texts written by other people (i.e. the communication partner(s)), data privacy was a major issue in the processing of the data. Other processing steps included language identification, normalization and the addition of annotations, as will be shown below.

2. Data Annotation

In the last three years, we performed the following manual and automatic processing steps on (parts of) the data:

- Mask the non-consented messages
- Detect partially duplicated chats
- Refine the tokenization
- Perfect the anonymization
- Manually (because of the non-standard nature of the data and the frequent code-switching) identify the language(s) in every chat
- Automatically identify the most probable language of every content message
- Manually identify the dialect(s) of randomly sorted messages from chats mainly written in Romansh
- Map the emoji symbols (Unicode code points) to a description string beginning with *emojiQ*[1]
- (Re)map emojis from the Private Use Areas – of potentially different providers – to the official Unicode code point and *emojiQ* description thereof[2]
- Convert the SQL data to Paula XML and the latter with Pepper to Annis format (Zipser and Romary, 2010; Krause and Zeldes, 2016)
- Manually normalize some messages in all four official languages of Switzerland as a source for further data processing
- PoS tag and lemmatize some chats written in Swiss German dialect(s) after manual normalization
- Automatically normalize and PoS tag the French messages
- Merge the manually and automatically annotated data

[1]We provide both emoji symbols and descriptions to our users for representation and query purposes; for example, users can search for all emoji descriptions containing the word *smiling* or *cat*.

[2]Some older WhatsApp messages used provider-specific encodings for the newest emojis that were not yet part of the standard Unicode or of their system's implementation.

In the following sections, we will motivate some of these steps and show how the manual and automatic processing mutually benefit from each other.

2.1. Privacy

The issue of privacy is one of our primary concerns when working with user-generated content. Therefore, the first step in processing the collected chats was to mask the messages of participants that had not given their consent. Another crucial point was to anonymize the data contained in the consented messages. Not only did we rotate the first names and masked the surnames of private persons appearing in the chats, but we also masked sequences of three and more digits that typically represent phone and card numbers. In addition, we also covered the addresses, e-mails, other communication and game accounts credentials we found as potential hints about a person's identity. This anonymization process went hand in hand with the incremental ad-hoc tokenization refinements, since both steps were executed automatically based on experience with previous data from Switzerland and then manually checked by student helpers.

2.2. Language and Dialect Identification

Switzerland is a multilingual country in the sense that different languages are spoken in different areas with German, French and Italian being official national languages that allow to perform all everyday tasks, whereas speakers of the fourth national language, Romansh "are obliged to function in German for many everyday purposes" (Barbour, 2004, 293). The situation for the German speaking community is characterized by a phenomenon known as diglossia (Ferguson, 1996). While German speakers in Switzerland normally use Standard Swiss German in formal situations and in written text, they almost always use their dialect in informal situations, definitely in spoken language but mostly also in written informal texts such as WhatsApp messages.
In order to identify the languages in the corpus, we first had student helpers manually identify the language(s) for every chat in order to get an overview and a quick starting point for the doctoral and postdoctoral students. In a second step, we used a normalized random-walk-based clustering system to identify the language of each message satisfying the following conditions: the approach builds a co-occurrence graph where each node is a lowercased token; the edges between nodes are the number of messages in which both tokens appear, normalized by the token frequency in the corpus. The tokens must have between 2 and 35 characters, at least half of which are alphabetical; additionally emojis, URLs and anonymization placeholders (e.g. *[LastName]*) are ignored. The only resource this system needs is a short list of words for each language to choose from; these words must 1) appear in the corpus, and 2) pertain unequivocally to one language. In our case, we had around 15 words for German, English, French, Swiss German, Italian, Romansh, and Spanish.
Table 1 shows, for each Swiss national language, the number of chats, messages with permission, and the tokens contained in them. Note that the volume of data for Romansh and Italian is way below what is available for other languages. This allowed us to have the Romansh data manually annotated for the language varieties by a native speaker (cf. Table 2).

Language	Chats	Messages	Tokens
Swiss German	275	506,984	3,611,033
French	141	197,255	1,397,375
Italian	87	42,559	293,567
Romansh	77	29,094	283,909

Table 1: Number of chats, messages and tokens per language.

Language	Code	Messages
Romansh	roh	2,157
Jauer	roh_ja	54
Putèr	roh_pt	206
Grishun	roh_rg	18
Surmiran	roh_sm	8,031
Sursilvan	roh_sr	4,263
Sutsilvan	roh_st	8
Vallader	roh_vl	1,893

Table 2: Number of manually identified Romansh messages per dialectal variety.

2.3. Manual Normalization

Text normalization aims at converting non-canonical text to a standardized form that is more suitable for further natural language processing (NLP) tasks. Once the main languages of the chats were identified, we started to manually normalize messages from chats 1) written in at least one of the four national Swiss languages, and 2) having demographic information from all participants. In a second phase, we relaxed the second condition but still restricted the selection to chats with permission from all participants.

Language	Messages	Tokens
Swiss German	6,441	54,371
French	6,999	51,234
Italian	4,399	40,625
Romansh	7,117	77,818
Total	24,956	224,048

Table 3: Number of manually normalized messages and tokens per language.

Several linguistics students and researchers normalized these 25,000 messages for a total of approximately 225,000 tokens with an adapted version of the online glossing tool described in Ruef and Ueberwasser (2013). In the main lines, we followed the normalization guidelines defined in the previous sms4science project (Stark et al.,); these guidelines[3] were further developed to cover new features, e.g. linguistic phenomena not yet encountered in older sms

[3] https://sms.linguistik.uzh.ch/SMS4science/Normalization

database or specific to WhatsApp messages. After this first phase of guideline adaptation and annotator training, we reached satisfying levels of inter-annotator agreements on small subsets (see Table 4). The remaining messages were normalized by one annotator only. We used the Swiss German part of this parallel data set (original ↔ normalized) to train an automatic normalization system (see Section 3. below).

Language	Messages	Tokens	Annot.	IAA
Swiss German	104	1,007	3	0.954
French	400	2,903	3	0.944
Italian	400	2,146	2	0.948
Romansh	100	512	2	0.892

Table 4: Inter-annotator agreement scores for normalization: Fleiss' Kappa when more than 2 annotators, otherwise Cohen's Kappa.

2.4. Automatic PoS Tagging

For the part-of-speech and lemma annotation of Swiss German messages, we ran the TreeTagger[4] (Schmid, 1994) with the Standard German model on the manually normalized forms to provide the doctoral students with data on the morpho-syntactic level.

Regarding the messages that had before been automatically identified as French, we used the MElt[5] sequence labeler (Denis and Sagot, 2012) together with its normalization wrapper to automatically normalize, tag and lemmatize them. We evaluated the resulting MElt part-of-speech tags a posteriori by manually checking a sample of 1,314 tokens from 160 randomly selected messages, correcting the tags as well as the normalized forms. We observed an accuracy of 95.74% for normalization and 84.93% for PoS tagging. The low PoS tagging accuracy is partly due to the peculiarities of WhatsApp messages (e.g., abbreviations, slang, typos, character reduplication and other irregularities), but also to the errors inherited from the normalization step.

Note that the normalization of the French messages achieves a substantially higher accuracy than the normalization of Swiss German messages (see Section 3.), because the input texts are closer to standard language than the German texts. While the German texts deviate from the standard because of it being CMC data and because of the dialectal situation in Switzerland, the French texts only deviate as CMC data. This allowed us to use an existing pipeline without adjustments and still obtain acceptable results in French.

Future work includes part-of-speech annotation for Italian based on Cimino and Dell'Orletta (2016).

3. Automatic Normalization of Swiss German Dialects

In section 2.4. we mention how we have normalized French messages using the pre-trained off-the-shelf tool MElt. In this section we describe how we have used a manually normalized subset of the corpus to train models for automatic text normalization of WhatsApp messages written in the Swiss German dialects. It is known that the more variation we have in a text, the harder it is for NLP tools and systems to interpret and process human language. Several factors contribute to the high degree of variation that characterizes user-generated content written in Swiss German. In addition to the irregularities that are typical of WhatsApp messages, we observe a high degree of regional, inter-speaker and intra-speaker variation, which is due to the lack of a standardized orthography and to the fact that Swiss German is not one single dialect, but rather a family of dialects. For example, the Swiss German words *viel, viil, vill* and *viu* map to the single normalized form *viel*.

We considered automatic text normalization as a machine translation task, where the source language is Swiss German and the target language is its normalized form. We carry out this process both with phrase-based statistical machine translation (Aw et al., 2006) (Section 3.1.) and neural encoder-decoder methods (Section 3.2.). We use character-level approaches to normalize words in isolation rather than text segments or complete sentences. Both approaches require a parallel corpus of manually normalized texts. A subset of the corpus *What's up, Switzerland?* (the WUS corpus hereafter) has been used to train and evaluate models that can learn from the data how to automatically normalize previously unseen texts. This subset consists of 54,229 alignment units, whereby most units are one-to-one alignments of a source and a target word token. One-to-many alignments are typically merged constructions of article and preposition, and verb forms merged with an enclitic subject pronoun (*hani* → *habe ich* 'have I'), an enclitic object pronoun (*heschen* → *hast ihn* 'have [you] [verb] him'), or with both (*hämmers* → *haben wir es* 'have we [verb] it'). The rare cases of many-to-one alignments are due to user's typos and arbitrarily split verb prefixes and compound nouns.

3.1. Character-level Statistical Machine Translation

Text normalization is mostly performed with character-level statistical machine translation (CSMT), which has been applied to Dutch user-generated content (De Clercq et al., 2013), Slovene tweets (Ljubešić et al., 2014) and Swiss German dialects (Samardžić et al., 2015; Scherrer and Ljubešić, 2016).

A fundamental prerequisite of statistical machine translation is a parallel corpus of aligned sentences, in which word alignment is performed by applying IBM models with the Expectation-Maximization (EM) algorithm (Brown et al., 1993). We carry out SMT by employing the Moses toolkit[6] (Koehn et al., 2007), which combines a translation model – responsible for the adequacy of the translation from source to target sequence – and a language model (LM) – responsible for the fluency of a sequence in the target language – to find the best translation of a source sequence. To achieve better context-sensitive source-target mappings, Moses relies on phrase-level translation models (Koehn et al., 2003).

[4] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[5] https://team.inria.fr/almanach/melt/

[6] http://www.statmt.org/moses/

These models allow to build a phrase table to store aligned phrase pairs, in the source and target languages, that are not necessarily linguistically motivated but are consistent with the alignments of single words established by the IBM models.
In CSMT, we simply replace words with characters as the symbols that constitute a phrase. CSMT is a suitable approach for those tasks in which many word pairs in the source and target languages are formally similar, such as the Swiss German word *Sunne* normalized as *Sonne* ('sun'), or are characterized by regular transformation patterns that are not captured by word-level systems, such as the pattern *ii* → *ei*, which produces the transformations *Ziit* → *Zeit* ('time') and *Priis* → *Preis* ('price'). One further advantage of CSMT is that it can be highly effective when little training data is available. This is because once a transformation pattern has been learned, it can be applied to translate unknown words that would be out of vocabulary (OOV) for a word-level SMT system.

3.2. Neural Encoder-Decoder Models

Recurrent neural networks (RNN) have been recently proposed as a new approach to machine translation. We use a neural architecture with long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) and a soft attention mechanism (Bahdanau et al., 2014). The encoder-decoder (ED) model consists of two RNNs: an encoder converts the input character sequence into a sequence of vectors that are then selected adaptively by a decoder to produce an output sequence. Each time a target character is predicted by the decoder, the soft attention mechanism focuses on the most relevant part of the input.
The main innovation of our approach consists in an additional word-level language model which is integrated in the character-level ED using a synchronization mechanism. This method is inspired by Ruzsics and Samardžić (2017), who addressed the task of morphological segmentation by integrating the basic encoder-decoder component with a language model trained on sequences of morphemes.

3.3. Results and Analysis

In both the CSMT and ED approaches, we use a corpus split of 80% training, 10% tuning and 10% test set. We use accuracy as evaluation metric, which measures the percentage of correctly normalized units compared with the manual normalization (reference). The baseline has been established following Samardžić et al. (2015): for each word in the test set, the most frequent normalization in the training set is chosen. In case of tie, the normalized form is chosen randomly, and words that have not been seen in the training set are simply copied. This method produces an accuracy score of 84.45%, which represents a very strong threshold.

Baseline	CSMT	ED
84.45%	86.35%	87.61%

Table 5: Automatic normalization accuracy scores.

Moreover, we use the target side of a corpus of SMS (Stark et al.,) to train additional language models. We achieve an accuracy score of 86.35% with the CSMT approach, when using character-level LMs trained on both the WUS and the SMS corpus. With respect to the ED approach, an ensemble of 5 models produces an accuracy score of 87.61% when using word-level LMs trained on both corpora, in addition to the character-level WUS language model, which is built in the neural framework. Both approaches perform better than the strong baseline, and the ED model outperforms the CSMT model by taking advantage of the integrated word-level LMs. These produce, for example, improvements in the normalization of foreign words (e.g., source *cream*, where CSMT erroneously forces normalization and gives *kream*), one-to-many mappings in which single source words are normalized as two or more target words (e.g., *söuis* → *soll ich es* 'shall I [verb] it'), and source words whose reference normalization is formally very different (e.g., *wg* → *wohngemeinschaft* 'shared apartment').

4. Conclusion

Starting from a raw corpus of WhatsApp messages, we have processed them both manually and automatically to identify their language, and to normalize and annotate their tokens, creating a multi-level annotated corpus that is an excellent base for further linguistic research and experiments in natural language processing. At the end of the project, the deeper (more specific) annotation levels shown above will be integrated into our Annis corpus, ready thus to be browsed and queried by a broader audience.

5. Acknowledgements

This research is funded by the Swiss National Science Foundation, project "What's up, Switzerland? Language, Individuals and Ideologies in mobile messaging" (Sinergia: CRSII1_160714) and done in collaboration with the URPP "Language and Space", University of Zurich.

Bibliographical References

Aw, A., Zhang, M., Xiao, J., and Su, J. (2006). A Phrase-based Statistical Model for SMS Text Normalization. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Barbour, S. (2004). National language and official language. In Ulrich Ammon, et al., editors, *Soziolinguistik*, pages 288–295. de Gruyter, Berlin, New York.

Brown, P. E., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Cimino, A. and Dell'Orletta, F. (2016). Building the state-of-the-art in POS tagging of Italian Tweets. In *Proceedings CLiC-it 2016 and EVALITA 2016*, volume 1749.

De Clercq, O., Desmet, B., Schulz, S., Lefever, E., and Hoste, V. (2013). Normalization of Dutch user-generated content. In *Proceedings of RANLP 2013*, pages 179–188, Hissar, Bulgaria.

Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

Ferguson, C. A. (1996). Diglossia. In Thom Huebner, editor, *Sociolinguistic Perspectives. Papers on Language in Society, 1959 – 1994*, pages 25–39. Oxford University Press, Oxford.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (Volume 1)*, pages 48–54, Edmonton, Canada.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 demonstration session*, pages 177–180, Prague, Czech Republic.

Krause, T. and Zeldes, A. (2016). Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Ljubešić, N., Erjavec, T., and Fišer, D. (2014). Standardizing tweets with character-level machine translation. In *Proceedings of CICLing 2014*, Lecture notes in computer science, pages 164–175, Kathmandu, Nepal. Springer.

Ruef, B. and Ueberwasser, S. (2013). The taming of a dialect: Interlinear glossing of Swiss German text messages. In Marcos Zampieri et al., editors, *Non-standard Data Sources in Corpus-based Research*, pages 61–68, Aachen.

Ruzsics, T. and Samardžić, T. (2017). Neural sequence-to-sequence learning of internal word structure. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 184–194, Vancouver, Canada. Association for Computational Linguistics.

Samardžić, T., Scherrer, Y., and Glaser, E. (2015). Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of The 4th Biennial Workshop on Less-Resourced Languages*. ELRA.

Scherrer, Y. and Ljubešić, N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 248–255.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Stark, E., Ueberwasser, S., and Ruef, B.). Swiss SMS corpus, University of Zurich. https://sms.linguistik.uzh.ch.

Ueberwasser, S. and Stark, E. (2017). What's up, Switzerland? A corpus-based research project in multilingual Switzerland. *Linguistik Online*, 84/5:105–126.

Zipser, F. and Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*, La Valette, Malta, May.

Effects of Relationship Goals on Linguistic Behavior in Online Dating Profiles: A Classifier Approach

Tess van der Zanden, Chris van der Lee, Alexander Schouten, Maria Mos, Emiel Krahmer
Tilburg University
t.vdrzanden@uvt.nl, c.vdrlee@uvt.nl, a.p.schouten@uvt.nl, maria.mos@uvt.nl, e.j.krahmer@uvt.nl

Abstract

This study investigates whether online daters looking for a long-term relationship behave linguistically different in the dating profile texts they write from casual relationship seekers. To determine language use differences, we first analyze 12,310 authentic online dating profile texts using the Linguistic Inquiry and Word Count (LIWC) program on three linguistic categories: the occurrence of physical appearance and status-related words, positive emotion words and personal pronouns. In addition, we employ an exploratory language identification method to investigate how profile texts of long-term and casual relationship seekers can be discriminated based on a selection of linguistic features. Both the LIWC and the word-based classifier method indicate there are linguistic differences between the profile texts of the two groups of relationship seekers. Most notably, long-term relationship seekers are more likely to use words that emphasize internal characteristics that are more important when looking for a long-term relationship partner.

Keywords: online dating, relationship goals, corpus research, text analysis, LIWC

1. Introduction

Over the past two decades, online dating's popularity has steadily increased (Pew Research Center, 2015). People's motivations and goals to date online may differ: some may aim for a life-lasting relationship, where intended intimacy goals may eventually develop into long-term commitment (Sternberg, 1986). Others may seek casual, potentially sexual dates which may involve personal contact without the intention to become high-involved, intimate relationship partners. When people are more interested in a partner for the short-term, intimacy goals are less vital and factors such as physical attractiveness are more important. These intended relationship goals determine how particular characteristics in a potential date or partner are valued (Buss & Schmitt, 1993; Sternberg, 1986).

On many dating sites, users can explicitly indicate their desired relationship goal as a basic characteristic on the dating profile. A dater's relationship goal can also be deduced from both the presence as well as the content of a profile owner's picture (Gallant, Williams, Fisher, & Cox, 2011). Long-term relationship seekers are more likely to display a profile picture. Moreover, on their profile pictures, casual relationship seekers tend to wear less clothes. Both relationship seeking groups emphasize specific attributes and in this way convey information about their intentions: where short-term relationship seekers emphasize physical appearance and attractiveness, long-term relationship seekers seem to self-disclose more by posting pictures.

It is conceivable that relationship goals are also reflected in the linguistic behavior online daters employ in their profile texts. Although this topic has been virtually unexplored directly, previous studies have shown that stable characteristics of a profile owner, such as gender, age and personality, influence linguistic behavior (e.g., Davis & Fingerman, 2016; Groom & Pennebaker, 2005). Profile owners' gender, the most studied characteristic, structurally affects language use in dating profiles. For example, men talk more about object properties and jobs, whereas women tend to talk in a more personal way and express themselves more emotionally (e.g., Groom & Pennebaker, 2005).

It may well be that not only stable characteristics such as an online dater's gender, but also more dynamic aspects of that person, such as the motivations to date online, affect language use in dating profile texts.

In order to examine whether intentions influence linguistic behavior in online dating profiles, we use a large sample of authentic profile texts of both casual and long-term relationship seeking men and women. We rely on the Linguistic Inquiry and Word Count program (LIWC; Pennebaker, Booth, & Francis, 2007) to analyze the language use in these profile texts, which allows us to assess the word use in various predefined and psychologically motivated categories. In addition, we use a classification approach based on Van der Lee and Van den Bosch (2017) to investigate if it is possible to discriminate between texts of casual and long-term relationship seekers based on word *n*-grams, and to automatically extract those content-specific features from a text that are most distinctive for these two categories.

1.1 Hypotheses

Following the assumption that casual relationship seekers pay more attention to external characteristics that are important for lower involved relationships, such as physical appearance (e.g., 'good-looking', 'fit'), and long-term relationship seekers focus more on internal characteristics, such as status (e.g., 'income', 'director'), it is expected that based on their own relationship goal, profile owners emphasize such characteristics in their dating profiles.

H1. Online daters looking for a casual relationship use more words related to physical appearance than long-term relationship seekers.
H2. Online daters looking for a long-term relationship use more words related to status than casual relationship seekers.

Compared to casual relationship seekers, long-term relationship seekers are aiming more for emotional involvement, commitment and the formation of intimate

and close relationships (Gibbs, Ellison, & Heino, 2006). Therefore, we expect long-term relationship seekers to use more words related to positive emotions, including words that express emotional closeness (e.g., 'love', 'loyal') and emphasize positive characteristics (e.g., 'intelligent', 'self-confident').

H3. Online daters looking for a long-term relationship use more positive emotion words than casual relationship seekers.

Differences in the use of automatically produced personal pronouns can reveal a great deal about an individual's attitudes, goals and roles within relationships. How often they are used can give away psychological and mental states and processes of profile owners (Pennebaker, 2011). The expectation is that online daters who seek casual partners, focus more on the self and consequently use more I-references (e.g., 'I', 'me'). Furthermore, we expect long-term relationship seekers to engage more in affiliative behavior and attempt to connect with others than casual relationship seekers by using more you- (e.g., 'you', 'your') and we-references (e.g., 'we', 'our').

H4. Online daters looking for a casual relationship use more I-references than long-term relationship seekers.
H5. Online daters looking for a long-term relationship use more you-references than casual relationship seekers.
H6. Online daters looking for a long-term relationship use more we-references than casual relationship seekers.

2. Method

2.1 LIWC

The sample included 12,310 profile texts from a popular Dutch dating website. Together with the profile text, the profile owner's self-indicated gender, age, education level and relationship goal was extracted. The Ethics Committee (ETC) of the *Tilburg School of Humanities and Digital Sciences* authorized us to collect and analyze the profile texts as long as the profile owners' anonymity was guaranteed and no identifiable information was provided.

The total sample consisted of profiles from long-term and casual relationship seekers by heterosexual men and women which were written in Dutch by profile owners who indicated to live in the Netherlands. Long-term relationship seekers were site users who selected the option 'long-term relationship' as the preferred relationship goal, and casual relationship seekers were those who indicated to look for a 'date'. In total, there were 10,696 profile texts of long-term relationship seekers and 1,614 of casual relationship seekers. From the total sample of profiles, 64.2% was written by site members who indicated themselves men, and the mean age of the profile owners was 42 years and 8 months (SD = 11.7). Profile texts were written by profile owners with either a low (42.3%) or a high level of education (57.7%).

LIWC is a program that calculates proportions of specific categories of words within text files. For the textual analysis of the Dutch dating profiles, the Dutch LIWC dictionary vocabulary was used (Zijlstra et al., 2004). This inventory contains an internal dictionary of around 7,000 words and each word is categorized into one or more linguistic categories.

By default, the Dutch LIWC analyzes each text file on 70 established linguistic categories. For our analyses, we used three sets of categories. The first set looked into the use of words related to physical appearance and status. To do so, the predefined LIWC categories Occupation, School, Achievements, Money and Job were combined into one measure of Status-Related words. The LIWC categories Physical Functioning, Body Parts and Sexuality were grouped under the umbrella category Physical Appearance. The second set measured the use of positive emotion words, while the third set looked into the use of personal pronouns, calculating the number of I-, you- and we-references in the profile texts.

The existing Dutch LIWC dictionary was slightly adjusted by the authors. Some inflectional variants and simple derivations of Dutch lemmas were added to the dictionary as inflected words in the Dutch version are not reduced to their word stem and assigned into the same category (Zijlstra et al., 2004). For example, the Dutch adverb/adjective form *aardig* ('kind') occurred in the vocabulary, but the inflected form *aardige* ('kind') which occurs in attributive use, did not. Furthermore, we added 14 Dutch translations of a supplementary English list of words composed by Davis and Fingerman (2016) related to physical attractiveness.

For the analyses with LIWC, only the first 100 words of all 12,310 profiles were analyzed, as this is also what other site users initially see when scrolling through profiles. Moreover, we only included profile texts in the sample with a word count of fifty or higher.[1] A multivariate analysis of variance was conducted with relationship goal as independent variable and the proportion of words matching with the words in the previously described linguistic categories as the six outcome variables (α = .05).

2.2 Word-Based Classifier

The additional classifier analysis is based on the approach of Van der Lee and Van den Bosch (2017). Six different machine learning methods are used: linear SVM (support vector machine), Naive Bayes, and four variants of tree-based algorithms (decision tree, random forest, AdaBoost en XGBoost). In contrast with LIWC, this classifier does not deal with a predefined list of words but uses aspects from the profile texts as direct input and extracts word *n*-grams that are distinctive for either of the text groups.

For the word-based classifier, 1,614 profile texts of

[1] Profile texts of long-term relationship seekers (M = 81.0, SD = 12.9) were significantly longer than those of casual relationship seekers (M = 79.2, SD = 13.5), $F(1,12309) = 26.8$, $p < .001$, $\eta p^2 = .002$. This does not influence our results since LIWC operates with proportion scores.

both relationship groups were used. This means that for casual relationship seekers, the entire subset was used, whereas for the long-term relationship seekers a smaller subset of the group of 10,696 texts was randomly selected.

For the word-based classifier a ten-fold cross validation method was used, that was run ten times using ten different seeds. To control for text length effects, the classifier used ratio scores as features rather than absolute values. In addition, most of the stop words from the regular NLTK stop words list were not considered as content-specific features, with the exception of personal pronouns that were selected as potentially interesting features. Notice that the classifier operates on lemma level, indicating that before running, words were converted to lemmas by means of Frog (Van den Bosch, Busser, Daelemans, & Cansius, 2007). Those content-specific features that will eventually be provided as important will also be distinctive lemmas.

3. Results

3.1 LIWC

Overall, LIWC recognized 72.2% of the words used in the profile texts (SD = 7.11).

3.1.1 Physical appearance and status-related words

Hypothesis 1 was not confirmed since casual (M = 0.32, SD = 0.73) and long-term relationship seekers (M = 0.37, SD = 0.79) did not differ in their use of words related to physical appearance, $F(1, 12310) = 3.69$, $p = .055$. Results on the use of status-related words were in line with hypothesis 2: long-term relationship seekers (M = 1.06, SD = 1.29) wrote more words related to status in their profile texts than casual relationship seekers (M = 0.97, SD = 1.24), $F(1, 12310) = 6.24$, $p = .012$, $\eta p^2 = .001$.

3.1.2 Positive emotion words

Hypothesis 3 was confirmed since long-term relationship seekers (M = 6.01, SD = 3.23) expressed more positive emotion words than casual relationship seekers (M = 5.65, SD = 3.12), $F(1, 12310) = 17.78$, $p < .001$, $\eta p^2 = .001$.

3.1.3 Personal pronouns

Contrary to hypothesis 4, not the casual relationship seekers (M = 6.83, SD = 3.36) but the long-term relationship seekers (M = 7.28, SD = 3.35) referred more often to the self, $F(1, 12310) = 25.51$, $p < .001$, $\eta p^2 = .002$. Hypotheses 5 and 6 concerned the effect of the type of relationship sought on the use of you- and we-references. The data showed that long-term and casual relationship seekers used equally often you-, $F(1, 12310) = 1.50$, $p = .221$, and we-references, $F(1, 12310) = 3.55$, $p = .059$.[2]

3.2 Word-Based Classifier

The XGBoost algorithm appeared to be the most effective one. For this algorithm, the accuracy score was 59.6%, improving accuracy above chance level with 9.6%.

The features presented in Table 1 are per relationship seeking group the ten most prominent distinctive content-specific features with their English translations, that had a total frequency score of hundred or more. These features are obtained based on the importance scores given by the best performing tree-based algorithm, in our case XGBoost. The importance scores are also listed in the table.

Content-specific features					
Long-term			Casual		
Dutch	**English**	**IS**	**Dutch**	**English**	**IS**
betrouw-baar	trustwort-hy	.032	date	date	.087
samen	together	.028	zin	feel like	.017
mijn profiel	my profile	.027	weten	to know	.016
rustig	calm	.026	vrouw	woman	.011
eerlijk	honest	.026	gek	crazy	.010
ik	I	.024	geen	no	.009
dag	day	.023	even	for a while	.009
serieus	serious	.023	eten	to eat	.005
mijn	my	.022	komen	to come	.005
genieten	to enjoy	.022	sturen	to send	.005

Table 1: Top 10 most distinctive content-specific features and the importance score (IS) per feature.

Some of the most important content-specific features indicate lexical differences between the two relationship seeking groups that point at interpretable patterns. Results seem to show that long-term relationship seekers tend to emphasize long-term attributions and traits: with words as 'trustworthy', 'together' and 'honest' in the top ten. In the top ten of words distinctive for casual relationship seekers, more 'casual' actions and attributes come across, such as the words 'date', 'crazy' and 'to eat'.

4. Discussion

The aim of this study was to investigate whether online daters who look for a long-term relationship behave linguistically different in the profile texts they write from online daters who seek a date, a casual less-involved relationship. Results of two types of textual analyses – LIWC and the word-based classifier – indicated there to be differences in language use between the two groups of relationship seekers. This seems to suggest that not only profile owners' stable characteristics (e.g., gender, age), but also dynamic characteristics, such as goals and intentions, can influence linguistic behavior in dating profiles.

In line with our expectations, long-term relationship seekers write more about status-related topics than casual relationship seekers, suggesting that promoting internal characteristics and long-term attributes, such as status, is considered to be more important for those looking for a partner for the long-term. The higher use of positive

[2] The interaction of relationship goal and gender was only significant for the average text length, with $p = .011$ and $\eta p^2 = .001$, and the use of I-references, with $p = .007$ and $\eta p^2 = .001$.

emotion words by long-term relationship seekers corresponds with Gibbs et al. (2006) who suggest long-term relationship seekers use more of these positive emotion words because this contributes to the pursuit of their intended high-involved, committed relationship.

The last linguistic category in LIWC we looked into was the use of personal pronouns by long-term and casual relationship seekers. In contrast with our hypothesis, the long-term relationship seekers referred more often to the self and not the casual relationship seekers. A possible explanation for this significantly higher use of I-references by long-term relationship seekers may be that referring to the self does not only indicate relationship autonomy but can also be interpreted as a sign of increased levels of self-disclosure. Self-disclosure promotes intimacy and closeness (Gibbs et al., 2006). Considering that long-term relationship seekers are more willing to involve in intimate and close relationships, it is perhaps not surprising that they self-disclose more personal information.

Complementary to our textual LIWC analyses, we used a word-based classifier method – with the profile texts as input – to automatically generate what discriminates language use of long-term and casual relationship seekers. Exploratory findings of the language classification machine learning method pointed at interpretable patterns for the two groups of profile texts. In line with earlier findings with LIWC, self-references such as *ik* 'I' and *mijn* 'my' were distinctive features in texts of long-term relationship seekers. In addition, where LIWC showed long-term relationship seekers emphasize long-term characteristics by using more words related to status, also the word-based classifier seems to show that long-term relationship seekers are more likely to mention internal and long-term attributions and traits: with words as 'trustworthy', 'honest' and 'together' in the top ten. Casual relationship seekers, on the other hand, were more distinctive in their use of words as 'date' and 'crazy'.

Results of this word-based classifier approach show some interesting preliminary findings that add to the earlier LIWC findings. The results are extracted on the basis of differences between the groups in this specific dataset. Running the same classifier on other datasets would be necessary to determine whether the observed differences in this dataset can be used as a classification predictor. Furthermore, these results form an interesting basis for future (hypothesis-testing) research, such as a scenario-based experiment in which participants have to imagine looking for either a long-term or a casual partner and write profile texts with that goal in mind.

Since the focus of this study was on isolated words and word combinations only, rather small effect sizes and a word-based classifier yielding an accuracy score of 59.6%, are not surprising. Not one relationship seeking group is restricted to the use of a particular linguistic element, but it is all a relative comparison in which daters of one group are more prolific in their use of words from a category or a particular content-specific feature compared to another group. In the future, it would be interesting to look further into linguistic differences at higher sentence or text levels, such as topics long-term and casual relationship seekers talk about and declarative and question sentence constructions.

Although both methods are classifiers operating on a word-count basis, there is, at least, one fundamental difference between their approaches. While the word-based classifier is entirely data-driven, LIWC works with a more top-down approach, in which predefined categories have been manually compiled and validated by psychologists, sociologists and linguists. Both methods come with pros and cons. It is with this reason that using both methods can work complementary, leading to a more complete picture of differences in language use of long-term and casual relationship seekers. To gain more insights into the predictive abilities of LIWC and to what degree LIWC and a word-based classifier accentuate different aspects of language, it would be interesting to use a comparable machine learning algorithm to classify this profile text dataset, with the words in the LIWC categories as part of the classification features.

5. References

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review, 100*(2), pp. 204-232.

Davis, E. M., & Fingerman, K. L. (2016). Digital dating: Online profile content of older and younger adults. *Journal of Gerontology Series B: Psychological Sciences and Social Sciences, 71*(6), pp. 959-967.

Gallant, S., Williams, L., Fisher, M., & Cox, A. (2011). Mating strategies and self-presentation in online personal advertisement photographs. *Journal of Social, Evolutionary, and Cultural Psychology, 5*(1), pp. 106-121.

Gibbs, J. L., Ellison, N. B., & Heino, R.D. (2006). Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in Internet dating. *Communication Research, 33*(2), pp. 152-177.

Groom, C. J., & Pennebaker, J. W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles, 52*(7), pp. 447-461.

Pennebaker, J. W., Booth, R. J. & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us.* New York: Bloomsbury Press.

Pew Research Center (2015, April 20). *5 facts about online dating.* Retrieved from https://www.pewresearch.org/fact-tank/2016/02/29/5-facts-about-online-dating/

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review*, *93*(2), pp. 119-135.

Van den Bosch, A., Busser, G.J., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99-114).

Van der Lee, C., & Van den Bosch, A. (2017). Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 190-199).

Zijlstra, H., Meerveld, T. van, Middendorp, H. van, Pennebaker, J. W., & Geenen, R. D. (2004). De Nederlandse versie van de 'linguistic inquiry and word count' (LIWC). *Gedrag Gezond, 32,* pp. 271-281.

Code-Mixing with English in Dutch Youths' Online Language: OMG SUPERNICE LOL!

Lieke Verheijen, Laura de Weger, Roeland van Hout
Radboud University (Nijmegen, the Netherlands)
lieke.verheijen@let.ru.nl; r.vanhout@let.ru.nl

Abstract

Youths in the Netherlands are hooked on their mobile phones. Their written computer-mediated communication (CMC) reflects their oral youth language, of which English elements have become a salient characteristic. The increasingly prominent role of English within Dutch society poses the question how this is reflected in Dutch youths' CMC. This paper presents a corpus analysis into Dutch youths' code-mixing with English in their written CMC. We analysed 8,619 English elements for language-internal (length, multiplicity, word category, integration, semantic field, intentionality, frequency) and language-external factors (CMC mode, age, gender). This study shows that English is a core feature of Dutch youths' online teen talk. It remains subordinate to Dutch, so English does not (yet) have the upper hand in Dutch youths' written CMC. Still, their use of full English sentences, discourse markers, and conversational words reveals a considerable knowledge of the English language, and the English elements signal youthfulness and dynamism.

Keywords: computer-mediated communication, social media, code-mixing, youth language, English, code-switching

1. Introduction

English borrowings are quite common in Dutch, especially in advertising, commercials, and business communication, but also in everyday speech (Zenner, Speelman, & Geeraerts, 2015). This shows the increasing status and impact of English within Dutch society and reflects the increasing number of Dutch speakers who have English as their second language. The relationship between English and Dutch has been qualified as a weak contact situation (Zenner et al., 2015), but the steady increase of bilingual primary and secondary education in the Netherlands in recent years (NOG, 2018) asks for a redefinition of the relationship between Dutch and English in the near future. Therefore, we need a better understanding of current multilingual practices among youths in the Netherlands. These practices include linguistic borrowing, but also code-mixing[1], i.e. the alternation between two or more languages within a single conversation or context. Code-mixing with English has become a salient aspect of Dutch youth language or 'teen talk' (Tagliamonte, 2016). Since Dutch youths often communicate online, this study focuses on Dutch-English code-mixing in their computer-mediated communication (CMC).

2. Research Goals

The present study aimed to explore Dutch youths' code-mixing in their written CMC, specifically to what extent, how, and why they code-mix with English.

3. Methodology

3.1 Materials

We quantitatively and qualitatively analysed a large corpus of Dutch written CMC. The corpus consisted of messages by youths between the ages of 12 and 23, of different genders (male, female) and age groups (adolescents: 12-17, young adults: 18-23). They were composed via four 'CMC modes' (SMS text messaging, Twitter, MSN chat, and WhatsApp) between 2009 and 2016. MSN chats, texts, and tweets were extracted from an existing reference corpus of written Dutch, called SoNaR ('STEVIN Nederlandstalig Referentiecorpus'; Treurniet & Sanders, 2012). WhatsApp chats were collected by requesting Dutch youths to voluntarily submit their messages via a website (https://cls.ru.nl/whatsapptaal/; Verheijen & Stoop, 2016). The composition of the corpus is presented in Table 1.

Genre	Years of collection	Age group	# words	# chats or contributors
MSN	2009-2010	12-17	45,051	106
		18-23	4,056	21
SMS	2011	12-17	1,009	7
		18-23	23,790	42
Twitter	2011	12-17	22,968	25
		18-23	99,296	83
WhatsApp	2015-2016	12-17	55,865	84 / 11
		18-23	140,134	132 / 23
Total	2009-2016	12-23	392,169	

chats: MSN, WhatsApp; # contributors: SMS, Twitter, WhatsApp

Table 1: CMC texts.

3.2 Method

3.2.1. Extraction of English Elements

Our first step was to extract all instances of code-mixing from the corpus. For this, we used the online version of *Van Dale's Great Dictionary of the Dutch Language*, a recognized authority among Dutch lexicons. Elements

[1] Throughout this paper, we use the term code-mixing to refer to elements from one language (English, L2) used in another language (Dutch, L1), when mainly the grammar of the L1 is at work. This contrasts with code-switching, in which the grammars of both the first and second language are active simultaneously. Alternatively, code-meshing and the related 'translanguaging' consider language alternation not as involving two separate grammars or language systems, but as one integrated system (Canagarajah, 2011).

were coded as English when their lemmas were (a) not included in this dictionary, but included in English dictionaries or (b) included in the Dutch *Van Dale* dictionary, but with an indication that they had recently been borrowed from English. Words were not considered instances of code-mixing when they were (a) included in the Dutch *Van Dale* dictionary without this indication or (b) proper names, such as titles of films or games. Determining whether a word or phrase was an English element was done entirely manually, so that inflected forms of English words (e.g. *chicks*, *checking*) and non-standard or 'dutchified' spelling variants of Standard English words (e.g. *nais* instead of *nice*) were also identified and coded as English. The corpus contained 8,619 English elements, which together made up roughly 10,000 words (since some elements consisted of multiple words) – 2.5% of the corpus.

3.2.2. Coding of English Elements

All instances of code-mixing were systematically and manually coded in Microsoft Access. Previous research (De Decker & Vandekerckhove, 2012; Verheijen, 2016) inspired us to examine the English elements for various language-internal factors, namely their length, multiplicity, word category, integration, semantic field, intentionality, and frequency, as well as for the language-external factors of CMC mode, age, and gender. Statistical analyses were afterwards conducted with IBM SPSS Statistics. In addition, we qualitatively analysed especially interesting cases of code-mixing that we encountered in the corpus.

4. Results

4.1 Language-Internal Factors

4.1.1. Length

The English elements were coded for five types of length: partial-word, single-word, textism, phrasal, and sentence. Table 2 presents the frequencies according to their length. By far the most frequent were single words, such as *nice*, *hey*, *shit*, *swag*, and *happy*. This supports the idea that Dutch youths' code-mixing consists of a matrix language Dutch, with English as the embedded language. Second most frequent were textisms, like *btw* (by the way), *idk* (I don't know), *jk* (just kidding), and *thx/thnx/tnx* (thanks). Then came phrases. Most of the phrases were short, simple, and fixed, such as *who cares* and *by the way*. Entire English sentences could be short (e.g. *Have fun!*, *I know*) or longer. Partially English elements, consisting of an English and a Dutch element within a single word, such as *awkwardheid*, *kankerchill*, and *5 uur ish*, occurred the least frequently.

Length	#	%
Single-word	6,124	71.1
Textism	1,265	14.7
Phrasal	691	8.0
Sentence	400	4.6
Partial-word	139	1.6
Total	8,619	100

Table 2: Length of English elements.

4.1.2. Multiplicity

Multiplicity is whether multiple English elements occurred together in an online message. The great majority (82.1%) appeared in an item without any other English elements. This again suggests that when code-mixing, Dutch youths used their native language as the matrix language, into which they usually embedded single English elements. Yet at times (17.9%) they inserted more English into one message, as in examples (1)-(2) (English is underlined):

(1) *kunje gwoon subtitles downloade en dan via da programma erbij zette*
'you can just download subtitles and then add them via that program'

(2) *Backpack weegt 20.06kg. Limiet is 20kg om hem in te checken. Gaan ze lopen bitchen om 60 gram*
'Backpack weighs 20.06kg. Limit is 20kg to check it in. Then they bitch about 60 grams'

4.1.3. Word Category

We coded the 6,124 single-word and the 139 partial-word English elements for their word category: noun, verb, adjective, adverb, interjection, or other. Table 3 shows the frequencies according to their word category, for types and tokens. Considering the frequencies for tokens, nouns were the most common word category (35.0%), and then interjections such as *hey*, *thanks*, and *shit* (24.4%). But for types, the distribution is different: the percentage of interjections (7.5%) is much lower when focusing on the variety used in our data. Instead, verbs (14.6%) are the second most frequent category for types, after nouns (57.8%). These results are partly in line with the borrowability hierarchy (Matras, 2007), which states that content words and especially nouns are borrowed cross-linguistically more often and easily than other word categories. Yet the use of English interjections and adverbs is salient, since according to the borrowability hierarchy, such discourse markers or 'utterance modifiers' are not the most obvious candidates for borrowing or code-mixing.

Word category	Tokens		Types	
	#	%	#	%
Noun	2,195	35.0	672	57.8
Verb	850	13.6	170	14.6
Adjective	1,015	16.2	161	13.9
Adverb	621	9.9	52	4.5
Interjection	1,531	24.4	87	7.5
Other	51	0.8	20	1.7
Total	6,263	100	1,162	100

Table 3: Word category of English elements.

4.1.4. Integration

The English elements were also coded for integration, of which we distinguished two kinds – graphemic and morphological. Most (80.2%) were non-integrated, but the rest (19.8%) was integrated in some way. Graphemic integration occurred with 10% of all English elements, mainly with adjectives, adverbs, and interjections. Many of these were commonly integrated in the same way, such as *fuck(ing) > fack(ing)* or *fock(ing)*, *relaxed > relaxt*, *yep / yup > jep / jup*, and *hey > heey*. But some cases of

graphemic integration were less straightforward, such as *thanks* > *fenks*, *nice* > *nais* or *naise*, and *mail* > *meel*: the integration seemed to serve a ludic function here. Morphological integration was present with 9.8% of all English elements. It occurred with verbs (*to check* > *checken*: infinitive, *gecheckt*: past participle), nouns (*app* > *appjes*: diminutive + plural), and adjectives (*cool* > *coole*: inflected). Morphologically integrated past participles often deviated from the Standard Dutch spelling. This was, for example, the case with *updaten*, which was spelled in various ways, e.g. *ge-update*, *geüpdate*, and *geupdate*. A curious, original case of code-mixing is *Ik zit nog steeds te wtf'en* ('I am still wtf'ing'). The textism *wtf* was used as a verb by adding the Dutch infinitive suffix *-en*. 'Double integration', i.e. both graphemic and morphological such as in *meeltjes* and *tjekken*, was rare in our data.

4.1.5. Semantic Field

We visualized the most frequent English elements in a word cloud, presented in Figure 1. This gives us a view of their semantic fields. Many much-used English elements belonged to three semantic fields: computer/technology, affective language, and conversational words. Together with swearing, the latter two fields greatly contribute to online teen talk.

Figure 1: Word cloud of most frequent English elements.

4.1.6. Intentionality

The English elements were divided according to their intentionality. Code-mixing was considered 'intentional' i.e. luxury when there is Dutch expression equivalent to the English element – words such as *nice*, *cool*, and *awesome*. 'Unintentional' code-mixing, on the other hand, is necessary: these English elements do not have an equivalent in Dutch and are used out of lexical need. The English in Dutch youths' CMC was mostly 'intentional' (79.9%), versus 21.1% 'unintentional'. In addition, intentionality interacted with some other factors relevant to code-mixing, such as length: English textisms (e.g. *lol*, *omg*, *wtf*) were almost always 'intentional', as were English sentences and phrases. Many single English words were 'unintentional', as there was no Dutch equivalent available, like *brownie*, *online*, and *high tea*.

4.1.7. Frequency

The top 10s of most frequent English elements and textisms are shown in Table 4. Head and shoulders above the rest of the English elements was the interjection *hey*, quite possibly because the Dutch *hé* has the same pronunciation and meaning. *Nice* came in second: this word is fashionable among Dutch youths to express a positive sentiment. One girl used the textism *lol* extremely often, 373 times; she was clearly an outlier and therefore excluded from our analysis. But even without the 'LOL-girl', *lol* was still the most frequent English textism used, followed by *omg* and *wtf*:

English elements		English textisms		
Lemma	**#**	**Lemma**	**Full version**	**#**
hey	661	lol	laughing out loud	184
nice	225	omg	oh my God	165
mail	202	wtf	what the fuck	111
lol	184	btw	by the way	72
yup	180	idk	I don't know	57
omg	165	thnx	thanks	42
thanks	142	thx	thanks	30
yep	140	k	okay	21
cool	134	np	now playing / no problem	13
mailen	124	ofc	of course	12

Table 4: Most frequent English elements and textisms.

4.2 Language-External Factors

4.2.1. CMC Mode & Age Group

There were differences in code-mixing for CMC mode and age group. MSN chats (2.47%) had the highest relative frequency of code-mixing, closely followed by WhatsApp (2.38%), then Twitter (1.96%), and finally SMS (1.40%). Adolescents (2.60%) used more code-mixing than young adults (2.02%) in all four CMC modes.

Figure 2 presents the mean percentage of English elements of participants divided per CMC mode and age group. For both age groups, WhatsApp messages contained more code-mixing than tweets and SMS text messages. Interestingly, the frequency of code-mixing in MSN chats differed markedly for the age groups: it was the CMC mode with the second highest frequency of code-mixing for adolescents, but with the lowest frequency for young adults.

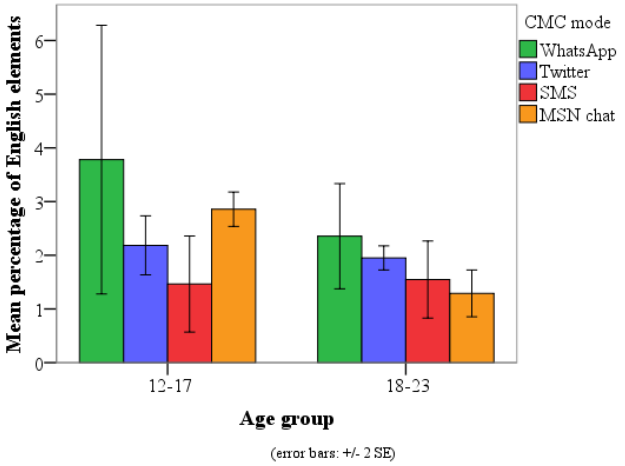

Figure 2: Code-mixing per CMC mode and age group.

A factorial ANOVA, taking into account individual variation in code-mixing between the 318 participants, confirmed that there were main effects of CMC mode, *F*

(3,283) = 4.50, $p < .005$, and age group F (1, 283) = 10.04, $p < .005$, on participants' percentage of English elements. The interaction between them failed to reach significance.

4.2.2. Gender

Boys (2.77%) used somewhat more code-mixing than girls (2.18%), but not significantly so: there was no main effect of gender in an analysis of variance with participants' percentage of English elements as the dependent variable. Still, boys' English elements were of a different nature: they used more words in relation to computers, games, and technology, corresponding to the topics they discussed.[2]

4.3 Some Qualitative Findings

The opposite of the 'usual' code-mixing also occurred: an entire sentence was written in English, with a Dutch element inserted into it, as in *so yes I am sogging* ('so yes I am procrastinating my studies'). The Dutch verb *soggen*, derived from the acronym *sog*, short for *studieontwijkend gedrag* ('study avoiding behaviour') was inflected to fit the English sentence. An English equivalent does not exist.

Example (3) below is a calque: the English phrase 'this shit gets real' has partially been translated into Dutch, but too literally. The issue is the demonstrative pronoun, of which there are two forms in Dutch, *dit* and *deze* ('this'). The gender of *shit* determines that only *deze* is grammatically correct here, but *dit* was used instead.

We also encountered memes in code-mixing, as in (4)-(6). A meme is a usually funny image, video, or piece of text that is copied and spread rapidly on the Internet, often with slight variations. (4) refers to the animated comedy series *Futurama*; (5) to an utterance in a YouTube video that went viral; and (6) to an American actor (Charlie Sheen) who was a fond user of the term 'winning'. If these memes were translated, the references to popular culture would be lost.

(3) *Vanaf daar wordt dit shit real*
'From that point on this shit gets real'
(4) *Can't tell if troll or just very very stupid*
(5) *Double rainbow all the way :p*
(6) *Xbox meenemen naar tentamen #winning*
'Bringing xbox to exam #winning'

5. Discussion & Conclusion

This corpus study has shown that code-mixing with English is present in various ways in Dutch youths' written CMC. It is obviously still subordinate to Dutch, the matrix language, because only about one in forty words (2.5%) of the corpus was an English element. Although the growing impact of English in Dutch society gives reason to dispute Zenner et al.'s (2015) assessment of the contact between Dutch and English being 'weak', our study suggests that English has not taken over at the expense of Dutch in youths' social media messages in the Netherlands. Still, their great use of English discourse particles (interjections, adverbs, and textisms) and use of full English sentences shows that they have an intimate knowledge of English.

A vast majority of the code-mixing in the Dutch written CMC were single English words. These mostly had the word category of nouns or interjections and were often conversational words, such as greetings, affective language, and swear words. English elements longer than one word tended to be (semi-)fixed phrases. Still, the Dutch youths revealed creativity in their code-mixing through graphemic and morphological integration, memes, and puns.

Code-mixing was sometimes out of lexical need: there was no Dutch equivalent for an object, concept, or action. This raises the question whether highly frequent 'unintentional' English elements really still represent code-mixing, or whether these should be considered loanwords and deserve to be listed in Dutch dictionaries without being designated as English. Yet code-mixing also helps to create an online youth language. Dutch youths use popular words such as *nice* and *thanks* and textisms such as *lol* and *omg* in their online messages as part of their 'teen talk' – such English words signal dynamism and are seen as 'cool'. In written CMC, this especially emerges in MSN and WhatsApp, near-synchronous CMC modes which strongly resemble spoken youth language. Ultimately, English elements help Dutch youths, and especially adolescents who were found to use more code-mixing with English in their written CMC, to distinguish themselves from older Dutch speakers and boost their youthful expressivity.

6. References

Canagarajah, S. (2011). Codemeshing in academic writing: Identifying teachable strategies of translanguaging. *Modern Language Journal*, 95(3), pp. 401--417.

De Decker, B., & Vandekerckhove, R. (2012). English in Flemish adolescents' computer-mediated discourse: A corpus-based study. *English World-Wide*, 33(3), pp. 321--351.

Matras, Y. (2007). The borrowability of structural categories. In Y. Matras & J. Sakel, *Grammatical Borrowing in Cross-Linguistic Perspective*. Berlin: de Gruyter, pp. 31--74.

NOG: Nationale Onderwijsgids (2018, April 30). *Tweetalig onderwijs vooral op vwo, maar groei is het sterkst op vmbo en havo.* https://www.nationale onderwijsgids.nl/voortgezet-onderwijs/nieuws/43628-tweetalig-onderwijs-vooral-op-vwo-maar-groei-is-het-sterkst-op-vmbo-en-havo.html

Tagliamonte, S.A. (2016). *Teen Talk: The Language of Adolescents*. Cambridge: Cambridge University Press.

Treurniet, M., & Sanders, E. (2012). Chats, tweets and SMS in the SoNaR corpus: Social media collection. In D. Newman (Ed.), *Proceedings of the First Annual International Conference on Language, Literature & Linguistics*. Singapore: Global Science and Technology Forum, pp. 268--271.

Verheijen, L. (2016). De macht van nieuwe media: hoe Nederlandse jongeren communiceren in sms'jes, chats en tweets. In D. van de Mieroop, L. Buysse, R. Coesemans, & P. Gillaerts (Eds.), *De macht van de taal:*

[2] Only the WhatsApp data were used for this analysis, because the gender of contributors to the SoNaR corpus is unknown.

Taalbeheersingsonderzoek in Nederland en Vlaanderen. Leuven / Den Haag: Acco, pp. 275--293.

Verheijen, L., & Stoop, W. (2016). Collecting Facebook posts and WhatsApp chats: corpus compilation of private social media messages. In P. Sojka et al. (Eds.), *Text, Speech and Dialogue: 19th International Conference*, TSD 2016, LNAI 9924. Springer, pp. 249--258.

Zenner, E., Speelman, D., & Geeraerts, D. (2015). A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show. *International Journal of Bilingualism*, 19(3), pp. 333--346.

Posters

Guided Tour: Donating and Editing WhatsApp Data Using the MoCoDa^2 Web Interface

Michael Beißwenger[1], Marcel Fladrich[2], Wolfgang Imo[2], Evelyn Ziegler[1]

[1] Universität Duisburg-Essen
[2] Universität Hamburg

michael.beisswenger@uni-due.de
marcel.fladrich@uni-hamburg.de
wolfgang.imo@uni-hamburg.de
evelyn.ziegler@uni-due.de

The demo presents results from the corpus project *MoCoDa^2* (*Mobile Communication Database*, https://www.mocoda2.de), which was funded by the Ministry for Innovation, Science, Research and Technology of the German federal state North Rhine-Westphalia and in which a team of researchers from two universities has created a database and web front-end for the repeated collection of written CMC from mobile messaging services such as *WhatsApp*.

MoCoDa^2 adopts a donation-based collection strategy. Unlike other projects that build corpora from data retrieved from mobile messaging applications, in MoCoDa^2 users are not only involved as donators but also as editors of their donated data. After submitting stored logfiles to the project server via email, donators can log into an online interface which provides them with assistant functions for editing their donations: They pseudonymize the data following a pseudonymization guideline, they add textual descriptions for media files embedded into their interactions, and they enhance the data with metadata on the interlocutors and their social relations – information which can only be provided by individuals who were involved in the interactions themselves and which is essential to transform raw data into valuable contributions to a corpus that shall be a useful resource not only for quantitative but also for qualitative research on CMC.

The demo will allow participants of the conference to test how to donate and edit *WhatsApp* data using the *MoCoDa^2* Web interface in a live setting. An additional poster will describe the technology behind the resource and the guidelines for the description of metadata and for pseudonymization.

Developing a Typology of Expertise in Health Forums: Issues and Challenges

Damien De Meyere
Institut langage et communication, UCLouvain
damien.demeyere@uclouvain.be

Nowadays, health forums are draining a significant number of anonymous contributors who are placed on an equal footing in terms of credibility. The popularity of such platforms remains controversial, although it has been shown that they gather 'lay experts' sharing their own medical experience (Boudier et al., 2012), a kind of testimonial that is highly valued by fellow patients (Paganelli & Clavier, 2011). In that context, the aim of our research is to gain a better understanding of the forms of expertise at work within these communities.

While studying user roles and expertise in collaborative platforms is not new, most works analyse activity logs without considering the textual content (Zhang et al., 2007; Bouguessa et al., 2008). To the best of our knowledge, works focusing on content rely on a manual analysis of selected threads (Coulson et al., 2007) or interviews (Lederman et al., 2014). Finally, some works aim at identifying posts written by health specialists (Abdaoui et al., 2016). Our goal is to develop a set of linguistically motivated measures that may be used to automatically detect potential (lay) experts in online communities. In this respect, Doctissimo, the most visited health website in France, gave us access to all the messages posted on the 'Health' forums between 2000 and 2016 (22 million messages).

Assuming that platforms like Doctissimo gather contributors with complementary health-related skills, we developed a dashboard that combines both quantitative (number of messages, dispersion between subforums, volume of activity in time, etc.) and content-based measures (average number of words, number of user mentions, etc.). While these measures show the dynamics of exchanges between users, it is still difficult to determine actual fields of expertise without extensive reading of posts. One way to summarise a user's contribution is to extract the most frequently used keywords (Civan-Hartzler et al., 2010). That is why we developed a named-entity recognition system that relies on a dictionary built upon the French components of the UMLS and terminologies used by the medical community. Even though we can see to what extent Doctissimo members tend to use the specialised vocabulary, terminological resources do not cover linguistic variation at work when non-specialists refer to medical concepts. To bring out the peculiarities of the language used on Doctissimo, all contiguous sequences of 1/2/3 words were extracted and ranked using keyness statistics (Pojanapunya & Watson, 2018) and reference corpora. The comparison of these two analyses can be used to determine whether a user who posts in a particular sub-forum actually tends to master the terminology related to this particular domain of medicine, or what are the mechanisms of linguistic variation at work when non-specialised users refer to medical concepts on online discussion platforms.

Our poster will first present the development stages of our activity analysis dashboard, with a particular focus on the processing of textual data and the keyword analysis. Then, even though it is still an ongoing project, we will show that we can already pinpoint different behaviours that can be related to certain forms of expertise, and to what extent the linguistic analysis confirms trends shown by quantitative analyses.

References

Abdaoui, A., Azé, J., Bringay, S., Grabar, N. and Poncelet, P. (2016). Expertise in French health forums. *Health Informatics Journal*.

Boudier, F., Bensebaa, F. and Jablanczy, A. (2012). L'émergence du patient- expert : une perturbation innovante. *Innovations*, 3(39), pp. 13-25.

Bouguessa, M., Dumoulin, B. and Wang, S. (2008). Identifying authoritative actors in question-answering forums: The case of yahoo! answers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA*, pp. 866-874.

Civan-Hartzler, A., McDonald, D. W., Powell, C., Skeels, M. M., Mukai, M. and Pratt, W. (2010). Bringing the Field into Focus: User-centered Design of a Patient Expertise Locator. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '10)*, pp. 1675‑1684.

Coulson, N. S., Buchanan, H. and Aubeeluck, A. (2007). Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group. *Patient education and counseling*, 68(2), pp. 173‑178.

Lederman, R., Fan, H., Smith, S. P. and Chang, S. (2014). Who can you trust? Credibility assessment in online health forums. *Health Policy and Technology*, 3(1), pp. 13‑25.

Paganelli, C. and Clavier, V. (2011). Le forum de discussion : une ressource informationnelle hybride entre information grand public et information spécialisée. In *Les forums de discussion: agoras du XXIe siècle? Théories, enjeux et pratiques discursives, L'Harmattan (Langue et parole)*, pp. 39-55.

Pojanapunya, P. and Watson, T. R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), pp. 133‑167.

Zhang, J., Ackerman, M.S. and Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th ACM International World Wide Web Conference (WWW'07)*, pp. 221-230.

A Pragmatic Analysis of Discourse Markers in Bengali CMC

Mandana Kolahdouz Mohamadi[1], Mainul Hasan Khan[2]

[1] PNU University, Iran
[2] Gazole Mahavidyalaya, West Bengal, India
manadana.mohamadi@gmail.com

Discourse Markers (DMs) are words or phrases that stick a piece of manuscript or speech together, and therefore, as the smallest discourse unit, can play a fundamental role. In recent years, in addition to English, various studies have been conducted regarding DMs in different languages such as Wouk 2001 (Indonesian); Durán and Unamuno 2001 (Spanish); Low and Deterding 2003 (Singapore English); Wang 2011 (Japanese *ano* and Chinese *nage*); Siti Nurbaya Mohd 2012 (Malaysian) and Palacio and Gustilo 2016 (Filipino).

Fung and Carter (2007) classified DMs into four macro-level functions, namely: (1) structural; (2) referential; (3) cognitive; and (4) interpersonal; but Palacio and Gustilo (2016) modified this framework and provided two macro-level functions, i.e. textual and relational. Their textual DMs correspond to Fung and Carter's structural and referential ones.

Textual DMs mainly manage the structure and coherence of a text, such as cause and effect (*now*, *right*, *well* …), topic shifts (*so*, *and*), sequential relationships (*first*, *next*, *finally*), continuation of topic (*yeah*, *and*, *cos*, *so*) and summary of opinions (*so*). The interpersonal category as an emotive/interactive function indicates attitudes of the speaker and responses like agreement, confirmation, and acknowledgement (*oh*, *alright*, *yeah*…). The cognitive category reflects thinking processes (*well*, *I think*, *I see*, *and*), reformulates (*I mean*, *in other words*), elaborates (*like*, *I mean*), marks hesitation (*well*, *sort of*) and assesses the listener's knowledge (*you know*) (Fung and Carter, 2007, p. 415).

Based on the above mentioned, we conclude that relatively limited research has been carried out on 'written spoken' genres, especially in digital genres. Therefore, the present study aims to contribute to the field of pragmatics and discourse analysis by investigating Bengali and English DMs in (mainly public) Facebook posts and in comments by Bengali users who were native speakers of Bengali, had academic qualifications and were living in West Bengal, India. Based on the authors' experience, Facebook is one of the commonly used apps by Indian people. This study intends to answer the following research questions:

1. Which Bengali and English DMs are used by Bengali users in their Facebook posts and comments?

2. Which major Bengali and English macro-level functions are used by Bengali users?

To shed light on these questions, we carried out a content analysis and reviewed 200 posts and their comments by Bengali users (twenty posts per user) and classified DMs into three macro-level functions, namely: (1) textual, (2) cognitive and (3) interpersonal.

The results indicated that Bengali users tend to use both positive DMs (such as *wow, bah, hmm, baba, haa*) and negative DMs (such as *chiiii, isshhhh* and *baba*). The negativity or positivity of DMs is context-dependent; e.g. in ***baba*** *madam r u know Bengali also, baba* has a positive meaning. It should be noted that in our data, *hmm* was a marker of affirmation, and *hmmmm* indicated thinking. English DMs were more frequent than the Bengali ones, and in English DMs, the textual (47%) and in Bengali the interpersonal (15.38%) category had the highest frequency. The limited amount of data makes it difficult to generalize the linguistic behaviour, therefore, future researchers can expand the list and categories of DMs.

References

Durán, E. M., and Unamuno, V. (2001). The discourse marker a ver (Catalan, a veure) in teacher-student interaction, *Journal of Pragmatics,* 33(2), pp. 193-208.

Fung, L., and Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogical settings, *Applied Linguistics,* 28(3), pp. 410-439.

Low, E. L., and Deterding, D. (2003). A corpus-based description of particles in spoken Singapore English. In D. Deterding, Low, E. L., and Brown, A. (Eds.). English in Singapore: Research on Grammar (pp. 58-66). Singapore: McGraw Hill.

Palacio, M. A., and Gustilo, L. (2016). A Pragmatic Analysis of Discourse Particles in Filipino Computer Mediated Communication. *GEMA Online Journal of Language Studies,* 16(3), pp. 1-19.

Siti Nurbaya Mohd. N. (2012). Discourse Markers in Turn-initial Positions in Interruptive Speech in a Malaysian Radio Discourse. *Multilingua,* 31, pp.113-133.

Wang, Y. (2011). A Discourse-pragmatic Functional Study of the Discourse Markers Japanese *ano* and Chinese *nage*. *Intercultural Communication Studies,* 20, pp. 42-61.

Wouk, F. (2001). Solidarity in Indonesian conversation: The discourse marker *ya*, *Journal of Pragmatics,* 33(2), pp. 171-191.

Leiwand Oida[1]: Geolocating Regional Linguistic Variation of German on Twitter

Bettina Larl[1] & Eva Zangerle[2]
[1] Linguistics, University of Innsbruck, [2] Databases and Information Systems, University of Innsbruck
Bettina.Larl@uibk.ac.at

Keywords: web corpus, geolocation, language variation

Twitter has been used for collecting language data and linguistic research in a variety of languages. (Goncalves & Sànchez 2014; Eisenstein, O'Connor, Smith & Xing 2014; Yuan, Guo, Kasakoff, Grive 2016). The proposed poster demonstrates the process of building a large Twitter corpus containing geolocated Tweets from the Deutscher Sprachraum (German language area) and investigates how German language varieties are used on Twitter.

German is the widest spread language within the European Union. German is a pluricentric language with three standard varieties: German Standard German, Swiss Standard German and Austrian Standard German. The official borders between Germany, Austria and Switzerland also form the official boundaries between the three standards. In addition to those national varieties, there are multiple varieties on the regional and dialectal spectrum. (Ammon 2015; Clyne 1992)

Easy access and its open API has made Twitter a popular source of data for research in various scientific fields and Twitter data shows great potential for linguistic research in multiple areas of expertise. Of particular interest for this poster are the tracking and exploring of regional linguistic variation of German on Twitter: Is there, for example, a connection between the language output and the geographic location tweets were sent from? To address such questions, a Twitter corpus of geotagged German Tweets within the Deutscher Sprachraum has been built. (Larl & Zangerle 2017)

This poster explores and describes the process of building the geotagged Twitter corpus of German tweets as well as giving a first glimpse into version.1 of the corpus.

The corpus version.1 currently contains tweets collected over a period of 30 (+1) months (January 2015 to June 2017).[2]

The Tweets were collected using the public Twitter Streaming API. 85,810,255 geolocated Tweets could be retained within a geographic rectangle (5.85, 46.016667 and 17.1, 55.016667) that covers the Deutscher Sprachraum. These tweets were re-filtered by removing those geolocalised outside of Austria, Germany, Switzerland or Italy (South Tyrol). Twitter's own language detector found 71 different languages within this data set. Subsequently, the corpus was filtered to only retain Tweets that were identified as German. The data was further refined by removing Tweets with missing latitude and/or longitude coordinates or other such deficiencies. In total 18,645,263 German Tweets, sent from within Austria, Germany, Switzerland and the German speaking part of Italy South Tyrol, could be processed and added to the corpus. The data has been tokenised with the SoMaJo-Tokeniser (Proisl, Uhrig 2016) and tagged with the SoMeWeTa-Tagger (Proisl 2018). The Metadata consists of coordinates, town name, country, date, time, ID. Within the corpus you can find Tweets from 452.501 individual users.

The corpus includes texts, hyperlinks and emoticons, as those can be seen as linguistic features. (Beißwenger 2015)

This poster describes the process from data to corpus and explains the various challenges that were encountered

[1] [Viennese; eastern Austria] Awesome, Dude!

[2] The collection process is still ongoing but will end at the end of June 2018. This will result in introducing another 12 months of Tweets – Tweets sent from July 2017 to June 2018 – to the corpus. This corpus version.2 will then contain Tweets with a character limit of 140 and such with a character limit of 280.

along the way. Furthermore, a first version of the corpus on CQP-web (restricted access only!) will be available for preview on sight.

References

Ammon, Ulrich; Bickel, Hans; Ebner, Jakob; et. al. [ed.] (2004). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Walter de Gruyter: Berlin.

Ammon, Ulrich (2015). *Die Stellung der deutschen Sprache in der Welt.* Walter de Gruyter: Berlin/München/Boston.

Ammon, Ulrich (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten.* Walter de Gruyter: Berlin/New York.

Barbaresi, Adrien (2016). Collection and Indexing of Tweets with a Geographical Focus. In: *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC). Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp.24-27.

Beißwenger, Michael (2015). Sprache und Medien: Digitale Kommunikation. In: *Studikurs Sprach- und Textverständnis. E-Learning-Angebot der öffentlichrechtlichen Universitäten und Fachhochschulen und des Ministeriums für Innovation, Wissenschaft und Forschung (MIWF) des Landes Nordrhein-Westfalen.*

Beißwenger, Michael, Horsmann, Tobias, Zesch, Torsten (2017). Part-of-speech Tagging for Corpora of Computer-mediated Communication: A Case Study on Finding Rare Phenomena. In: Fišer, Darja, Beißwenger, Michael (Eds.): *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World.* Ljubljana: Ljubljana University Press (Translation Studies and Applied Linguistics), pp. 192-219.

Bouvier, Gwen (2015). What is a discourse approach to Twitter, Facebook, YouTube and other social media: Connecting with other academic fields? *Journal of Multicultural Discourses*, 10(2), pp. 149-162.

Clyne, Michael (1992). German as a Pluricentric language. In: Clyne, Michael [ed.] (1992): *Pluricentric Languages. Differing Norms in Different Nations*. Mouton de Gruyter: Berlin, New York, pp. 117-148.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, 9, e113114.

Goncalves, Bruno & Sànchez, David (2014). Crowdsourcing Dialect Characterization through Twitter. *PLoS one,* 9(11), e112074.

Larl, Bettina & Zangerle, Eva (2017). Geolocating German on Twitter. Hitches and glitches of building and exploring a Twitter corpus. *9th International Corpus Linguistics Conference, 24 to Friday 28 July 2017, University of Birmingham.*

Proisl, Thomas (2018). SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), pp. 665–670.

Proisl, Thomas, Peter Uhrig (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In: *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task.* Berlin: Association for Computational Linguistics (ACL), pp. 57–62.

Scheffler, Tatjana (2014). A German Twitter Snapshot. In: *Proceedings of LREC, Reykjavik, Iceland.*

Storrer, Angelika (2013). Sprachstil und Sprachvariation in sozialen Netzwerken. In: Frank-Job, Barbara, Mehler, Alexander, Sutter, Tilmann [ed.] (2013): *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchen an Beispielen des WWW*. Springer Fachmedien: Wiesbaden, pp. 331-366.

Yuan, Hauang, Guo, Diansheng, Kasakoff, Alice, Grive Jack (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59, pp. 244-255.

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

Varying Background Corpora for SMT-Based Text Normalization

Claudia Matos Veliz, Orphée De Clercq, Veronique Hoste
LT^3, Language and Translation Technology Team, Ghent University
{Claudia.MatosVeliz, Orphee.DeClercq, Veronique.Hoste}@UGent.be

Social media text has become a huge source of information for researchers, companies and institutions in the past decade. One of the main characteristics of social media is the use of non-standard language among their users. Since Natural Language Processing (NLP) tools have been trained on traditional text material, this has led to an increased interest in the task of text normalization (Clark and Araki, 2011; Pennell and Liu, 2014).

In this work we applied text normalization techniques to two different languages, English and Dutch; and performed experiments on noisy text coming from different social media genres. We tackled the normalization problem using a Statistical Machine Translation (SMT) approach, taking advantage of its use of contextual information during translation (Aw et al., 2006). The objective is to go from noisy to standard text using SMT techniques. In order to do so, we relied on existing Dutch (Schulz et al., 2016) and English (De Clercq et al., 2014) corpora that were manually normalized. Three social media genres are included in both languages: text messages (AskFM and DutchSMS), message board posts (Youtube and Netlog) and tweets. Regarding the level of noise, in both languages the text messages required most normalization operations.

When applying SMT to the normalization task it is crucial to select a language model (LM) that is trained on a background corpus which is close to the standard in order to correctly transform the noisy text input. For social media one could thus suspect that depending on the level of noise of the data, the use of different corpora for training the SMT model, should lead to better results. For the English experiments we relied on three different background corpora for constructing our LMs: the OPUS corpus (Tiedemann, 2009), Europarl (Koehn, 2005), and the combination of both. Similarly, for Dutch we used an in-house subtitles dataset, Europarl, and the combination of both. Due to the unavailability of social media text corpora for the task, we needed to find a resource close enough to the target domain. We believe that by using corpora based on subtitles and Europarl we can cover spoken language which is close to the user-generated content that we can find in social media texts. We trained LMs at character (unigram and bigram) and token level. For building the SMT model we used Moses (Koehn et al., 2007). All LMs were built using the SRILM toolkit (Stolcke, 2002) with Witten-Bell discounting which has proven to work well on small data sets (Tiedemann, 2012). We experimented with our data using 80% for training the model and 20% for development and test. To evaluate the performance of each LM, Word Error Rate (WER) was calculated.

Experiments revealed that best results were achieved with SMT at the token level for all genres. Regarding the different corpora that were used to construct the LM, we found that Europarl gave the best results for the genre with the least noise (tweets), i.e. WER of 4% and 6.3% for English and Dutch respectively. This could be expected, since the word usage in both sides, the LM and the test data, is very close. The same is true for the genre comprising the most noise (text messages), where we obtained a WER of 9.5% for English using OPUS, and a WER of 12% for Dutch using a combination of Europarl and subtitles. Considering our results, it seems to be important to make variations in the background data for building the LM, depending on the amount of noise and vocabulary that is present in the social media genre.

References

Aw, A., Zhang, M., Xiao, J., and Su, J. (2006). A phrasebased statistical model for sms text normalization. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 33–40. Association for Computational Linguistics.

Clark, E. and Araki, K. (2011). Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia-Social and Behavioral Sciences*, 27, pp. 2–11.

De Clercq, O., Schulz, S., Desmet, B., and Hoste, V. (2014). Towards shared datasets for normalization research. In Nicoletta Calzolari, et al., editors, *LREC 2014 - Ninth International Conference on Language Resources and Evaluation*, pp. 1218–1223. European Language Resources Association (ELRA).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177–180. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, 5, pp. 79–86.

Pennell, D. L. and Liu, Y. (2014). Normalization of informal text. *Computer Speech & Language*, 28(1), pp. 256–277.

Schulz, S., De Pauw, G., De Clercq, O., Desmet, B., Hoste, V., Daelemans, W., and Macken, L. (2016). Multimodular text normalization of Dutch user-generated content. *ACM Transactions on Intelligent Systems and Technology*, 7(4), p. 24.

Stolcke, A. (2002). Srilm-an extensible language modelling toolkit. In *Seventh international conference on spoken language processing.*

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors*, Recent Advances in Natural Language Processing*, volume V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tiedemann, J. (2012). Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 141–151. Association for Computational Linguistics.

Is This Common Sense?
Discursively Creating Community on a Japanese Online Messaging Board

Giancarla Unser-Schutz
Rissho University
giancarlaunserschutz@ris.ac.jp

This poster reports on an ongoing analysis of a corpus of posts from the popular Japanese online messaging board, *Hatsugen Komachi* [Small Talk Town], focusing specifically on how users discursively create a sense of community and shared values through the term *jooshiki* [common sense]. Contemporary Japan is said to be characterized by the supposed fragmentation of shared values (Yamada, 2009). Although this is often presented as a social crisis, it does not mean that people are uninterested in achieving a sense of community; on the contrary, the attention given to these issues signals people's desire to create such connections. By allowing for interactions with a wider net of people, social media can play an important role in building such senses of community. However, building communities involves negotiation as people deal with different ideas; advice giving can be particularly risky as it entails evaluating behavior, thus requiring face work to maintain positive interactions. Consequently, it can be anticipated that *Hatsugen Komachi*, a semi-moderated Q&A site (Harper, Raban, Rafaeli, & Konstan, 2008) where most interactions involve the requesting and dispensing of advice, offers abundant opportunities to examine how people discursively create a sense of community.

Given that such negotiations often involve legitimization through calling to *jooshiki* and its antonym, *hijooshiki* [lacking common sense] (Unser-Schutz, in press), a corpus of 391 posts discussing *(hi)jooshiki*—as determined by its appearing in the subject line—from 2013 to 2017 was compiled using Sketch Engine. Tags were added for posters' gender (*male*, *female*, *unspecified*), which is often specified in posts or through posting to the advice category *from men*. To ascertain how it frames posts, how *jooshiki* was used in each post's subject line was analyzed, followed by frequency and concordance analyses of posts and responses, focusing on how users repeated, reframed and rekeyed *(hi)jooshiki* to create interpersonal involvement (Tannen, 2006), which can contribute to a sense of shared values. Preliminary analyses show that although *jooshiki* is most frequently used initially to confirm the normalcy of behavior in the form of questions (e.g., "is ~ common sense?": ~45% of all uses in subject lines), how it reappears in responses depended on whether the original poster's stance was supported. The original posts and their responses were frequently marked with pragmatic discourse markers indicating alignment with other users, like the sentence final particle *yo-ne*, used to confirm that information is shared (McGloin, et al., 2014); this can be taken as an indication of the face work conducted to maintain relationships. Consequently, it can be said that the negotiation of normalcy—and thus shared values and beliefs—is done through both direct (e.g., questions) and indirect (e.g., pragmatic discourse markers) channels. As a forum largely marketed and described as being for women (Inazawa, 2011), this research also complements and fills in gaps in previous research on Japanese messaging boards, which often do not assess the role of gender or focus on forums largely assumed to be used by men, such as 2channel (e.g., Matsumura et al., 2004, Shiki, 2017).

References

Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of answer quality in online Q&A sites. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 865–874.

Matsumura, N., Miura, A., Shabanai, Y., Ohsawa, Y, & Nishida, T. (2004). The dynamism of 2channel. *AI & Society*, 19(1), pp. 84-92.

McGloin, N., Hudson, M.E., Nazikian, F., & Kakegawa, T. (2014). *Modern Japanese Grammar Workbook*. Routledge: London.

Inazawa, Y. (2011). Owari ni. In Ootekomachi Henshuubu (Ed.), *Ta'nin no Nanigenai Hitokoto ni Tasukeraremashita.* Chuokoron-Shinsha: Tokyo, pp. 242--245.

Shiki, Y. (2017). Terebi bangumi o wadai to shita 2channerujoo no komyunikeeshon ni kansuru kentoo: Twitter to hikaku o tooshite. *Keio Media and Communications Research*, (67), pp. 83-96

Tannen, D. (2006). Intertextuality in interaction: Reframing family arguments in public and private. *Text & Talk*, 26(4/5), pp. 597-617.

Unser-Schutz, G. (In press). Persuasion through commonality: Legitimizing actions through discourse on common sense in a Japanese advice forum. In A.S. Ross & D.J. Rivers (Eds.), *Discourse of (De)Legitimization: Participatory Cultures in Digital Contexts*. London: Routledge.

Yamada, M. (2009). *"Futsū" to iu Kibō*. Tokyo: Seikyusha.

Lexical Normalization for Dutch Social Media Texts

Rob van der Goot
University of Groningen
r.van.der.goot@rug.nl

Lexical normalization is the task of translating ill-formed or non-standard text to a more standard register. This can be helpful for many natural language processing pipelines, since they are usually trained on standard texts. These natural language processing systems simply break down when they encounter the noisy text from social media domains. Below we show an example of a normalized Dutch tweet:

tgaat goed , vdg rustig aaan .
Het gaat goed , vandaag rustig aan .

There is already some previous work on normalization for Flemish (De Clercq et al., 2013; Schulz et al., 2016). On this dataset, the performance of a state-of-the-art normalization model (van der Goot & van Noord, 2017) is much lower compared to the English corpora: 42.5% vs. 86.4%. Upon inspection of the different corpora this is due to the fact that this corpus also includes transformation of some Flemish words into Dutch, and to the difference in size of the training data. But even when using the same amount of training data, the performance difference remains.

Besides this, the corpus also makes no distinction between punctuation and normalization edits; this corpus actually contains 1,322 tokenization replacements and only 708 normalization replacements. This results in tokenization being far more important for the final evaluation. On top of that, the corpus is not publicly available and capitalization use is not corrected.

We will annotate a new dataset of 1,000 noisy sentences taken from the SoNaR corpus (Oostdijk et al., 2013) with a normalization layer. This can be used to train a normalization model and confirm if Dutch is really a more difficult language to normalize. 150 sentences will be annotated by two annotators to obtain an inter-annotator agreement. This also enables inspection of the type of disagreements.

We will train the existing normalization model MoNoise (van der Goot & van Noord, 2017). This normalization model is modular, because the normalization task comprises of different normalization replacements. The most important modules to generate candidates are:
- Aspell: lexical and phonetic edit distances
- Lookup list: generated from the training data
- Word embeddings trained on Dutch tweets data collected between 2012-2016, using the same method as described in Tjong Kim Sang and van den Bosch (2013): the top-n closest words in the embedding space are used as candidates

The model uses features from the generation modules as well as some additional features. From the additional features, the N-gram features are by far the best predictor, these include unigram and bigram probabilities from both standard and non-standard texts. A random forest classifier is used to predict the best normalization candidate based on these features. At the conference, we will present a live demo of the normalization model.

References

Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, & Lieve Macken. (2016). Multimodular text normalization of Dutch user-generated content. *ACM Transactions on Intelligent Systems Technology*, 7(4), 61:1–61:22.

Orphée De Clercq, Sarah Schulz, Bart Desmet, Els Lefever, & Véronique Hoste. (2013). Normalization of Dutch user-generated content. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, & Ineke Schuurman (2013). The construction of a 500 million word reference corpus of contemporary written Dutch. In *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*.

Tjong Kim Sang, Erik, & Antal van den Bosch. Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3.

van der Goot, Rob, & Gertjan van Noord. (2017). Monoise: Modeling noise using a modular normalization system. *Computational Linguistics in the Netherlands Journal*, 7.

The Significance of Authenticity in a Multimodal Online Genre: A Metapragmatic Analysis of YouTube Consumer Reviews

Michael Wentker
University of Duisburg-Essen
michael.wentker@uni-due.de

Online consumer reviews are a relatively recent CMC (computer-mediated communication) genre and have thus far primarily been studied in their text-based form (cf. Vásquez 2014). Such text-based reviews (on sites such as Amazon, Netflix, Yelp and TripAdvisor) seem to be favoured by information-oriented users, while this study indicates that users on the multimodal platform YouTube primarily watch online reviews for entertainment purposes. In other words, the evolution of multimodal reviews seems to coincide with a functional shift from information- to entertainment-seeking purposes. This becomes especially visible when users comment on the functionality and credibility of reviews. Such metapragmatic commentary (cf. Bublitz & Hübler 2007) serves as rich and valuable data for investigating users' notions of authenticity, a central concept when users, implicitly or explicitly, evaluate reviews with regard to their purposes.

This research is interested in the metapragmatic construction and discussion of authenticity in the follow-up comments of YouTube consumer reviews. Adopting a first-order/audience perspective, this study illustrates what it is that makes a review authentic for the users. More specifically, it systematically analyses the communicative triggers that give rise to metapragmatic discussions of authenticity, and the linguistic strategies used for its negotiation. Furthermore, this study provides insights into how claims of authenticity are linked to both the discursive construction of (enacted) reviewer identities as well as the perceived breach of genre-specific norms. Furthermore, the findings are used to assess the role that multimodality plays in such a process.

Following a discourse-analytic approach, this pilot study focuses on a small corpus of reviews and their follow-up comments. A qualitative analysis provides a synchronic snapshot of how authenticity is metapragmatically discussed among YouTube users, paying special attention to the socio-technical idiosyncrasies and affordances (cf. Herring 2007, Page et al. 2014) of YouTube as a CMC mode.

References

Bublitz, Wolfram & Axel Hübler. (2007). *Metapragmatics in Use*. Amsterdam/Philadelphia: John Benjamins.
Herring, Susan C. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet,* 4, pp. 1–37.
Page, Ruth, David Barton, Johann W. Unger & Michele Zappavigna. (2014). *Researching Language and Social Media. A Student Guide.* London/New York: Routledge.
Vásquez, Camilla. (2014). *The Discourse of Online Consumer Reviews.* London: Bloomsbury.

A Multi-Layered Corpus of Namibian English

Frederic Zähres
Bielefeld University
fzaehres@uni-bielefeld.de

English was declared the sole official language in Namibia upon its independence in 1990, although the country has never been under British rule. Even in 2011, only 3.4% of Namibians reported that they speak English as their primary home language (cf. NSA 2012). English does, however, play a major role in official and inter-ethnic communication in Namibia and is favored especially by young Namibians born and raised after independence (cf. Stell 2014a; 2014b; 2016). Research on Namibian English is still in its infancy and until now, the studies are based on questionnaire data, sociolinguistic interviews, and data from experimental research designs (cf. Buschfeld & Kautzsch 2014; Kautzsch & Schröder 2016; Schröder & Schneider fc.). The present investigation adds to this field by providing a data source of naturally-occurring language that is particularly relevant for studying sociophonetic aspects of this emerging variety of English.

This project envisions at compiling a multi-layered corpus containing YouTube video data of Namibian YouTubers, which is complemented with written data from the comment sections and further social media accounts of the respective YouTubers as well as qualitative interview data and metadata consisting of basic sociolinguistic variables. This collection of CMC data can, on the one hand, be analyzed using acoustic phonetic methods, but its digital ethnographic nature also allows an approach to qualitative analysis that is driven by key notions of third-wave sociolinguistics (cf. Eckert 2012). Such a corpus therefore combines three fields of study, CMC, sociolinguistics, and phonetics, in a completely new way. Third-wave sociolinguistic studies are usually based on ethnographies conducted in small communities, and the data collected in such settings are most often not usable for studying phonetic and phonological aspects of language production. Researchers working in sociophonetics have usually supplemented ethnographic data with audio-recorded sociolinguistic interviews (cf. e.g. Drager 2009). By incorporating CMC data into a sociolinguistic study, the complete dependence on interview data for sociophonetic investigation can be circumvented and naturally occurring language can be used as an additional data source. Also, by using language data from multiple contexts, both naturally occurring and based on interviews, sociolinguistic ethnography can be expanded to use digital forms of communication. This is especially relevant for the study of Namibian English, as outer and expanding circle varieties of English are generally under-researched compared to inner-circle varieties such as British or American English. We know only little about how language is used in digital contexts by young Namibians, and the combination of CMC research with cutting-edge sociolinguistic approaches will help to shed light on this question.

I will present the pilot corpus, which consists of 300 minutes of YouTube videos by 15 self-identified Namibian content creators including orthographic transcriptions of the language used in the videos as well as the comment sections of the respective videos (as of 31 July 2018). I will provide two case studies to test the usability of this database: The first one is a sociophonetic case study, which analyzes the audio layer and investigates whether the NURSE-WORK vowel split described in recent work based on sociolinguistic interviews with Namibians (cf. Kautzsch, Schröder & Zähres 2017) is also found in the CMC data. This is significant because Namibian English has traditionally been aligned with South African Englishes but may now be establishing itself as an independent variety that diverges from South African Englishes. The NURSE-WORK split will be analyzed with acoustic phonetic methods, using standard programs and procedures from that field, in particular Praat (Boersma & Weenink 2018) and R. The second case study will also make use of the other layers of the corpus by investigating features identified as typical for NamE in Kautzsch (in prep.), a study using the online newspaper corpus CNamOn. I will compare the use of bare infinitive constructions containing go and between CNamOn and the various layers of the present corpus. The results confirm previously observed features in naturally-occurring data and the ongoing nativization process of English in Namibia.

References

Boersma, Paul and David Weenink. (2018). *Praat: Doing Phonetics By Computer*. Version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/

Buschfeld, Sarah and Alexander Kautzsch. (2014). English in Namibia: A First Approach. *English World-Wide,* 35(2), pp. 121-160.

Drager, Katie. (2009). *A Sociophonetic Ethnography of Selwyn Girls' High*. Doctoral Dissertation, University of Canterbury.

Eckert, Penelope. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology,* 41, pp. 87-100.

Kautzsch, Alexander. (In prep). Namibian English on the web: Lexical and morpho-syntactic features in a Corpus of Namibian Online Newspapers (CNamON). In Alexandra Esimaje, ed. *Corpus Linguistics in Africa*. Amsterdam: John Benjamins.

Kautzsch, Alexander and Anne Schröder. (2016). English in Multilingual and Multiethnic Namibia: Some Evidence on Language Attitudes and on the Pronunciation of Vowels. In Christoph Ehland, Ilka Mindt, and Merle Tönnies, eds. *Anglistentag 2015 Paderborn. Proceedings*. WVT: Trier, pp. 277-288.

Kautzsch, Alexander, Anne Schröder and Frederic Zähres. (2017). The Phonetics of Namibian English: Investigating Local Features in a Global Context. Paper presented at the IAWE Conference, June 30 – July 2, 2017, Syracuse University, NY.

Namibia Statistics Agency (NSA). (2012). Namibia 2011 Population and Housing Census Main Report. <http://nsa.org.na/page/publications>

Schröder, Anne and Klaus P. Schneider. (Fc). Variational Pragmatics, Responses to Thanks, and the Specificity of English in Namibia. *English World-Wide,* 40(1).

Stell, Gerald. (2014a). Social identities in post-Apartheid intergroup communication patterns: Linguistic evidence of an emergent nonwhite pan-ethnicity in Namibia? *International Journal of the Sociology of Language,* 230, pp. 91-114.

Stell, Gerald. (2014b). Use and Functions of English in Namibia's Multiethnic Settings. *World Englishes,* 33(2), pp. 223-241.

Stell, Gerald. (2016). Trends in linguistic diversity in post-independence Windhoek: A qualitative appraisal. *Language Matters,* 47(3), pp. 326-348.

Appendix

Author Index

Keyword Index