Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora19)

Julien Longhi, Claudia Marinica

Conference Proceedings

Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)

9-10 September 2019

Cergy-Pontoise University, France

Editors:

Julien Longhi

Claudia Marinica

Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)

Editors: Julien Longhi, Claudia Marinica

Cergy-Pontoise, 2019

Conference web site: https://cmccorpora19.sciencesconf.org/
This publication is available from: https://cmccorpora19.sciencesconf.org/resource/page/id/15

This publication was supported by

Preface

This volume presents the proceedings of the 7th edition of the annual conference series on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019). This conference series is dedicated to the collection, annotation, processing, and exploitation of corpora of computer-mediated communication (CMC) and social media for research in the humanities. The annual event brings together language-centered research on CMC and social media in linguistics, philologies, communication sciences, media and social sciences with research questions from the fields of corpus and computational linguistics, language technology, text technology, and machine learning.

The 7th Conference on CMC and Social Media Corpora for the Humanities was held at IDHN (Institute of digital humanities) on Septembre, 9th and 10th, in Cergy-Pontoise, France. This volume contains extended abstracts of the papers (14), and abstracts of posters presented at the event (4). The program also includes two invited talks: one keynote talk by Marty Laforest (Université du Québec à Trois-Rivières, Canada), and one by Julien Velcin (University Lumière Lyon 2).

The contributions in these proceedings cover a wide range of both topics and languages. Some contributions focus on standards and best practices of CMC corpora, others on the pragmatics of CMC, others on geographic linguistic variation, or applied linguistics, with discursive, semantic, or computational point of views.

We wish to thank all colleagues who have contributed to the conference and to this volume with their papers and posters. 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), Bolzano, Italy (2017), Antwerp, Belgium (2018), we hope that the Cergy-Pontoise 2019 conference will mark another step for the CMC corpora community.

Committees

Scientific Committee

Chair

Julien Longhi (IDHN, AGORA lab, Paris Seine University, France)
Claudia Marinica (IDHN, ETIS lab, Paris Seine University, France)

Co-Chairs

Michael Beißwenger Universität Duisburg-Essen
Steven Coats (University of Oulu, Finland)

Members

Markus Bieswanger (Universitaet Bayreuth, Germany)
Tomaž Erjavec (Jožef Stefan Institute, Slovenia)
Darja Fišer (University of Ljubljana, Slovenia and Jožef Stefan Institute)
Aivars Glaznieks (EuRac Research, Italy)
Axel Herold (Berlin-Brandenburgische Akademie der Wissenschaften, Germany)
Lisa Hilte (University of Antwerp, Belgium)
Gudrun Ledegen (U. Rennes 2, France)
Harald Lüngen (Institut für Deutsche Sprache, Germany)
Céline Poudat (U. Nice, France)
Muge Satar (Newcastle University, United Kingdom)
Stefania Spina (University for Foreigners, Italy)
Egon W. Stemle (EuRac Research, Italy)
Angelika Storrer (Universitaet Mannheim, Germany)
Reinhild Vandekerckhove (University of Antwerp, Belgium)
Lieke Verheijen (Radboud University, Netherlands)

Organizing Committee

Chair

Julien Longhi (IDHN, AGORA lab, Paris Seine University, France)
Claudia Marinica (IDHN, ETIS lab, Paris Seine University, France)

Members

Zakarya Després (IDHN, Paris Seine University, France)
Laurene Renaut (AGORA lab, Paris Seine University, France)
Olivier Belin (IDHN, LT2D lab, Paris Seine University, France)
Boris Borzic (IDHN, ETIS lab, Paris Seine University, France)
Abdelouafi El Otmani (AGORA lab, ETIS lab, Paris Seine University, France)

International steering committee (conference series)

Michael Beißwenger (Universität Duisburg-Essen, Germany)
Darja Fišer (Univerza v Ljubljani, Slovenia)
Steven Coats (University of Oulu, Finland)

Table of Contents

Online auction listings between community and commerce

Annette Gerstenberg, Valerie Hekkel, Freya Hewett
University of Potsdam, Germany
{gerstenberg,hekkel,hewett}@uni-potsdam.de

Abstract

We present a corpus of online auction listings gathered from the platform eBay. It covers two time-frames (2005 and 2017/2018) and two different situational settings (private vs. professional sellers) which define the four sub-corpora. We present the listings as instances of a hybrid genre whose variability can be traced along three dimensions (printed classified adverts, community language and marketing language). The linguistic features identified as indicative for these dimensions are investigated in their distribution over the four sections and for their suitability for genre prediction, using a classification tree.
Keywords: cmc, marketing language, genre, stance

1. Introduction

Our study sheds light on a particular kind of online text: the online auction listing, rarely dealt with in studies on computer-mediated communication (CMC), even though recent studies on the topic include a very broad range of sub-genres showing "transfer, emergence, and transformation" (Giltrow, 2013, p. 718). Online auction listings are particularly interesting in the context of genre change and variability: they have their roots in printed classified adverts (1st dimension) while taking advantage of the increased freedom of the online medium, and they combine the competing task of reaching out to the community of "users" (2nd dimension) while also successfully selling a product (3rd dimension). In what follows, we trace these dimensions using a four section corpus of 1348 eBay listings, gathered in two periods (2005 and 2017/2018) and in two situational settings (*private* for 2005, *private* vs. *professional sellers* for 2017). The fourth corpus section consists of auction listings that were posted in 2018 and that include the French stance marker *vraiment*. We analyse the distribution of the linguistic features from the three dimensions over the four corpus sections and also use various methods to identify the features which best predict genre attribution.

2. Three dimensions of genre variation

2.1 Samples

The first wave of the corpus (e05p) consists of 300 listings randomly gathered from the French eBay site (Gerstenberg, 2007): an empty search was submitted, which returned all active listings on the site and therefore a wide range of the different categories. In order to exclude professional sellers, the listings were pruned so that only users with less than 200 ratings were included, and each user was featured only once in the corpus. Additionally, listings with extensive delivery or returns information were excluded, as well as listings from shops. In 2017, we replicated this corpus (e17p) and created an additional corpus with listings from professional users (users with a shop and more than 200 listings, manually controlled, corpus e17x). These 3 sections of the corpus have the same distribution of the eBay product categories (*maison* 'home': 41 listings, *vêtements* 'clothing': 122 listings, etc.). In 2018, we used the web scraping tool ParseHub to automatically collect more listings. As Biber and Egbert (2016) suggest, stance markers are indicative for genre selection. This turned out to be the case with the French stance marker *vraiment* 'really'. This adverb was occasionally used in e05p, in ads featuring a very personal style: it is used to emphasise the credibility of the individually liable seller, who claims to offer an object in "really good" condition ("really rarely used"). As a single word, it has the advantage to be easily included in the ParseHub query and to return more listings than alternatives such as verbal assertions ("I think it's a nice piece"). In this way, more than 10,000 listings from potentially private sellers could be gathered. In order to assure the private nature of the listings, they were filtered down to a maximum of one listing and 1000 ratings[1] per user (only 503 listings remained). Additionally, listings containing descriptions that were copy-pasted from elsewhere or expressions that indicated a professional activity[2] were excluded; finally, we manually coded for usage of *vraiment* as an actual stance marker which was the case for 356 ads, included in corpus e18v. This information is summarised in Table 1 below, along with information on the average length of the listings.

Sub-corpus name	Year	Type of seller	No. of listings	Average length of listing (tokens)
e05p	2005	private	300	43
e17p	2017	private	300	49
e17x	2017	professional	300	177
e18v	2018	private	356	97

Table 1: Summary of sub-corpora

As can be seen, the average amount of tokens in the listings in each sub-corpus has increased since 2005, and the difference between professional and private users is also

[1] After more than 15 years of eBay activity in France, eBay has increased in popularity. In 2015, 4m sellers were active (eBay France 1995–2019) and we therefore adjusted our upper limit for ratings for the e18v corpus, as even private users have a higher number of ratings.

[2] *mon stock, mes photos, ma boutique, mes autres, shipping, tracklist, welcome, ask, regroupez, regroupe*

reflected in the length of the listing (Table 1). The functionality of eBay has also changed since 2005, with more sophisticated templates for adding metadata to the listing, for example. These structural changes, the growing popularity of the platform, and of course the ubiquity of technology in modern life also inevitably play a role in the evolution of the genre.

2.2 Dimensions

As we are building on a study from 2005, we have used the same dimensions that were used for the analysis of the sub-corpus e05p (cf. Gerstenberg, 2007, p. 374): classified adverts, (online) community language and online marketing. Classified adverts are typically found in a newspaper or directory and are sold on a word or line basis (Danna, 2006, p. 327). They often feature a heading which symbolises the category which the classified ad belongs to, such as job offers, property or objects for sale. As classifieds are now found on many online platforms, we make a clear distinction between printed classified adverts and their online counterpart: in this study the term 'classifieds' refers to the traditional printed advert. Online community language refers to the unique linguistic traits used online by groups of people united by a common theme. On a platform such as eBay, the "community" is shaped by the profile of ratings and earlier transactions, which give a sense of credibility, and a fairly personal username. The communication on eBay can be reciprocal, as every potential seller is a potential buyer, emphasised by the Q&A functions embedded in the listing form. Thanks to the interactive structure (Janoschka, 2004, p. 59), users can virtually join more distinct communities which are associated with the products and thus consist of "collectors of 19th century porcelain", "parents at a virtual flea market for children's clothes", "bikers with an interest in tailpipes", and other niche sub-communities. We use the term 'online marketing language', to refer to language used to encourage the "addressees' communicative integration" and "emotionally motivating strategies" which create a positive image of the item (Janoschka, 2004, p. 132; 146). The corpus architecture makes it possible to determine if different linguistic means of these strategies are employed in the professional vs. private sections.

2.3 Linguistic features

Using a top down approach appropriate for the given data, various frequent features considered to be indicators for the three dimensions were defined (Gerstenberg, 2007) and manually tagged (if applicable, one per listing, Table 2). Only features which occurred in more than 5% of the listings were included in the quantitative analysis.

Classifieds (1st dimension). Due to the character limit in printed classifieds, abbreviations such as *tbe* 'très bon état, very good condition' are often made use of; it is also common practice to add 'for sale' (both tagged with <ann>) or a positive attribute (*super*) at the beginning of a classified ad (<bon>).

Dimension and feature	Example	Total n listings
1 ann[once]	*Vds* '[I] sell'	110
1 bon [attributes]	*Magnifiques lunettes* 'great glasses'	140
2 ego [1st person]	*Je l'ai portée une fois* 'I wore it one time'	246
2 stn [standard:no]	*elle est neuve jamais porter* 'she's new, never worn' [non standard *-er* < *-ée*]	72
2 pre[sentative]	*il y a deux étiquettes* 'There are two labels'	104
3 att[ributes]	*état superbe* 'super condition'	312
3 enc[hères, 'biddings']	*bonnes enchères* 'happy bidding'	97
3 imp[erativ]	*n'hésitez pas à me contacter* 'don't hesitate to get in touch'	176

Table 2: Three dimensions: indicative linguistic features

Another loan from traditional written communication (letters) is starting with a greeting (*bonjour*), with 48 occurrences in the whole corpus (and thus excluded from the statistical analysis), most frequently in e18v (n=25).

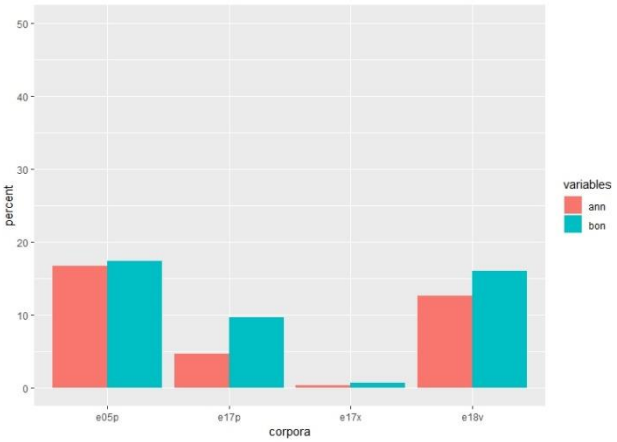

Figure 1: Elements commonly found in classifieds

The features *ann* and *bon* are in fact extremely rare in the professional corpus (e17x, Figure 1). The decrease in *ann* in e17p in comparison to e05p could suggest that the genre is moving away from its offline counterpart, whilst the high proportion of ads in e18v with these characteristics show that a sub-group of users are still using these traditional genre markers (Biber & Conrad, 2009).

Community language (2nd). Online interaction within established communities is typical for certain genres (Crowston & Williams, 2000), which may have certain characteristics such as a permissive attitude towards orthographic and grammatical standards. Non-standard variants and variants frequent in spoken French (e.g. morphology, orthography, *ne* deletion in negation constructions; Koch & Oesterreicher, 2011) were coded

according to whether they adhere to the standard or not (<stn>). Additionally, we coded for the presentative constructions *il y a* and *c'est* 'there is', typical for colloquial French (<pre>). The personal involvement of the sender was coded using the presence of the first-person singular pronoun *je* (<ego>). Prototypical CMC features such as emojis rarely occur in the corpus (28 instances).

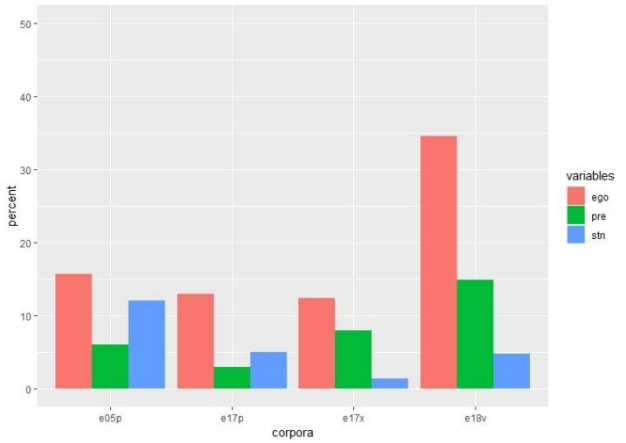

Figure 2: Elements which evoke community language

The personal pronoun *je* 'I' was highly present in all corpora (Figure 2), including e17x, with the highest rates in e18v. Presentative constructions *c'est* and *il y a* 'it is / there is/ are' are frequent in e17x and e18v. Non-standard variants (<stn>) slightly decreased between e05p and e17p. They were very rarely used in the control corpus e17x.

Online marketing language (3rd). Unlike classifieds, where space is limited, an online auction listing can be as long as the user desires, allowing the user to make use of many techniques to tempt the prospective buyer.
Possible influences from the advertising world were coded in the corpus: a restricted set of the most frequent positive evaluative attributes (such as *magnifique*, *parfait*: tag <att>), the most common imperative forms used to encourage the buyer to look at the seller's other listings (*consultez*, *regardez*), for example, or assuring the buyer to not hesitate in getting in touch with the seller (<imp>). *Bonnes enchères* 'happy bidding' features a different distribution and was coded separately (<enc>).

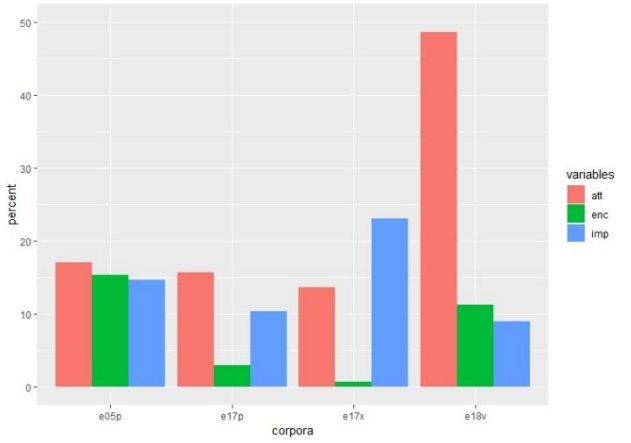

Figure 3: Features of marketing language

The feature *att* has a strong presence in e18v (Figure 3; it often, but not exclusively, co-occurs with *vraiment*). Imperative forms are most frequent in the commercial listings (e17x: 23%, Fig. 3). The closing remark typically found on eBay *bonnes enchères* 'happy bidding' appears in the e05p corpus (15%), while commercial ads don't use it (<1% of all listings in e17x). It is less popular in e17p (3%) but again frequently occurs in e18v (11%).

3. Prediction of genre variation

Based on the results in section 2, we tested the predictive power of the features *att*, *enc*, *imp*, *ann*, *bon*, *ego*, *stn* and *pre* for the task of classifying the correct sub-corpus. A classification tree (*tree* package in R; Ripley, 2018) identifies *att*, *bon* and *ann* as the features with the greatest predictive power. Nonetheless, the misclassification error rate reaches a relatively high value of 58.84%. In addition to this, the corpus e17p is not represented in the classification. Due to this high overall error rate, we did not subdivide the data into a training set and a test set, and all ads were used for training the model (Figure 4).

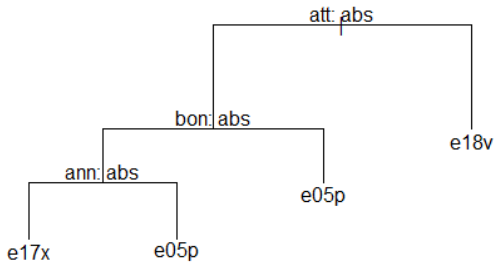

Figure 4: Classification tree predicting corpus target variable. The branches on the left-hand side represent the positive outcome of each test (variable is absent), the right-hand side the negative outcome (variable is present)

The error rate can only slightly be reduced by means of a *randomForest* model (Liaw & Wiener, 2002) (ntrees=450, mtry = 2, Out Of Bag Error = 55,33%).

	Pred:e05p	**Pred:e17p**	**Pred:e17x**	**Pred:e18v**
Ref: e05p	96	2	128	74
Ref: e17p	39	0	191	70
Ref: e17x	5	1	244	50
Ref: e18v	36	22	77	221

Table 3: Confusion matrix of predicted (Pred) and actual (Ref) target features of the randomForest model

As Table 3 shows, for e05p, a misclassification occurred more often than a correct classification – even in the training data. Rules to predict the target corpus e17p did not lead to any correct classification.

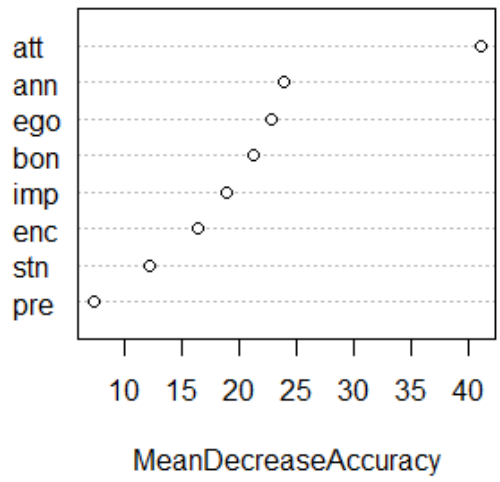

Figure 5: Feature importance

A ranking of the predictor features according to their importance for the classification (Figure 5) reveals a marginal role for *pre*, as its removal would lead to a decrease of the mean accuracy value (calculated by dividing the decrease of accuracy averaged over all trees by the standard deviation) of only 7.37.
The features *imp*, *enc*, *bon*, *ego*, *ann* and *stn* play a moderate role in the classification, whereas the removal of *att* would entail a mean decrease of accuracy of 41.05.

	e05p	e17p	e17x	e18v
ann	17.97	10.01	20.97	-1.68
bon	15.64	-1.95	22.19	2.03
ego	-6.68	11.35	11.17	19.86
stn	9.04	1.05	13.24	-0.51
pre	0.66	10.04	-5.62	5.22
enc	5.34	7.42	18.35	-3.39
imp	0.85	9.81	18.72	10.57
att	15.72	11.77	17.95	39.63

Table 4: Importance values of the predictor features for the corpus target feature

Table 4 shows how important the features are as predictors for the individual sub-corpora. The two forms of evaluative attributes tagged in the corpora – *bon* refers to attributes appearing at the very beginning of an ad, *att* is anywhere else – or the absence thereof, are most important for e17x (*bon*) and e18v (*att*).
The high overall error rate can be partly explained by the shared common genre, which can also be inferred from a k-modes clustering (*klaR* package; Weihs et al., 2005). Here, clusters are created based on similar combinations of the categorical predictor features.
A comparison of these clusters with the target corpus feature (Table 5) shows a clear prevalence of the first cluster, which covers around 83% of the analysed ads. This prevalence shows a high degree of inter-corpus similarity, possibly due to the presence of non-distinctive features. At the same time, the attribution to clusters other than cluster 1 states a considerable intra-corpus variation and thus a non-exclusiveness of corpora and feature combinations.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
e05p	249	4	10	37
e17p	265	2	5	28
e17x	230	2	2	66
e18v	293	29	18	16

Table 5: Cross calculation of the four corpora and the four clusters

4. Conclusion

As early as 1998, Shepherd and Watters state that the key evolutionary aspect in cybergenres "appears to be the functionality afforded by the new medium" (p. 2), and this applies especially for advertising contexts, per se characterised by innovation and variability over time (Gerstenberg 2006). The difference between the distribution of linguistic features after 12 years, from the e05p corpus to its sibling e17p, reflects (micro-)historical dynamics. All three dimensions and their respective linguistic features decrease from 2005 to 2017. Two features show a different pattern, that is, the use of first-person pronouns and evaluative adjectives. Interestingly, the frequency of these features, *ego* and *att*, show similar values in e17p and e17x. First-person pronouns are highly present in all listings, with the highest frequency in e18v. This feature, also considered to be a stance marker (Biber & Egbert, 2016, p. 108) shows that even professional users communicate in a rather personal manner, while they do not use the stance marker *vraiment*, which turns out to be exclusive for a distinct niche of the eBay listings as represented in e18v. In this sub-corpus, another community specific feature has managed to survive over the 13 years, that is, the phrase 'happy bidding', which clearly decreased between the private corpora of 2005 and 2017 (e05p, e17p), and was extremely rarely used in the professional control corpus (e17x). Our control corpus does in fact feature the lowest number of features from dimensions 1 and 2 (classifieds and community, respectively). The presentative constructions are an exception to this, which indicate an informal style, but also an increasing amount of textual description – space restrictions becoming less relevant, especially in e17x and e18v. The 3rd dimension (marketing language) has a fairly homogeneous distribution across the sub-corpora with the only exception being imperative forms, a prototypical feature of professional marketing language. The use of non-standard variants typical for CMC (*stn*) is very rare in all four sections, showing that this genre does display some formal tendencies.
We have seen some developments from 2005, as users move away from traditional 'classifieds language' (see Figure 1), to 2017, where, for some aspects, private and professional listings are aligned. Constructing the corpus using the stance marker *vraiment* in the query turned out to be a valid tool for automatically collecting text types which show persistent features of community language (e18v).
The relatively low predictive power of the features, as shown in chapter 3, emphasises how hybrid and fluid this

genre is. Although a larger dataset would be necessary to make more conclusive remarks, our corpus has shown that the eBay genre has evolved over the last 13 years. A detailed analysis of the distribution of the individual features and of their respective predictive power helps to precisely detect the levels of genre fusion and hybridisation, which is displayed in the use of personal pronouns in professional as well as in private marketing language, and in the professional use of positive attributes in private ads. However, some features remain exclusive for a niche of users; those who wish 'happy bidding'.

5. References

eBay.fr-corpus = Gerstenberg, Annette & Freya Hewett. (2019). A collection of online auction listings from 2005 to 2018 (anonymised). doi.org/10.5281/zenodo.3361253. University of Potsdam: LA-bank.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics*, 44(2), pp. 95–137.

Crowston, K., & Williams, M. (2000). Reproduced and Emergent Genres of Communication on the World Wide Web. *The Information Society*, *16*(3), pp. 201–215.

Danna, S. (2002). Classified Advertising. In J. McDonough, K. Egolf (Eds.), *The Advertising Age Encyclopedia of Advertising*. New York: Routledge.

eBay France. (1995–2019). À l'occasion de ses 15 ans en France, eBay dévoile ses ambitions et annonce des partenariats stratégiques (18 sept. 2015). eBay France: eBay Inc.

Gerstenberg, A. (2006). Geschichte der Sprache der Werbung in der Romania. In G. Ernst, M-D. Gleßgen, C. Schmitt & W. Schweickard (Eds.), *Romanische Sprachgeschichte*: *Ein internationales Handbuch zur Geschichte der romanischen Sprachen*. HSK 23.2. Berlin, New York: de Gruyter, pp. 2161–2175.

Gerstenberg, Annette. (2007). Zur Klassifizierbarkeit französischer Privatwerbung im Internet nach sprachlichen Merkmalen. In M. Döring, D. Osthus & C. Polzin-Haumann (Eds.), *Sprachliche Diversität: Praktiken – Repräsentationen – Identitäten.* Bonn: Romanistischer Verlag, pp. 373–394.

Giltrow, J. (2013). Genre and Computer-Mediated Communication. In S. C. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of Computer-Mediated Communication.* Berlin: Mouton de Gruyter, pp. 717–737.

Janoschka, A. (2004). *Web Advertising*. Amsterdam: Benjamins.

Koch, P., Oesterreicher, W. (2011). *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. Berlin: de Gruyter.

Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

ParseHub. (2015–2019). *A web scraping tool that is easy to use*. Toronto: ParseHub.com.

Ripley, B. (2018). *tree: Classification and Regression Trees*. R-project.org, package tree: R package version 1.0-39.

Shepherd, M., Watters, C. (1998). The Evolution of Cybergenres. *Hawaii International Conference on System Sciences*, *31*(2), pp. 1–13.

Weihs, C., Ligges, U., Luebke, K., Raabe, N. (2005). klaR Analyzing German Business Cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data Analysis and Decision Support.* Berlin: Springer, pp. 335–343.

On downgrading and upgrading strategies used in the act of self-praise in French and US LinkedIN-summaries. A contrastive pragmatic analysis

Els Tobback
University of Antwerp
E-mail: els.tobback@uantwerpen.be.

Abstract

In Politeness Theory, self-praise has traditionally been interpreted as a potentially face threatening act, which infringes the 'Modesty Maxim' proposed by Leech (1983). Certain discourse genres, however, like application letters or job interviews serve, by definition, to promote the professional as skilful. This paper takes up the analysis of self-praise in LinkedIn-summaries written by French and US communication professionals. More specifically, it focuses on the use of upgrading and downgrading pragmatic strategies from a contrastive perspective. On the basis of a corpus of 200 summaries, it shows on the one hand that downgrading is a far less frequently used strategy than upgrading in both corpora. On the other hand, the data show that, overall, US communication professionals are less reluctant than the French in uttering self-praise in a strong, more or less "bragging" way.

Keywords: impression management, self-promotion, self-praise, Politeness theory, LinkedIN

1. Introduction

In every social interaction, individuals are concerned with the perceptions others have on their behalf. In order to influence their image positively, they (intentionally or unintentionally) make use of verbal and non-verbal strategies, which have been described as **self-presentation** (e.g. Barrick et al., 2009) or **impression management** (IM) tactics (e.g. Peeters & Lievens, 2006; Johnson et al., 2016) in social psychology. These strategies, such as 'other enhancements', exemplification, entitlements,… (Peeters & Lievens, 2006), have been studied extensively in organizational settings, and especially in the context of job interviews. In this context they appear to be omnipresent (Johnson et al., 2016), and they have generally been found to have a positive effect on the interviewers' evaluations of the job applicants (e.g. Peeters & Lievens, 2006).

One of the most frequently used IM tactics appears to be **self-promotion**. While its aim is to be considered as competent and capable, self-promotion is realised by "pointing with pride to one's accomplishments, speaking directly about one's strengths and talents, and making internal rather than external attributions for achievements" (Rudman, 1998: 629). Self-promotion has been described as "especially useful in situations in which the self-enhancer is not well-known or is competing against others for scarce resources (e.g. during a job interview)" (Rudman 1998 id). Although the effect of self-promotion on recruiters' decisions has often been described as positive (Kacmar et al., 1992; Stevens & Kristof 1995), some research argue that it is not always effective because self-promotion "violates norms of politeness and humility, and thus is often considered socially inappropriate" (Fragale & Grant, 2015: 63).

Interestingly, also in pragmatics, it is the interpretation of self-promotion as being socially risky that is the starting point for the few pragmatic studies that have been devoted to this topic in the framework of Politeness theory (e.g. Speer, 2012; Dayter 2014, 2016; Matley, 2018). At least according to a traditional view on Politeness, self-promotion, which is in this context rather referred to as "self-praise", has been analysed as a potentially face threatening act (FTA), which infringes one of the Politeness Principles, viz. the Modesty Principle formulated by Leech (1983): "Minimize the expression of praise of self; maximize the expression of dispraise of self".

This led several authors to describe the pragmatic strategies that are used in face-to-face interactions (Speer, 2012) and CMC (Dayter, 2014, 2016; Matley, 2018) to commit this FTA. Among these strategies, we find the classical opposition between **direct** and **indirect** utterances of the FTA, the latter form of self-praise being materialised in the form of a complaint (Dayter 2014, 2016). Next, in the category of explicit self-praise, **unmitigated** self-praise has been opposed to different forms of **modified** self-praise, such as the use of disclaimers (e.g. "I shouldn't compliment myself but […]" (Speer, 2012)) or qualifications (e.g. "even if I do say so myself" (Speer 2012)), reported third-party compliments (Speer, 2012), shifting the credit for accomplishments to a third party, adding an element of self-denigration or making reference to hard work (Dayter 2014, 2016). Unlike the previously mentioned authors, Matley (2018), in his study on Instagram posts, not only describes **mitigation** strategies, aimed at reducing the potentially face threatening character of self-praise, but also **aggravation** strategies, where the face-threat [is] intentionally exacerbated, 'boosted' or maximised in some way" (Matley 2018: 3).

2. Research focus and methodology

This paper focuses on the use of self-praise in a specific form of CMC, viz. the summary of LinkedIN-profiles, from a contrastive (cross-cultural) perspective. More specifically, it brings the analysis of a corpus of some LinkedIn-summaries written by French or US communication professionals: 104 summaries are written in French by authors situated in France and 101 are written in English by authors situated in the US.

These LinkedIN-summaries are obviously quite closely related to discourse during job interviews. Indeed, according to Van Dijck (2013: 208), LinkedIN is "often nicknamed 'facebook in a suit', referring to people's typical job interview attire". In a comparable fashion to job-interviews, LinkedIn-members are generally not

(well-)known by the potential readers and they are equally competing against others for scarce resources (cf. *supra*). Moreover, just like for job interviews, the principle aim of LinkedIn-profiles is to highlight skills and to promote strengths to peers and anonymous evaluators (van Dijck, 2013). Hence, unsurprisingly, self-praise appears to be a central speech act in the LinkedIN-summaries. As a matter of fact, apart from some contact details such as email addresses, LinkedIN-résumés as a whole apply for the definition of self-praise, since the entire text aims to attribute credit to the speaker for being **skilful** as a professional, something which is expected to be positively viewed by the speaker and the potential audience (Dayter, 2014). This also means, theoretically speaking, that this kind of discourse almost by definition infringes the Modesty maxim formulated by Leech (1983). However, just like is the case in job interviews, self-praise in LinkedIN-summaries may probably be expected to be interpreted as less face threatening than in day to day interaction since promoting oneself as a skilful professional is the very essence of the LinkedIn-résumé and therefore socially accepted and expected. Nevertheless, the question as to which forms of self-praise may be interpreted as acceptable (or even positive) or, on the contrary, as more readily face-threatening by the reader, be it a potential client, commercial partner or recruiter, has hardly been tackled in previous research (Tobback, in press)

This study is first of all a continuation of a pilot study (Tobback, in press) devoted to the qualitative description of all the pragmatic (semantic-)strategies used in the LinkedIN-summaries by some 90 communication professionals located in France or the United States. This pilot study revealed that the performance of the act of self-praise in LinkedIn summaries can, just as in other contexts (e.g. Dayter, 2014) and just like other types of FTAs, be achieved explicitly or more indirectly, through the use of 'substitution' processes (cf. Kerbrat-Orecchioni, 1992). It also allowed us to identify a fairly wide range of modifiers ("additive" processes) that either downgrade or upgrade the act of self-praise in LinkedIn summaries.
In this contribution, we will focus exclusively on the quantitative analysis of these "additive" processes used in the two corpuses, the objective being to examine to what extent the use of downgrading or upgrading strategies differs depending on the country of origin (the culture) of the author.

3. Results

Although French and US communication professionals do not present completely opposed behaviors with regard to the use of downgrading (3.1) and upgrading (3.2) strategies, the data overall reveal that US communication professionals are somehow less reluctant than French in uttering self-praise in a strong, more or less "bragging" way.

3.1 Downgrading strategies

Downgraders are all kinds of syntactic, lexical and phrasal devices which "tone down the impact an utterance is likely to have on the hearer" (Trosborg 1995: 209). Typical examples are past/conditional verb forms, downtoners, understaters or hedges (cf. Trosborg 1995).
Overall, both corpora contain a very low number of downgrading strategies. However, in the US corpus, this number is even (significantly ($p < 0.001$, chi2= 19.8) lower than in the French corpus, as shown in table 1:

	France		US	
Presence of Downgrader	51	7%	23	2%
No Downgrader	696	93%	935	98%
Tot	747[1]	100%	958	100%

Table 1 – Downgrading strategies

Among the mitigating strategies, we may mention a few propositional and speech act hedges (Fraser 2010), and quantifying modifiers. Another type of mitigating strategy concerns the use of verbs that do not stress the actual possession of skills but rather the (still ongoing) process or their development and even the origin of the acquired skill (e.g. 'It is also within UNI that I started to develop my skills in communication strategy' (French corpus).

3.2 Upgrading strategies

Following Trosborg (1995), 'upgraders' have been taken as all elements that are likely to strengthen the impact of the speech act on the receiver. More specifically, in the case of the self-praise speech acts expressed in the LinkedIn-summaries, all elements that do not just neutrally/objectively mention or downgrade one of the core elements such as a quality/skill or a concrete work experience…. have been analysed as upgraders. These upgraders may be either quantifying or qualifying elements (cf. Tobback, in press). Furthermore, we distinguished between three levels on a gradability scale, extending between a low degree and a high degree of upgrading, thereby applying loosely Martin & White's (2005) gradability scale for attitudinal meanings, but both to quantifying and qualifying upgraders.

3.2.1. Overall presence of upgraders

In contrast with the downgraders, the part of corpus extracts containing one or more upgraders appears to be much higher (table 2). However, here again, a significant difference appears between the French and the US corpus, the latter containing far more (56%) extracts with upgraders than the former (37%):

	France		USA	
Presence of upgrader	275	**37%**	534	**56%**
No upgrader	472	63%	424	44%

[1] The absolute figures correspond to the numbers of excerpts taken into account for the analysis. These excerpts result from the first coding of the corpus, based on the type of pragmatic strategy (different types of direct and indirect strategies) used by the author to convey the impression of skillfulnesss (Tobback, in press).

Tot	747	100%	958	100%

Table 2 – upgrading strategies (Chi2 = 60.301 ; p < 0.001)

3.2.2. Qualifying vs quantifying upgraders

In view of table 3, French and American authors seem to have different preferences when it comes to the way they strengthen the expression of their skilfulness. Indeed, whereas French authors clearly favour quantitative upgraders (65% of cases), American authors use qualifying upgraders much more often (57%), these differences appearing to be highly significant (p < 0.001; chi2= 49.9).

	FR		US	
# qualifying upgraders	133	35%	478	57%
# quantifying upgraders	243	65%	356	43%
Tot	376	100%	834	100%

Table 3 – Quantifying vs qualifying upgraders

If we admit that qualitative expressions give a more subjective flavour to the description of oneself, we might say that they appear as more "bragging-like" and possibly have a stronger "self-promoting" effect. By contrast, quantitative expressions, especially high numbers, long periods, etc. – combined with the relevant nouns – appear as a more objective way, based on "facts and figures", so to speak, of highlighting one's skills. As such, quantitative modifiers turn out to be less direct boasters, since they are based on the implicature 'more is better'.

3.2.3. Strength of the upgraders

In order to further refine the comparison of the French and American corpuses, we have tried to distinguish three positions on a gradability scale: upgraders with "low", "medium" and "high" intensity. Low-intensity upgraders refer to elements (usually adjectives or adverbs) that simply add a qualifying or quantifying element to the basic information, without themselves being reinforced (e.g. ***results driven*** *English language training*). Medium-intensity upgraders are upgraders that are themselves modified by an upgraders or that convey a stronger sense compared to low intensity upgraders (e.g. ***very*** *active vs active).* High intensity upgraders roughly correspond to the category of "maximizers" (cf. Martin & White 2005: 142) (e.g. ***excellent*** *writer, working* ***extremely*** *well*).

In this case, a contrast was observed between the qualifying upgraders, on the one hand, and the quantifying upgraders, on the other hand. Indeed, with respect to the qualifying upgraders, both French and US authors are found to use in about half of the cases low-intensity upgraders, about one-third medium-intensity upgraders, and about 20% high-intensity upgraders. Given these results, we therefore note that if French authors use qualifying upgraders - which they do much less often than their American colleagues - they do so on average with the same degree of intensity as American authors.

In the case of the quantifying upgraders, the previously observed contrasts between the French and the US corpus show up again (table 4). More specifically, it can be seen that the 'cultural' differences mainly appear in the use of low and medium intensity upgraders. French authors use low-intensity upgraders in more than half of the cases (54%), compared with 38% for American authors, who use medium-intensity upgraders more frequently (44% of cases, compared with 31% in the FR corpus). On the other hand, few differences are noticeable for the higher-intensity upgraders, with percentages of 15% in the French corpus and 18% in the American corpus.

	FR		USA	
Low	110	54%	116	38%
Medium	62	31%	135	44%
High	31	15%	56	18%
Tot[2]	**203**	100%	**307**	100%

Table 4 – strength of quantifying upgraders

(Chi2= 13.76; p= 0.001)

4. References

Barrick MR, Shaffer JA and DeGrassi SW (2009). What You See May Not Be What You Get: Relationships Among Self-Presentation Tactics and Ratings of Interview and Job Performance. *Journal of Applied Psychology*, Vol. 94, No. 6, 1394–1411.

Dayter D (2014). Self-praise in microblogging. *Journal of Pragmatics* 62: 91-102.

Dayter D (2016). *Discursive Self in Microblogging: speech acts, stories and self-praise*. John Benjamins Publishing Company.

Fragale AR an Grant AM (2015). Busy brains, boasters' gains: Self-promotion effectiveness depends on audiences cognitive resources. *Journal of Experimental Social Psychology* 58: 63–76.

Fraser B (2010). Pragmatic competence: The case of hedging, in: Kaltenböck Gunther, Wiltrud Mihatsch and Stefan Schneider (eds) *New approaches to hedging*. Emerald Group Publishing Limited, pp. 15-34.

Johnson G, Griffith J and Buckley R (2016). A new model of impression management: Emotions in the 'black box' of organizational persuasion. *Journal of Occupational and Organizational Psychology* 89: 111–140.

Kacmar M, Delery J and G Ferris (1992); Differential effectiveness of applicant impression management tactics on employment interview decisions. *Journal of Applied Social Psychology* 22: 1250–1272.

Kerbrat-Orecchioni C (2005). *Le discours en interaction*. Paris: Armand Colin.

Leech G (1983); *Principles of pragmatics*. Longman, London.

Leech G (2014). *The pragmatics of Politeness.* Oxford: Oxford University Press.

Martin J and White PRR (2005). *The Language of Evaluation.* NewYork: Palgrave Macmillan.

Matley D (2018). "This is NOT a# humblebrag, this is just a# brag": The pragmatics of self-praise, hashtags and politeness in Instagram posts. *Discourse, context &*

[2] Since quantifying upgarders are in the vast majority of cases identified in the extracts where the author expresses his competence indirectly (AUTHORS, in revision), the presentation of the results has been limited to these 'indirect strategies'.

media 22: 30-38.

Peeters H and Lievens F (2006) Verbal and nonverbal impression management tactics in behavior description and situational interviews. *International Journal of Selection and Assessment* 14(3): 206-222.

Rudman-Rutgers L (1998) Self-Promotion as a Risk Factor for Women: The Costs and Benefits of Counterstereotypical Impression Management. *Journal of Personality and Social Psychology* 74(3): 629-645.

Speer SA (2012). The interactional organization of self-praise: epistemics, preference organization, and implications for identity research. *Social Psychology Quarterly* 75(1): 52-79.

Stevens CK and Kristof AL (1995) Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology* 80: 587–606.

Tobback E (in press) Telling the world how skilful you are: self-praise strategies on LinkedIn, *Discourse and Communication.*

Van Dijck J (2013) 'You have one identity': performing the self on Facebook and LinkedIn. *Media, Culture & Society 35*(2): 199-215.

Mapping the itineraries and interests of internet users with the RedditGender corpus

Marie Flesch
ATILF, Université de Lorraine - CNRS
E-mail: marie.flesch@univ-lorraine.fr

Abstract

This quantitative corpus study aims at mapping the itineraries and interests of internet users on the American community website Reddit. It is based on RedditGender, a 19 million-word corpus that includes comments posted by 1,044 cisgender and transgender Redditors. A list of the forums in which each Reddit user left comments was extracted from the corpus. Forums were then classified by topic. Statistical analyses were performed in order to gauge the mobility of Redditors on the site, the length of their comments, and their interests. Preliminary results reveal significant gender differences in interest and use of Reddit. They provide insight in the way transgender and cisgender internet users inhabit the virtual space. They can also inform linguistic studies of gender and CMC, and help better understand differences of usage between genders.

Keywords: gender, CMC, Reddit, corpus linguistics, quantitative linguistics, sociolinguistics

1. Introduction

Social media and community websites offer researchers an unprecedented opportunity to study gender differences and similarities in interests (Thelwall & Stuart, 2018). Offline, researchers have found that women tend to discuss personal relationships more than men, in studies of overheard conversations (Bischoping, 1993; Dunbar, Marriott, & Duncan, 1997), and surveys (Sehulster, 2006). This trend seems to also exist online, with females writing more about social processes and home than males, who talk more about work, sports, politics and religion (Schwartz, Eichstaedt, Kern, et al., 2013; Wang, Burke & Kraut, 2013; Thelwall & Stuart, 2018).

Several methods have been used to identify internet users' interests: word frequency analysis (Schwartz, Eichstaedt, Kern, et al., 2013; Thelwall & Stuart, 2018), analysis of hashtags and usernames (Holmberg & Hellsten, 2015), and content analysis (Evans, 2016). On Reddit, analysis of participation rates of males and females in the 100 most popular forums was also conducted (Thelwall & Stuart, 2018). It showed, for instance, that male participation rates are higher in humor, gaming, news and politics forums.

This study aims at investigating internet users' interests on Reddit from a different perspective. It uses a corpus built from the comment history of 1,044 Redditors to try to map their interests through the forums they commented on, and to understand how they occupy the virtual space. It attempts to go beyond the gender binary which dominates the way gender has been studied (Eckert, 2014) by taking into account contributions by cisgender males and females, whose gender identity matches their birth sex, but also by transgender and non-binary people. Transgender males and females' gender identity does not correspond with the gender they were assigned at birth. Non-binary people identify outside of the gender binary. The term "non-binary" encompasses different gender identities, including genderfluid, agender, bigender or demigender.

2. Reddit as a geek and masculine space

Reddit is an American community-based website. Founded in 2005, it is now the 5th most popular website in the United States (Top Sites in United States, n.d). Reddit comprises of more than a million forums or "subreddits", moderated by volunteers. The two-thirds of its user base is male (Barthel et al., 2016), and the site has been found to be a "center of geek culture" and a "hub for anti-feminist activism" (Massanari, 2017). In 2014, it greatly contributed to the "The Fappening", a leak of intimate photographs of stars, and was the scene of Gamergate, a campaign of harassment of female and minority game developers and journalists (Massanari, 2017). It is also one of the birthplaces of the misogynistic Incel movement (Beauchamp, 2019). In spite of this, Reddit has been described as one of the best online resources for transgender people (Woodstock, 2018).

3. The RedditGender corpus

The RedditGender corpus was compiled in 2017 by the author in order to conduct a sociolinguistic study of gender and CMC. It will be made available to researchers after her PhD defense. The corpus uses Redditors' comment histories as its building blocks. Internet users who have a Reddit account can freely post comments in a discussion thread, or initiate a thread by writing a text ("self-post") or posting a link to an article or a video from another site. The RedditGender corpus only contains comments posted within discussion threads, and no self-posts. It features 460,533 comments posted on 7,937 forums by 1,044 Redditors: 372 cisgender males, 372 cisgender females, 100 transgender males, 100 transgender females, and 100 non-binary people. Gender was not inferred from usernames, but was obtained through Redditors' self-declarations such as "I'm a 40s hetero male in middle management" or "I'm a young 20s black woman in Washington". The corpus is encoded in XML-TEI TXM

and can be queried with the TXM software.

4. Classification of forums

A list of the forums Redditors' commented on was generated by a research engineer using Python, with the number of comments posted by each person on each subreddit. The 7,937 forums were classified into 12 categories, some of which are based on Thelwall and Stuart's (2018):

1. **General interest** (1275 subreddits). General questions and topics, images, animals, places. Examples: r/Advice, r/AskReddit, r/CasualConversation, r/ImagesofCalifornia, r/motivation, r/AskNYC.
2. **Humor** (476 subreddits). Memes, humor and "circlejerks" subreddits. Examples: r/oldpeoplefacebook, r/shittyadvice, r/WastedGifs, r/Hiphopcirclejerk, r/CrappyDesign, r/EbayWTF.
3. **Gaming** (869 subreddits). Examples: r/FreeGamesOnSteam, r/gamedev, r/gaming, r/gtaonline, r/hearthstone.
4. **News**, **politics and religion** (383 subreddits). Examples: r/hillaryclinton, r/immigration, r/MarchAgainstTrump, r/NewPatriotism, r/PhilosophyofReligion, r/syriancivilwar.
5. **Education and science** (681 subreddits). Examples: r/TranslationStudies, r/Archeology, r/math, r/AskLiteraryStudies, r/college, r/Harvard.
6. **Technology** (419 subreddits). Examples: r/microsoft, r/Nexus7, r/Python, r/smarthome, r/TechnologyProTips, r/SQL.
7. **Mass entertainment** (1124 subreddits). Music, radio, TV, movies, books. Examples: r/Concerts, r/KingCrimsons, r/NPR, r/visualnovels, r/disney, r/criminalminds.
8. **Hobbies** (663 subreddits). Hobbies, food and shopping. Examples: r/DIY, r/Fiat, r/FoodPorn, r/gardening, r/Gin.
9. **Personal advice** (793 subreddits). Beauty, fashion, relationships, health. Examples: r/interracialdating, r/LatinaBeauties, r/relationships, r/lonely, r/LushCosmetics, r/malehairadvice.
10. **Sports and fitness** (328 subreddits). Examples: r/FitToFat, r/Gymnastics, r/olympics, r/SFGiants, r/xxketo.
11. **X-rated** (499 subreddits). Porn, violence, guns and drugs. Examples: malegonewild, r/MorbidInterests, r/progun, r/SexyHalfAsians, r/Drugs.
12. **Others** (236 subreddits). Subreddits that did not fit in any of the above categories. Examples: r/FuckChuck, r/ZombiesSurvivalTactics, r/ButtonAftermath, r/GTAorRussia.

5. Methods

Statistical analyses were performed with the R software in order to try to answer several questions: who comments on the more forums, or, in other words, who moves the most through the site? Who writes the longest comments? To what extent are interests gendered on Reddit? Our hypotheses were that cisgender males would be the most mobile, given that Reddit is a male-dominated platform, that transgender Redditors would be the least mobile, because they are a minority on Reddit, and that women would write longer messages than men, as it was found by Thelwall and Stuart (2018).

We also hypothesized that "traditional" gender interests would be reflected in the analysis of thematic categories, with, for example, cisgender females gravitating towards personal advice subreddits and cisgender males commenting more on gaming, news and politics, sports and x-rated forums. Several datasets were used to perform theses analyses:

1. Mean number of comments posted on each forum Redditors commented on
2. Mean length of comments
3. Number of comments posted on three "big" subreddits, per 1,000 words
4. Percentage of comments posted on subreddits pertaining to each thematic category

We tested our hypotheses with regression models. All the data collected are count-based, are very skewed and display a significant amount of variation. For datasets 1 and 2, we used negative binomial regression, with allows for greater variability than models based on the Poisson distribution (Hilbe, 2014). Since datasets 3 and 4 contain large numbers of zero counts, we created zero-inflated regression models with the `zeroinfl` function of the `pscl` R package (Zeileis, Kleiber, & Jackman, 2008; Jackman, 2017). Zero-inflated models are mixture models which give two sets of coefficients, one modelling the probability of a zero count, and the other the probability of counts higher than zero (Faraway, 2016).

Zero-inflated negative binomial models were compared to zero-inflated Poisson models with the boundary likelihood ratio test, using the `lrtest` function from the `lmtest` R package (Zeileis & Hothorn, 2002). Results showed that in all cases zero-inflated negative binomial models were preferable to zero-inflated Poisson models.

6. Results: Number of comments per forum and length of comments

Results reveal significant gendered differences in number of comments per forum. Descriptive statistics are shown in Figure 1. Negative binomial regression was used to model the number of comments written by Redditor per forum. Transgender males ("ftm") and non-binary Redditors were found to be the least "mobile" on the site. They write more comments per forum than other categories, thus frequenting a smaller number of forums. Transgender females ("mtf"), on the other hand, comment on the most forums.

Significant differences between subcorpora were also found in comment length. Mean comment length was computed for each user and subcorpora (Figure 2), and

negative binomial regression was again used. No significant difference was found between transgender and cisgender females. According to the model, cisgender males wrote the shortest comments, while transgender males and non-binary Redditors wrote the longest.

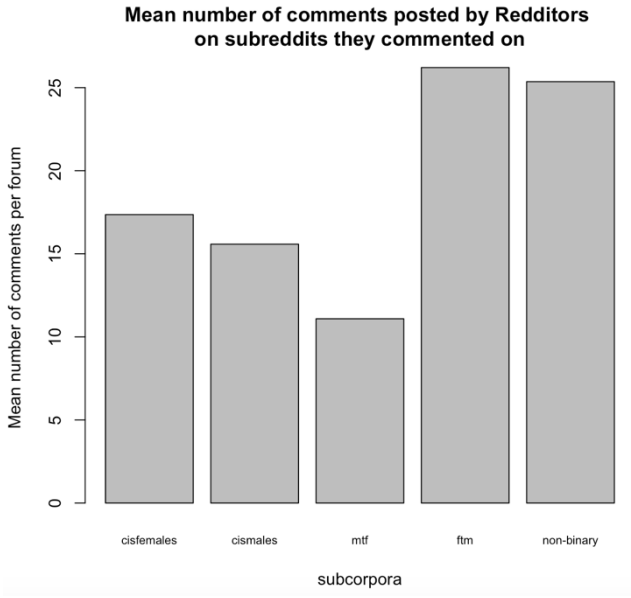

Figure 1: Bar chart representing the mean number of comments posted by Redditors on subreddits they commented on in RedditGender, per gendered subcorpus

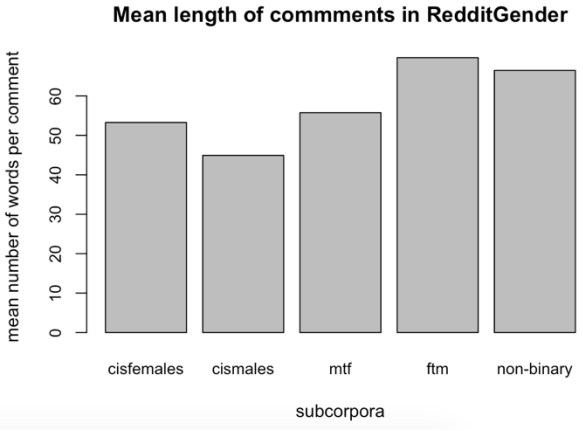

Figure 2: Bar chart representing the mean length of comments in RedditGender, per gendered subcorpora

7. Results: Analysis of three subreddits

Analyses performed on individual forums reveal distinct patterns of use. Three of the most popular subreddits in RedditGender were examined: r/AskReddit, a forum where users can ask questions on various topics, r/politics, which is dedicated to the discussion of politics, and r/relationships, which deals with personal relationships. Zero-inflated negative binomial regression was performed, with the mean number of comments posted on each site per 1,000 words as the dependent variable, and gender and age of Redditors as the predictors. Results show that cisgender males are the most likely to comment on r/AskReddit and r/politics, but that among Redditors who comment on these two forums, cisgender females are the most frequent posters. Older Redditors comment more on r/politics than younger Redditors. Cisgender females are the most likely to comment on r/relationships. Non-binary Redditors and transgender males and females are the least likely to post comments on the three subreddits.

8. Results: Thematic interests of Redditors

Descriptive statistics, per gendered subcorpus, are shown in Table 1. Zero-inflated negative binomial regression was performed on each thematic category. The main findings are summarized below:

1. **General interest.** Cisgender males are as likely to comment as cisgender females. Transgender people are the least likely to write comments on these subreddits. Among those who comment on general interest subreddits, cisgender males are those who post the most comments.
2. **Humor**. Cisgender males are the most likely to participate in humor subreddits. However, when they comment on these forums, transgender females, cisgender females and non-binary individuals comment more than cisgender males.
3. **Gaming.** Cisgender males and transgender females are the most likely participants to gaming subreddits. Transgender males and cisgender females are respectively 7.83 and 10.37 less likely to comment on these subreddits than cisgender males. There is no significant difference between Redditors who comment on gaming subreddits.
4. **News and politics.** Cisgender males are far more likely to comment on news and politics forums than other gender group. Among those who comment on political and news subreddits, cisgender females are those who post the least.
5. **Education and science.** Cisgender males are the most likely to post comments. Transgender people are less likely to participate than cisgender females. Among Redditors who write comments on education and science forums, non-binary individuals are the least prolific contributors. There is no significant difference between the other groups.
6. **Technology.** The model shows that cisgender males are the most likely to post comments. A transgender male is 11.78 times less likely than a cisgender male to participate in technology subreddits. Among Redditors who participate on tech related forums, cisgender females and transgender males are the least frequent contributors.
7. **Mass entertainment.** Cisgender males and transgender females are the most likely to write comments on music, movies, books or radio subreddits. When they do participate to these

forums, cisgender females post comments at the same frequency as transgender females and cisgender males. Non-binary Redditors and transgender males are the least frequent posters.

8. **Hobbies**. Transgender males and non-binary Redditors are the least likely to post comments. Transgender males are also, among Redditors who participate in hobbies related discussions, the least frequent contributors.
9. **Personal advice**. For this category, we used simple negative-binomial regression instead of a zero-inflated model, because there were only 22 zero observations. This is due to the fact that most Redditors of RedditGender were found in personal advice forums, generally gender-related, such as r/AskMen or r/asktransgender. The negative binomial model shows that cisgender males are the least frequent posters, and that transgender males and non-binary people are the most frequent contributors. They post messages on personal advice forums at a rate respectively 2.86 and 2.73 times that of cisgender males.
10. **Sports and fitness**. Cisgender males are the most likely to leave comments on these forums, and the most frequent posters. Non-binary individuals are the least likely to post, and when they do, they do it less often than the other groups.
11. **X-rated**. Cisgender males and transgender females are the most likely to write comments. Among Redditors who participate to porn, drugs, guns and violence subreddits, cisgender females are those who comment the least. There is no significant difference between the other groups.
12. **Others.** Cisgender males and transgender females are the most likely to comment. Non-binary people who comment on these subreddits are the least frequent posters, with no significant difference between the other groups.

	m	f	mtf	ftm	nb
General interest	26.7	22.2	21.5	13.1	8.6
Humor	5.3	5.7	7.8	4.1	8.7
Gaming	7.9	3.5	6.1	3.2	6.5
News & politics	7.4	3.6	6.3	2.6	3.7
Science & education	5.0	4.1	3.6	2.6	2.4
Technology	3.3	0.8	2.2	0.5	1.1
Mass entertainment	6.6	6.5	7.8	2.6	1.9
Hobbies	4.8	4.6	3.3	0.8	1.2
Personal advice	23.4	44.3	37.0	67.1	63.9
Sports & fitness	5.8	3.0	1.6	1.8	0.7
X-rated	2.4	0.5	1.6	0.6	1.0
Others	1.4	1.2	1.2	1.1	0.3
Total	100	100	100	100	100

Table 1 : Percentage of comments posted on each categories per gendered subcorpus (cis-males, cis-females, trans-females, trans-males and non-binary individuals)

9. Discussion

Our analysis seems to confirm the status of Reddit as a male space. As we hypothesized and as Thelwall & Stuart (2018) showed, male Redditors write shorter comments and are the most likely to post comments on subreddits dealing with "traditional" male interests, such as gaming, news and politics, science and education, technology, porn, violence and drugs, as well as humor and mass entertainment. On the other hand, they participate much less in personal advice and hobbies forums than other Redditors.

The zero-inflated models showed that, in some cases, there seem be a "barrier of entry" in these subreddits; when they overcome it, some or all other gender groups participate as much as or even more than cisgender males. This is the case, for example, for gaming subreddits, where no significant difference was detected between groups among the Redditors who posted comments. Analysis of the r/politics and r/AskReddit subreddits and of education and science, mass entertainment, humor, and "x-rated" subreddits also show that males are not always the most frequent contributors among Redditors who comment – they are only the most likely to write at least one comment on these forums. "Traditional" feminine gender interests were reflected in the way cisgender females participate in r/relationships and personal advice subreddits.

Our analysis also reveals interesting and maybe unexpected patterns. Transgender males and non-binary Redditors often behave in the same way. They post long comments on a limited number of subreddits, and are the more frequent posters on personal advice forums. They are the least prolific posters to forums about general interest, education and science and mass entertainment. Transgender females, on the other hand, seem to inhabit the Reddit space in a different way. They comment on more subreddits, are as likely as cisgender males to comment on "x-rated", mass entertainment, and gaming subreddits, and comment on general interests subreddits as much as cisgender males.

It has been shown that the cyberspace offers transgender people the freedom to live their "real" identity, when they have to hide it offline, or suffer from marginalization and bullying (Whittle, 1998; Marciano, 2014). The offline world offer different challenges to transgender females than to transgender males: they do not "pass" as well as transgender males (meaning that they are often not correctly perceived as women by others), are more isolated than transgender males, and hide their gender identity more often (Rankin & Beemyn, 2012). Transgender females then could use Reddit as a "safe" space, where they can be who they really are, regardless of what they are talking about or to whom they are talking. We also cannot exclude the possibility that non-binary and transgender males are more likely to create several Reddit accounts to talk about different topics, without the fear of being "outed", which would explain their seemingly limited mobility on the website.

10. Conclusion

This study of gender differences in online interests shows significant differences in how male, female and non-binary Redditors use the geek and male-dominated community website. It provides insight into the way internet users inhabit the virtual space. It can also be used to understand linguistic differences between genders online, by providing context and theme of discussion as a variable for quantitative studies.

11. 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 Pew Research Center's Journalism Project website: http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Beauchamp, Z. (2019, April 16). The rise of incels: How a support group for the dateless became a violent internet subculture. Retrieved May 12, 2019, from Vox website: https://www.vox.com/the-highlight/2019/4/16/18287446/incel-definition-reddit

Bischoping, K. (1993). Gender differences in conversation topics, 1922-1990. *Sex Roles, 28*(1–2), 1–18. https://doi.org/10.1007/BF00289744

Dunbar, R. I. M., Marriott, A., & Duncan, N. D. C. (1997). Human conversational behavior. *Human Nature, 8*(3), 231–246. https://doi.org/10.1007/BF02912493

Eckert, P. (2014). The Problem with Binaries: Coding for Gender and Sexuality. *Language and Linguistics Compass, 8*(11), 529–535.

Evans, H. (2016). Do women only talk about "female issues"? Gender and issue discussion on Twitter. *Online Information Review, 40*(5), 660–672. https://doi.org/10.1108/OIR-10-2015-0338

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Heiden, Serge, Magué, Jean-Philippe, & Pincemin, Bénédicte. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010) (Vol. 2, p. 1021–1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. Online.

Holmberg, K., & Hellsten, I. (2015). Gender differences in the climate change communication on Twitter. *Internet Research, 25*(5), 811–828. https://doi.org/10.1108/IntR-07-2014-0179

Jackman, S. (2017). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.2. URL https://github.com/atahk/pscl/

Marciano, Avi 2014. Living the VirtuReal: Negotiating Transgender Identity in Cyberspace. *Journal of Computer-Mediated Communication 19*(4): 824–838.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329–346.

Rankin, S., & Beemyn, G. (2012). Beyond a binary: The lives of gender-nonconforming youth. *About Campus, 17*(4), 2–10. https://doi.org/10.1002/abc.21086

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Sehulster, J. R. (2006). Things we talk about, how frequently, and to whom: Frequency of topics in everyday conversation as a function of gender, age, and marital status. *The American Journal of Psychology, 119*(3), 407–432. https://doi.org/10.2307/20445351

Thelwall, M., & Stuart, E. (2018). She's Reddit: A source of statistically significant gendered interest information? ArXiv:1810.08091 [Cs]. Retrieved from http://arxiv.org/abs/1810.08091

Top Sites in United States - Alexa. (n.d.). Retrieved May 13, 2019, from https://www.alexa.com/topsites/countries/US

Wang, Y. C., Burke, M., & Kraut, R. E. (2013). Gender, topic, and audience response: an analysis of user-generated content on Facebook. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 31-34). New York, NY: ACM Press.

Whittle, S. (1998). The trans-cyberian mail way. *Social & Legal Studies, 7*(3), 389-408.

Woodstock, M. (2018, February 5). Gender Reveal [Audio podcast]. Retrieved from https://gender.libsyn.com/episode-6-gender-102-with-z-griffler.

Zeileis, A. & Hothorn T. (2002). Diagnostic Checking in Regression Relationships. *R News 2*(3), 7-10. https://CRAN.R-project.org/doc/Rnews/ http://www.jstatsoft.org/v27/i08/.

Zeileis, A., Kleiber, C., and Jackman, S.. (2008). *Regression models for count data in R. Journal of Statistical Software 27*(8).

A Contrastive Analysis of E-mail Requests in Chinese and French

Ting-Shiu Lin, Chia-Ling Hsieh
Université Lumière-Lyon 2, National Taiwan Normal University
tingshiu.lin@gmail.com, clhsieh@ntnu.edu.tw

Abstract

This paper examines Chinese and French e-mail requests and argues that the individualist/collectivist model, often used to account for cultural differences in face-to-face interaction, is not sufficient to explain the differences and similarities between Chinese and French e-mail communication. This traditional model predicts Chinese requests to be less direct and more sensitive to interpersonal factors. However, the present study found that Chinese e-mail requests are not more indirect than the French ones concerning strategy types and information sequencing, and both groups are equally sensitive to interpersonal variants in the choice of requestive strategies. It is thus concluded that e-mail, as a medium of communication, has developed its own interaction patterns that are different from those of face-to-face conversation.

Keywords: request, speech act, e-mail communication

1. Introduction

In studies on culture and communication, it is generally held that collectivist and individualist societies prefer different styles of communication: in comparison to speakers coming from individualist cultures, such as English and French, members of collectivist societies, such as Chinese and Japanese, tend to be less direct in communication and have a higher tendency to adjust their communication strategies according to their relationships with addressees (Fukushima, 2000; Gelfand et al., 2004; Holtgraves & Yang, 1992; Liu, 2003; Triandis, 1995).

However, the above-mentioned generalization is mainly based on observation of interaction patterns in face-to-face situations; to what extent it may be applied towards computer-mediated communications remains open to debate. Some studies suggest that the traditional individualist/collectivist model is able to predict communicative behaviors in collaborative editing on Wikipedia (Pfeil et al., 2006) and in the website design of certain enterprises (Chang, 2011), while others argue that this model is insufficient in depicting interaction patterns on social networking sites (Chang & Tseng, 2009; Hsieh, 2011), demonstrating that in addition to cultural factors, the medium of communication also plays an important role. The applicability of the individualist/collectivist model to communication via computer-mediated platforms thus warrants further research.

The present study examines e-mail requests composed in Chinese and French in order to understand whether the traditional individualist/collectivist model is sufficiently accurate for predicting interaction within the bounds of e-mail mediated communication. Specifically, the following questions will be addressed:

(1) Directness of communication:

(1-1) Do Chinese participants adopt more indirect strategies than their French counterparts in request e-mails?

(1-2) Do Chinese participants adopt a more indirect schema than their French counterparts in the information sequencing of request e-mails?

(2) Sensitivity to interpersonal variables:

Do Chinese participants exhibit higher sensitivity to interpersonal factors when selecting request strategies than their French counterparts?

2. Methodology

2.1 Material

Data from this study was collected through an online questionnaire conducted between December 12, 2011 and April 3, 2012. Given that previous studies inspecting authentic e-mail data encountered difficulty in controlling non-relevant factors, the present study adopted the questionnaire method in order to better control potential variables. For example, Aslan (2017) collected realistic request e-mails exchanged in an academic context, but could hardly control the interpersonal variables or imposition levels of the requested tasks. Similarly, in Biesenbach-Lucas' (2007) analysis of English e-mails written by non-native speakers, the data size was not extensive enough for the researcher to distinguish between the different language backgrounds of the participants. In view of the aforementioned, this study believes that the questionnaire method provides a better means to construct a database in which interpersonal variables, degrees of imposition, and the cultural/ linguistic backgrounds of the participants can be strictly controlled.

The questionnaire consists of two parts: general information concerning participants and e-mail composition. In accordance with our research goals, these e-mails must be addressed to recipients differing in social distance and relative power in regard to the sender, with all other factors being equal. Based on these criteria, this study incorporated four request scenarios into the questionnaire: requesting a close friend to correct homework, requesting an acquaintance to correct homework, requesting a recommendation letter from a familiar teacher, and requesting a recommendation letter from an unfamiliar teacher. The tasks requested in each of these scenarios were similar in degree of imposition—addressees had little obligation to offer assistance; moreover task completion would require some amount of time and effort on their behalf. In addition, all participants were asked to write to

same-sex recipients that they knew in real life, thus improving data authenticity and minimizing the potential effects caused by gender differences.

2.2 Subjects

51 Taiwanese and 71 French university students, living respectively in Taipei and in Paris, participated in the study. Participants responded to the questionnaire either at home via personal computer, in a library, or in the office of the authors. The questionnaire was completed by clicking on the link provided by the authors, composing the four e-mails, and submitting the information to Google Drive.

Among responses provided by the participants, 43 from the Chinese group (21 males, 22 females) and 47 from the French group (20 males, 27 females), all of which were written by participants aged 18-32, were counted as valid. We then randomly selected 20 responses given by participants of each gender from each language group, with all responses consisting of four e-mails corresponding to the four request scenarios. In the end, a total of 160 Chinese and 160 French e-mails were considered in our analysis.

2.3 Analysis

The content of the e-mails examined during the course of our study can be divided into three parts: subject line, letter format, and main content. The subject line indicates the theme of the e-mail and is the first thing that an addressee sees upon checking their e-mail. Examples from our study include *correction of a text*, *recommendation letter*, etc. The letter format refers to formulaic expressions found in traditional written letters, such as the salutation, complimentary closing, and signature. Finally, the main content constitutes the core of the e-mail, including the head acts and supportive moves within a request. In this paper, only request head acts, supportive moves, and information sequencing in the main content are to be analyzed; subject lines and letter formats will not be discussed. Classifications of the head acts and supportive moves appearing in our data are presented in section 3.

After classifying all strategies and counting the frequency of their occurrences, we applied Pearson's chi-squared test to determine whether language group exerts significant influence on the choice of head act strategies, supportive moves, and information sequencing in the e-mail request data. In addition, we adopted a generalized linear mixed model in order to investigate whether relative power and social distance have significant effects on the choice of head act strategies and supportive moves in each language group, with participants and strategy types considered as random effects. This model was also used to test the differences in degree of influence on strategic choice brought about by relative power and social distance in the two language groups.

3. Results

3.1 Head acts

3.1.1. Head acts: coding

For the purposes of this study, a head act is defined as the sentence in which the need for the addressee's help on a task, which is or has been clearly described, is expressed. Tasks demanded must be related to those outlined in our scenarios, including writing a recommendation letter or correcting homework.

Following Blum-Kulka et al. (1989a, 1989b), Lin (2009), and Van Mulken (1996), we classified the head acts according to their linguistic forms and grouped them into three categories according to their directness levels:

I. Direct strategies include imperatives (*Help me take a look at my final paper*), explicit performatives (*I beg you to write a recommendation letter for me*), hedged performatives (*I would like to ask you to take a look at my article*), and want statement (*I need you to write a recommendation letter for me*).

II. Conventionally indirect strategies include suggestory formulae (*I was thinking that maybe you could reread what I did*), query preparatory (ability) (*I was wondering if you could write this recommendation letter for me*), query preparatory (willingness) (*Would you accept to reread my work*), query preparatory (permission) (*May I bother you to take some time and take a look at it for me*), query preparatory (feasibility) (*Is it possible for you to help me correct it a little bit*), and query preparatory (others).

III. Non-conventionally indirect strategies refer to hints, with which the intention of the request is interpretable only within the given context. In our data, hints include the cases where an addresser does not directly ask for help in the text of the e-mail, but rather describes how they are in need of assistance, requesting to meet with the addressee, and implying that some sort of service may be required during the course of the meeting.

3.1.2. Head acts: analysis results

Examined first is the influence of language group on level of directness and selection of head acts. For the purpose of calculating language group influence on directness level, we combine counts of conventionally indirect strategies and non-conventionally indirect strategies into a single indirect strategies category, with the intention of improving the reliability of the chi-squared test. This is due to the fact that non-conventionally indirect strategies are low in frequency, and would serve to induce error should they be calculated separately (Table 1). In addition, imperatives, query preparatory (permission) and hints are not included in the comparison, since they give frequency values that are either zero or near to zero (Table 2).

The results of chi-squared tests indicate that the Chinese and French language groups differ significantly in both directness level ($x^2 = 50.4343$, $p < .001$), as well as the strategy types chosen ($x^2 = 78.9993$, $p < .001$) for request head acts. Head acts in Chinese language e-mails are more often realized through direct strategies, while French e-mails generally rely on conventionally indirect strategies (Table 1). The most common forms found in Chinese are hedged performatives and want statements, while in French e-mails, query preparatory (ability), query preparatory (willingness), and suggestory formulae constitute the most common head act choices (Table 2). Besides, the total

number of head acts found in Chinese e-mails is much greater than in French, implying that Chinese-speaking participants are more inclined to repeat requests than their French counterparts (Table 1).

Directness levels	Ch.	%	Fr.	%
Direct	144	**58.3%**	44	24%
Conv. indirect	102	41.3%	140	**75%**
Non-conv. indirect	1	0.4%	2	1%
Total	247	100%	186	100%

Table 1: Directness levels of head acts.

Head act strategies	Ch.	%	Fr.	%
Imperatives	22	8.9%	0	0%
Explicit perform.	14	5.7%	4	2.2%
Hedged perform.	58	**23.5%**	13	7%
Want statements	50	**20.2%**	27	14.5%
Suggest. formulae	13	5.3%	31	**16.7%**
Q. (ability)	17	6.9%	46	**24.7%**
Q. (willingness)	9	3.6%	33	**17.7%**
Q. (permission)	37	15%	0	0%
Q. (feasibility)	20	8.1%	16	8.6%
Q. (others)	6	2.4%	14	7.5%
Hints	1	0.4%	2	1%

Table 2: Head act strategies.

As to the effects of interpersonal factors, tests using a generalized linear mixed model reveal that only in Chinese e-mails does relative power have significant influence on either level of directness ($|z| = 2.099$, $p < .05$) or choice of head act strategies ($|z| = 2.105$, $p < .05$). Social distance, however, has no significant influence on either group. Moreover, although the Chinese group appears more sensitive to relative power than the French group in regards to choice of head acts, this difference is significant in neither directness level ($|z| = 1.823$, $p = .068$) nor in strategy selection ($|z| = 1.83$, $p = .067$).

3.2 Supportive moves

3.2.1. Supportive moves: coding

Supportive moves refer to the sentences preceding or following the head acts in a request e-mail. These function to reinforce requests by means of various external strategies. In our data, supportive moves are grouped according to function into 7 categories.

(1) Openings: Following the salutation in an e-mail, the addresser gives a self-introduction before arriving at the main topic. For example, *we met last month with our mutual friend [Given name][Family name]*.

(2) Grounders: The addresser describes the background of the request (the fact that he or she took a language course, applied for a scholarship, etc.), explaining why help is required or why the addressee is in a position to provide help, as well as disclosing any related information. For example, *I took basic English writing this semester*, *the program of this scholarship is highly related to your research area*, etc.

(3) Preparators: The addresser prepares the addressee for the coming request by explicitly stating that help is needed, or by inquiring about preparatory conditions for which the addressee may be able to provide aid. Preparators differ from head acts in that the addressee knows only that the other requires assistance; however the specific task to be completed is not clearly described. For example, *I need you to help me on something*.

(4) Rapport builders: The addresser makes small talk or tells jokes in order to create the impression of a friendly mood, effectively shortening the distance between themselves and the addressee. For example, *did your cold get better*, *don't laugh at me too much*, etc.

(5) Debt compensators: The addresser compensates the addressee for any potential assistance by apologizing, expressing gratitude, promising future rewards, offering verbal gifts (compliments, acknowledgements, etc.), or acknowledging the disturbance that such a request might cause. For example, *it would be really nice of you*, *next time when you're back to Taipei I'll invite you for a meal*, etc.

(6) Insistors: The addresser insists on receiving help from the addressee. This may be accomplished by a variety of methods, for example, persuading the target that the service requested is not very difficult to complete (*It wouldn't take you too much time*), expressing hope that everything will come to fruition as desired (*I hope that it will work*), etc.

(7) Imposition mitigators: The addresser lessens the burden of a request by asking for as little as possible or providing opportunities for the addressee to refuse. For example, *if you don't have time, it doesn't matter*.

3.2.2. Supportive moves: analysis results

Supportive moves	Ch.	%	Fr.	%
Openings	185	12.1%	87	8%
Grounders	760	**49.8%**	598	**55%**
Preparators	50	3.3%	43	4%
Rapport builders	94	6.2%	31	2.9%
Debt compensators	240	**15.7%**	177	**16.3%**
Insistors	104	6.8%	82	7.6%
Impos. mitigators	92	6%	67	6.2%
Total	1525	100%	1085	100%

Table 3: Supportive moves.

The results of chi-squared tests indicate that the Chinese and French groups differ significantly in terms of supportive move selection ($x^2 = 29.63$, $p < .001$). Although grounders and debt compensators are the two most commonly used strategies in both groups, Chinese participants employed openings and rapport builders at a

higher frequency than French subjects did (Table 3). Moreover, the total number of supportive moves appearing in Chinese e-mails is much higher than in French.

As to the interpersonal factors, social distance has marked effects on the usage of supportive moves in both Chinese ($|z| = 3.984$, $p < .001$) and French e-mails ($|z| = 2.963$, $p < .01$). Relative power, in contrast, has no significant influence on the types of supportive moves adopted in either group. Furthermore, neither the interaction of language group and relative power ($|z| = 0.678$, $p = .4976$) nor that of language group and social distance ($|z| = 0.311$, $p = .7555$) is significant.

3.3 Information sequencing

We investigated the ratio of supportive moves preceding the first head act to those appearing after in both Chinese and French e-mails and found that both Chinese and French participants tend to place more than half of supportive moves before the first head act. Although the percentage of fronted supportive moves is slightly higher in Chinese e-mails (59% vs. 57%), the difference between the two language groups is not statistically significant ($x^2 = 0.511$, $p = .4747$).

Furthermore, supportive moves in Chinese and French e-mails are organized in similar fashions: openings, grounders, preparators, and rapport builders all tend to appear before the first head act, while debt compensators, insistors, and imposition mitigators generally appear after.

4. Discussion

This study compared request strategies between Chinese and French e-mails, finding that the two language groups differed significantly in their choices of head acts and supportive moves. However, this difference cannot be accounted for by the traditional model of individualism/ collectivism. First of all, request strategies used in Chinese e-mails are in fact not more indirect than those in French e-mails. Chinese participants realized their request head acts most frequently through direct strategies; French participants preferred conventionally indirect strategies. In terms of supportive moves, both groups most often deployed grounders and debt compensators, but the total number of supportive moves used in Chinese e-mails is higher than in French. In other words, French participants tend to make more indirect requests than their Chinese-speaking counterparts, who in turn mitigate direct head acts with a larger number of supportive moves. This pattern does not match with that predicted in the individualist/collectivist model.

Furthermore, Chinese e-mails are not more indirect than French e-mails in terms of overall schema. On the contrary, the two groups structured their request e-mails in similar manners. In general, participants from both groups placed more than half of supportive moves before the first head act; percentages of fronted supportive moves likewise exhibit no marked difference. Moreover, the distribution of supportive moves is similar between Chinese and French e-mails. This demonstrates that indirect sequencing of information is not a politeness strategy reserved for Chinese addressers (cf., Chang & Hsu, 1998; Kirkpatrick, 1991). In e-mail requests, both Chinese and French writers preface head acts with rich preambles, preferring to structure their letters in an indirect manner.

In regards to interpersonal factors, we found that relative power and social distance influenced strategy choices in both Chinese and French e-mails. Relative power significantly affected the overall choice of head act strategy in Chinese requests, while social distance exhibited marked effects on the overall choice of supportive moves for both language groups. Moreover, the relationship between language groups and interpersonal factors is not statistically significant, indicating that interpersonal relationships have similar degrees of influence on the selection of request strategies regardless of whether the participants are Chinese or French speakers. This outcome conflicts with results drawn from previous studies on face-to-face communication. Such studies maintain that members of collectivist cultures are more sensitive to interpersonal factors than those of individualist cultures (Gelfand et al., 2004; Triandis, 1995), and that compared to their French counterparts, Chinese speakers are more likely to use different communication strategies when engaged in face-to-face talk with different interlocutors (Meng, 2006; Pu, 2003; Wang, 2009). In contrast, our study indicates that in constructing e-mail requests, French participants are equally as sensitive to interpersonal factors as Chinese, altering request strategies according to their relationship with the respective interlocutor.

5. Conclusion

This study analyzed Chinese and French request e-mails, with the conclusion that individualist/collectivist model was inadequate in accounting for differences between the two groups. In composition of request e-mails, Chinese-language participants are actually not more indirect than French-language respondents in terms of strategy choice or information sequencing, and both groups are in fact equally sensitive to interpersonal variants. E-mail, as a medium of communication, differs from face-to-face talk in that it is asynchronous as well as lacking in paralinguistic cues (Gu, 2011). Therefore, it is only natural that e-mail users have developed a style of interaction distinct from that of face-to-face conversation. Directions for revising the model of individualism/ collectivism in order to better describe e-mail communication necessitate further investigation.

6. References

Aslan, E. (2017). The impact of face systems on the pragmalinguistic features of academic e-mail requests. *Pragmatics and Society*, 8(1), pp. 61--84.

Biesenbach-Lucas, S. (2007). Students writing emails to faculty: An examination of e-politeness among native and non-native speakers of English. *Language Learning & Technology*, 11(2), pp. 59--81.

Blum-Kulka, S., House, J., & Kasper G. (1989a). Investigating cross-cultural pragmatics: An introductory overview. In S. Blum-Kulka, J. House, & G. Kasper

(Eds.), *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex, pp. 1--34.

Blum-Kulka, S., House, J., & Kasper G. (1989b). Appendix: The CCSARP coding manual. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex, pp. 273--294.

Chang, H.-J. (2011). Multinationals on the web: Cultural similarities and differences in English-language and Chinese-language website designs. *Journal of the American Society for Information Science and Technology*, 62(6), pp. 1105--1117.

Chang, H.-J., & Tseng, I.-C. (2009). Shejiao wangzhan de kuawenhua fenxi bijiao: Yi Taiwan de Wumingxiaozhan yu Meiguo de My Space wei li [A cross-cultural analysis of social network sites in Taiwan and U.S.A.: Comparisons between Wretch and My Space]. *Dianzi Shangwu Xuebao* [Journal of E-Business], 11(3), pp. 611--638.

Chang, Y.-Y., Hsu, Y.-P. (1998). Requests on e-mail: A cross-cultural comparison. *RELC Journal*, 29, pp. 121--150.

Fukushima, S. (2000). *Requests and Culture: Politeness in British English and Japanese*. Bern, Switzerland: P. Lang.

Gelfand, M. J., Bhawuk, D. P. S., Nishii, L. H., & Bechtold, D. J. (2004). Individualism and collectivism. In R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, & V. Gupta (Eds.), *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Thousand Oaks, CA: Sage, pp. 437--512.

Gu, Y. (2011). Modern Chinese politeness revisited. In F. Bargiela-Chiappini, D. Z. Kádár (Eds.), *Politeness across Cultures*. Basingstoke, UK: Palgrave Macmillan, pp. 128--148.

Holtgraves, T., Yang, J.-N. (1992). Interpersonal underpinnings to request strategies: General principles and differences due to culture and gender. *Journal of Personality and Social Psychology*, 62(2), pp. 246--256.

Hsieh, C.-L. (2011). Hanyu wanglu jiaoyou yantan zhi geti zhuyi yanjiu [A study on individualism in Chinese Internet dating discourse]. *Yuyan Jiaoxue yu Yanjiu* [Language Teaching and Linguistic Studies], 3, pp. 102--106.

Kirkpatrick, A. (1991). Information sequencing in Mandarin letters of request. *Anthropological Linguistics*, 33(2), pp. 183--203.

Lin, Y. H. (2009). Query preparatory modals: Cross-linguistic and cross-situational variations in request modification. *Journal of Pragmatics*, 41, pp. 1636--1656.

Liu, G.-H. (2003). A Contrastive Study of Request Strategies in English and Chinese. Doctoral Dissertation. Fudan University, Shanghai, China.

Meng, X. (2006). Jiaji zhong de hanfa zhiqianyu bijao [A contrastive study on Chinese/French apologies]. *Etudes Françaises* [French Studies], 3, pp. 74--79.

Pfeil, U., Zaphiris, P., & Ang, C. S. (2006). Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12, pp. 88--113.

Pu, Z. (2003). *Politesse en Situation de Communication Sino-Français: Malentendu et Compréhension*. Paris: L'Harmattan.

Triandis, H. C. (1995). *Individualism and Collectivism*. Boulder, CO: Westview.

Van Mulken, M. (1996). Politeness markers in French and Dutch requests. *Language Sciences*, 18(3-4), pp. 689--702.

Wang, M. (2009). Une Étude Comparative sur les Formules de Remerciement et d'Excuse en Chinois et en Français. Mémoire de Maîtrise. Université des Etudes Etrangères du Guangdong, Guangdong, Chine.

The gastronomic meal of the French through the tweets of Michelin star-rated chefs: characterization of the cultural heritage, and extraction of techniques and professional gestures

Julien Longhi* **, Zakarya Després, Claudia Marinica** ***, Vincent Marcilhac* ****, Felipe Diaz Marin******

*AGORA, EA7392, Université Paris-Seine

**Institut des Humanités Numériques de l'Université de Cergy-Pontoise

***ETIS UMR 8051, Université Paris-Seine, Université de Cergy-Pontoise, ENSEA, CNRS

*****Pôle de Gastronomie de l'Université de Cergy-Pontoise

{Julien.Longhi, Zakarya.Després, Claudia.Marinica, Vincent.Marcilhac, Felipe.Diaz-Marin}@u-cergy.fr

Abstract

This paper presents how the use of digital corpora – sourced from Twitter and Instagram accounts pertaining to two and three Michelin Star-rated Chefs respectively - may help reconstruct cooking techniques as well as professional gestures. Our approach, articulating data sciences and semiological analysis, proposes to take advantage of the heritage dimension embedded in the aforementioned social media. These social media are being used by two and three Michelin-star rated Chefs to convey knowledge and practices; these particular data are perceived as extremely useful for gastronomy experts to make sense of and make use of. We focus in particular on the most favoured culinary techniques, and on how these techniques transmit sensations and perceptions. From a methodological viewpoint, we narrow down on interdependencies between textual data and images/videos that illustrate techniques and gestures. For example, verbs contained in the text corpus shall enable extracting culinary techniques (i.e. *stuffing*, *puffing*, *infuse*) but also ways of experiencing gastronomy as such (i.e. *discover*, *taste*), which shows in turn that techniques are also related to the experience they provide. In addition, some adjectives may show qualities attributed through the preparation process, (i.e.: *confit*, *crispy*, *greasy*). The use of these adjectives demonstrates that the sensory description of the dishes stands at the core of Chefs' culinary rhetoric. Generally speaking, despite the brief format of messages, this research shows the interest of such an analytical method to extract and represent culinary techniques of the great Chefs.

Keywords: tweets, chefs, heritage, perception, gastronomy, textometry

1. Introduction

Since 2010, the gastronomic meal of the French has been inscribed on the representative list of the intangible cultural heritage of humanity: according to the definition from UNESCO [1], its important components are "the careful selection of dishes from a constantly growing repertoire of recipes; the purchase of good, preferably local products whose flavours go well together; the pairing of food with wine; the setting of a beautiful table; and specific actions during consumption, such as smelling and tasting items at the table." When we treat the gastronomic meal of the French as intangible heritage, we are not only interested in the dishes that are part of the French gastronomy, but also in the people who allow the preservation and the transmission of this heritage which, in the case material works of art, can be compared to curators or restorers. The Ministry of Agriculture also specifies that it is a heritage to be transmitted, and that it is important to safeguard the gourmet meal of the French. The paper is organized as follows: Section 2 presents the data we have processed. Section 3 presents the methodology and Section 4 presents the preliminary results. The last section concludes the article.

2. Building corpora

Among this list of restaurants awarded by the Michelin Guide, we selected those who had a Twitter account or whose chef manages his own Twitter account, or both. We extracted 47,923 tweets from these 61 accounts, which represents 775,754 words. We used a Python script based on the Tweepy library, a wrapper for the Twitter API that collects tweets from a given account. Some of those accounts are much more prolific than others: 7% of the accounts represent 25% of our corpus. We have chosen to focus on chefs who have obtained two or three stars in the Michelin Guide: the star attribution is based on identical criteria in order to guarantee the coherence of the selection. These criteria are five in number: quality of products, mastery of cooking and flavors, personality of the chef in his cuisine, quality / price ratio and regularity over time and over the entire card. The stars only evaluate "what is in the plate"; they only reward the quality of the cuisine. Three stars indicate that it is "a remarkable cuisine, worth the trip" and two stars means "an excellent table worth a detour". The choice of the Michelin Guide and the chefs rewarded by two or three stars as a sample of our study is justified by the fact that the Michelin Guide remains in France the reference gastronomic guide whose classification is authoritative and by the fact that the attribution of two or three stars brings great notoriety to restaurants and distinguished chefs, whose speech is relayed by the media.

3. Methodology

Textometry is an instrumented approach to corpus analysis, articulating quantitative syntheses and analyzes including text (Lebart & Salem, 1994). Textometry implements differential principles. The approach highlights similarities and differences observed in the corpus according to the representation dimensions considered (lexical, grammatical, phonetic, or prosodic ones, etc). Textometry

[1] https://ich.unesco.org/en/RL/gastronomic-meal-of-the-french-00437

establishes contextual and contrastive modeling (Pincemin 2012) and is particularly relevant to corpus exploitation in human and social sciences: detailed and global observation of different texts while remaining close to them, and highlights the fact that language is an important observation field for human and social sciences. The Iramuteq[2] software offers a set of analysis procedures for the description of a textual corpus.

One of its principal methods is Alceste. It segments 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). The choice of Iramuteq for this exploratory work is motivated by our double conceptual interest for the links between forms and themes on the one hand, and between forms and profiling of another side. This is part of the Theory of Discursive Objects (Longhi, 2015), which is based on the concept of discourse object (Longhi, 2008), the Theory of Semantic Forms (Cadiot and Visetti, 2001), which mobilizes the concepts of motifs, profiles and themes. Also, we retain the lexical classification (Ratinaud and Marchand, 2015) implemented in Iramuteq because it allows to bring out the specific themes of the leaders is to group lexical worlds and highlight the general themes of the corpus, the method seeking to "give an account of the internal order of a speech, to highlight its lexical worlds" [3]. It follows from the factorial analysis of correspondence. In our case the corpus is of reasonable size, and the software offers three major groups, as shown in Figure 1:

class 1	class 2	class 3
merci cuisine grand beau équipe	monnaiedeparis flaveur pomme tomate fleur	maurocolagreco thank new best great

Figure 1: Classification according to Iramuteq (see the complete classification in appendix)

We find three classes, or lexical worlds, in this corpus. A large part of the corpus, represented by the first class with almost 60% of the terms, is related to promotion, and another part, represented by the third class and by 24% of the terms, is also linked to promotion, but in English.
An example of tweets of this type is given in Figure 2.

Figure 2: romotion by Jean-François Piège.

However,almost 16% (second class) are messages containing actual culinary information, with terms relating to ingredients or techniques / recipes.

By filtering the corpus to form a sub-corpus specific to this class, we can have a less noisy vision of what could be the culinary heritage through the digital speeches of the great chefs. For this, we use the function Characteristics Text Segments (ST) of the relevant class and then export all ST in the class, to form, with their reunification, a new corpus.

4. Preliminary results

The analysis of this new corpus then makes it possible to use other functionalities, such as the similarity analysis, which produces graphs from the R igraph library. The input table is an attendance / absence table. The similarity matrix is calculated from one of the suggested indices. Most of the indices offered come from the R proxy library:

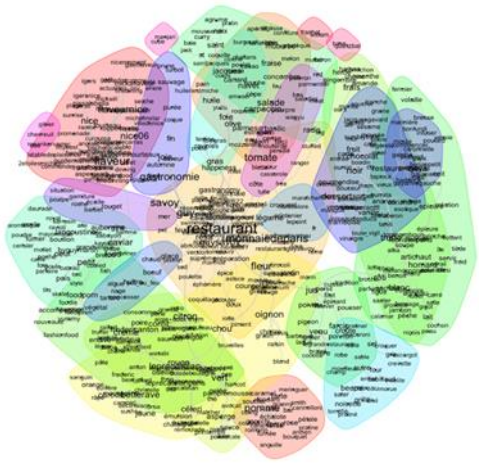

Figure 3: Similarity analysis (larger picture in Appendix)

We have a set of clusters that can find ingredients and their associations, parts of revenue perceptible through designations of dishes, or assortments or specifications.
An interesting way to experiment with this is to use the part-of-speech tagging function to look at the grammatical category of words used in the corpus, and the most common in each category (Figure 4).

[2] cf http://www.iramuteq.org

[3] cf https://datahist.hypotheses.org/11

verbs		nouns		adjectives	
stuff	137	tomato	349	black	261
laugh	98	flower	329	green	254
cook	97	gastronomy	321	small	223
smoke	89	apple	320	roasted	192
blow	72	lemon	270	fresh	190

Figure 4 : Top 5 most frequent words by grammatical categories, with their number of occurrences in the corpus (complete list in appendix)

We note that with the verbs, can be extracted culinary techniques (*stuff*, *blow*, *brew*) but also ways of living the dining experience (*discover*, *taste*) which confirms our idea that these data have a heritage interest. For example, the tweets in Figure 5 present dishes using techniques and products that fit into the heritage process presented.

Figure 5 : Tweets linked to culinary techniques

The category of nouns is interesting from a heritage point of view as well, because despite the corpus of two and three star chefs, the ingredients mentioned are not necessarily luxurious and we can thus think that the French culinary heritage is based on certain traditional products. terroir (*apple*, *onion*, *cabbage*). This link between heritage and terroir will have to be deepened. Moreover, it would be interesting to compare these results with the work carried out in the social sciences (notably the sociologist Claude Fischler) on the ingredients mentioned in the specialties mentioned by the three Michelin-starred chefs in the Michelin Guide: truffles, lobster, caviar or chocolate are among the most mentioned ingredients; but it is observed that more and more vegetables are mentioned in the specialties, which testifies to a new status of these products in the haute-cuisine. Finally, many adjectives of color are mentioned, indicating that the names used are often specified, either because the ingredient itself contains a color adjective (*green bean*), or because it is a specificity of the recipe. Other adjectives concern more the patrimonial dimension, through qualities attributed via the preparation (*confit*, *crispy*, *greasy*). The use of these adjectives shows that the sensory description of the dishes is the heart of chefs' culinary rhetoric. A more specific work of associations between these categories remains to be done, in order to grasp the combinatorial complexity of these parts of speech.

5. Conclusions

In this paper we present a work in progress with preliminary results around the French gastronomic heritage. To this end, we analyzed the tweets sent by chefs or two or three stars restaurants to the Michelin Guide. In the continuation of the works, an extension of the corpus to the one star chefs will be possible. It will nevertheless be necessary to check the consequences of such an extension in quantitative terms (number of accounts and quantity of data) and qualitative (specificities of this corpus and relevance for the general problem). Finally, an educational dimension can be seen, for cooking schools, since some tweets even give access to cooking recipes explanation videos. The scope of the analysis is therefore multiple: heritage, techniques, pedagogy, and transmission.

6. References

Cadiot P., Visetti Y.-M. (2001). *Pour une théorie des formes sémantiques* : PUF.

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

Longhi J. (2008). *Objets discursifs et doxa. Essai de sémantique discursive*. Paris : l'Harmattan.

Longhi J. (2015). *La Théorie des objets discursifs : concepts, méthodes, contributions*. Mémoire d'HDR : Université de Cergy-Pontoise.

Pincemin B. (2011). *Sémantique interprétative et textométrie – Version abrégée*, Corpus, 10 , pp. 259--269.

Ratinaud P., Marchand P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014), *Mots – Les langages du politique*, 108, pp. 57--77.

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

APPENDIX

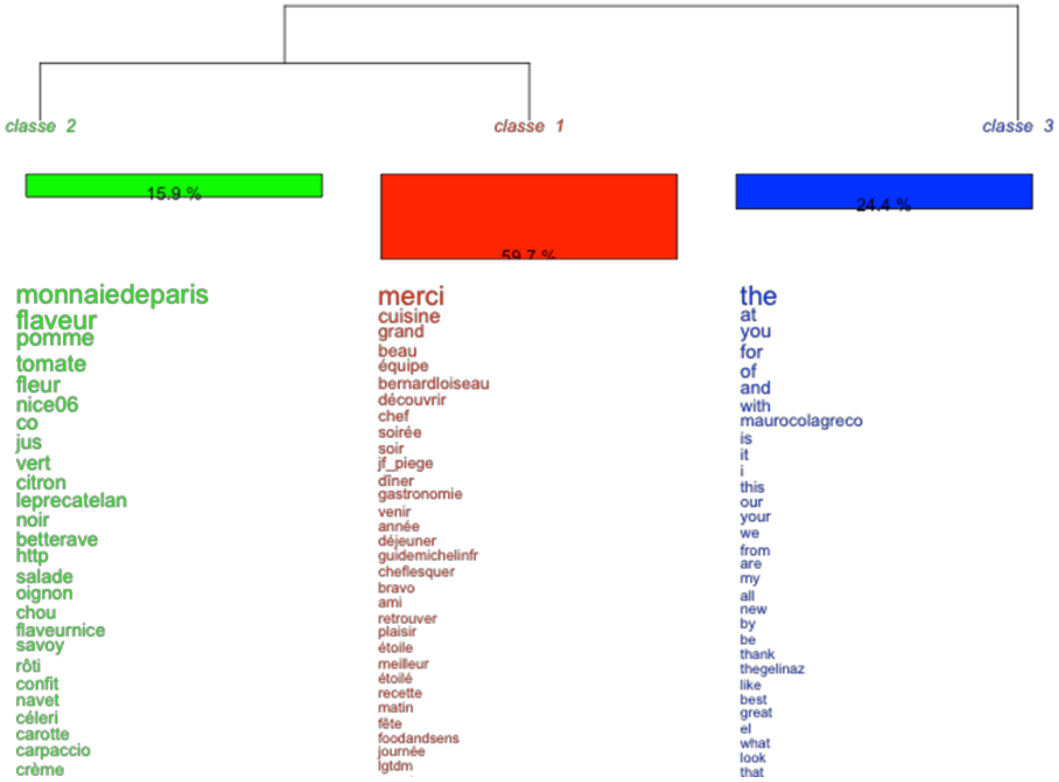

Figure 1: Complete classification according to Iramuteq

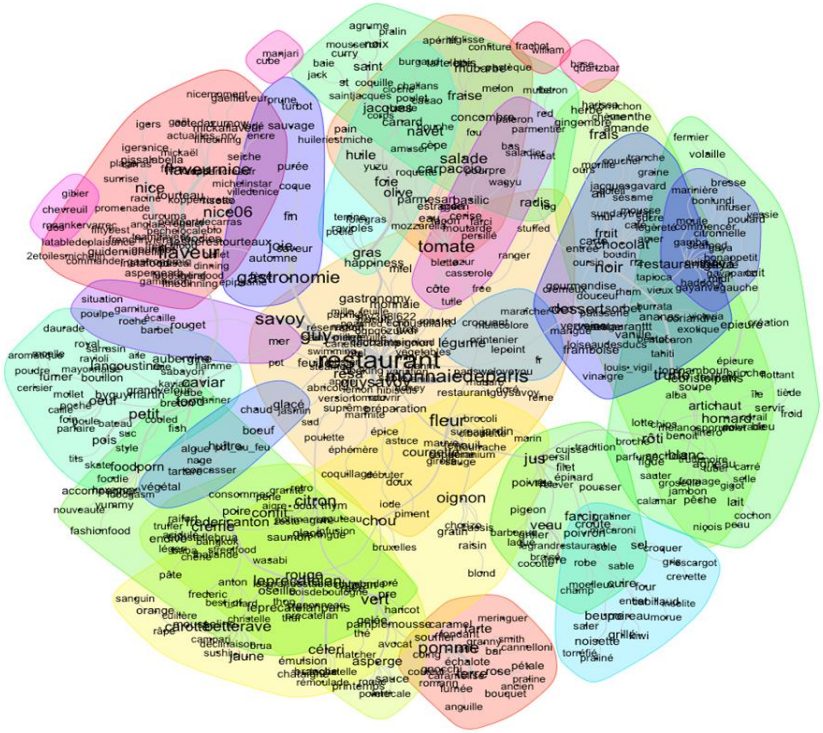

Figure 3: Similarity analysis

Forme	Freq.	Types
farcir	137	ver
rire	98	ver
cuire	97	ver
fumer	89	ver
souffler	72	ver
accompagner	67	ver
griller	65	ver
découvrir	61	ver
servir	55	ver
consommer	49	ver
commencer	41	ver
saler	39	ver
pousser	39	ver
déguster	34	ver
infuser	30	ver
mettre	29	ver
croquer	28	ver
parfaire	25	ver
concasser	25	ver
retrouver	22	ver
matcher	21	ver
goûter	21	ver
cuisiner	21	ver
truffer	20	ver
sauter	20	ver
régaler	19	ver

Forme	Freq.	Types
tomate	349	nom
fleur	329	nom
gastronomie	321	nom
pomme	320	nom
citron	270	nom
jus	244	nom
truffe	242	nom
nice	238	nom
caviar	236	nom
oignon	232	nom
salade	227	nom
dessert	227	nom
betterave	225	nom
chou	215	nom
crème	202	nom
joie	198	nom
terre	189	nom
carpaccio	182	nom
céleri	177	nom
homard	175	nom
navet	174	nom
carotte	173	nom
veau	169	nom
foie	164	nom
chocolat	163	nom
fraise	162	nom
légume	158	nom
ail	155	nom
oeuf	151	nom
tarte	149	nom
huile	148	nom

Forme	Freq.	Types
noir	261	adj
vert	254	adj
petit	223	adj
rôti	192	adj
frais	190	adj
confit	182	adj
blanc	173	adj
gras	162	adj
rouge	160	adj
jaune	138	adj
nouveau	117	adj
grillé	100	adj
glacé	96	adj
cuit	77	adj
croustillant	67	adj
sauvage	65	adj
bleu	63	adj
premier	62	adj
doux	62	adj
beau	60	adj
végétal	56	adj
gourmand	51	adj
tartare	50	adj
nouvelle	48	adj
chaud	45	adj
pourpre	44	adj
sablé	43	adj
mini	43	adj
feuilleté	43	adj
suprême	41	adj
gris	41	adj
mauve	40	adj
velouté	39	adj
poché	38	adj

Figure 4 : Most frequent words by grammatical categories, with their number of occurrences in the corpus

How FAIR are CMC Corpora?

Jennifer-Carmen Frey, Alexander König, Egon W. Stemle
Institute for Applied Linguistics, Eurac Research
E-mail: {JenniferCarmen.Frey, Alexander.Koenig, Egon.Stemle}@eurac.edu

Abstract

In recent years, research data management has also become an important topic in the less data-intensive areas of the Social Sciences and Humanities (SSH). Funding agencies as well as research communities demand that empirical data collected and used for scientific research is managed and preserved in a way that research results are reproducible. In order to account for this the FAIR guiding principles for data stewardship have been established as a framework for good data management, aiming at the findability, accessibility, interoperability, and reusability of research data. This article investigates 24 European CMC corpora with regard to their compliance with the FAIR principles and discusses to what extent the deposit of research data in repositories of data preservation initiatives such as CLARIN, Zenodo or Metashare can assist in the provision of FAIR corpora.

Keywords: research data management, computer-mediated communication corpora, reusability, FAIR principles

1. Introduction

Over the last few years, both the scientific community and the public demonstrated a growing awareness of the necessity to make research reproducible and research data reusable (see, for example, Cohen et al., 2018; Wieling, Rawee, & van Noord, 2018; or the proceedings of the second dedicated 4REAL workshop, Branco, Calzolari, & Choukri, 2018). As part of general research ethics, the scientific community commits to making research transparent, to sharing and reproducing results, and to enabling the repeated use of costly created research data. However, this has various implications for research data management that regard the way research data is collected and preserved. In order to address these issues, Wilkinson et al. (2016) published the FAIR Guiding Principles (FAIR)[1] for data management and stewardship as a result of a joint workshop on the matter. The principles provide a universal framework for data management based on findability, accessibility, interoperability and reusability that can be utilized to establish community-standards for research data management (Mons et al., 2017). Over the last few years, FAIR have received international support, for example, at the G20 International Summit in Hangzhou[2], and have been adopted within individual domains (e.g. Boeckhout, Zielhuis, & Bredenoord, 2018) as well as within important funding schemes like Horizon 2020 (European Commission, 2016). However, FAIR as such have barely been discussed in the field of language resources, although also costly created language corpora need clear and well-planned research data management.
In this work we take a look at FAIR in the context of language corpora of computer-mediated communication (CMC). We identify the FAIR principles' implications for the CMC community and describe the current state of affairs by reviewing a list of European CMC corpora and assessing their compliance with FAIR.

2. FAIR & CMC corpora

FAIR aim at describing the characteristics of research data that are beneficial for their re-use in the scientific community. They provide added value to the scientific community by facilitating knowledge discovery and ensuring the transparency and reproducibility of research results as well as the long-term preservation of funded research.
FAIR are divided into the four main groups F, A, I, R (Findability, Accessibility, Interoperability and Reusability), each of which is subdivided into sub-items, for example, F1 or A1.1. We will address them in turn and interpret the principles for CMC corpora.

2.1. Findability - F

The most important precondition for having reusable and FAIR research data is to inform others of their existence. This aspect is addressed by the Findable principle of FAIR. It requires that data is described with rich metadata (F2) and both data and metadata are assigned globally unique and persistent identifiers (F1) that link to each other (F3). Additionally, the data should be registered or indexed in a (usually field-specific) search engine (F4).
For CMC corpora, metadata can be provided on dedicated corpus web-pages or in research articles. However, in order to comply with FAIR, metadata should be "machine-actionable", this means they must be represented in a structured and machine readable format and have a persistent identifier. Research data repositories for language corpora such as CLARIN centres (Hinrichs & Krauwer, 2014) or other data repositories such as META-SHARE[3], zenodo[4], and figshare[5] provide the infrastructure to store metadata in one or multiple specific metadata formats and automatically assign persistent identifiers. To find CMC corpora, general purpose search engines like Google and Bing or specialized search engines for language resources like the CLARIN Virtual Language Observatory (VLO)[6] and the Open Language Archives

[1] https://www.go-fair.org/fair-principles/
[2] https://www.consilium.europa.eu/media/23621/leaders_communiquehangzhousummit-final.pdf
[3] http://www.meta-share.org/
[4] https://zenodo.org/
[5] https://figshare.com/
[6] https://vlo.clarin.eu/

Community (OLAC)[7] can be used.

2.2. Accessibility - A

According to FAIR, research data are accessible if they can be automatically retrieved (A1) by their unique identifier (e.g. PID, URL) using a free and open protocol (e.g. HTTP) (A1.1). However, the retrieval method should also handle authentication and authorisation for non-public data (A1.2). Furthermore, even when access rights are restricted, metadata should still be accessible (A2).
For CMC corpora, this means that access to the data does not depend on individual, personal communication (e.g. mail requests), but that the data can be retrieved autonomously by standardised methods – usually via the internet. Furthermore, conscious steps should be taken to secure the long-term preservation of the metadata. Note that all these points can usually be addressed by depositing data in a research data repository.

2.3 Interoperability - I

In order to be Interoperable, both data and metadata have to use widely accepted standards for knowledge representation that are properly and openly documented. Proprietary or undocumented formats should be avoided (I1). If vocabularies are used to populate certain fields, they should comply with FAIR (I2) and cross-references should be provided whenever possible (I3).
For CMC corpora, there is no explicit knowledge representation format for data, ultimately also because it is still unclear what is to be represented at all. But as long as the format is open, broadly used and well documented, we see this as a step in the right direction. In this respect, the TEI standard (Burnard & Bauman, 2007) and other typical formats for corpora such as XML, JSON or CSV and CMDI (Broeder, Van Uytvanck, Gavrilidou, Trippel, & Windhouwer, 2012) for metadata are good examples. Cross-references between different data are not always necessary, but become relevant in the presence of similarly named corpora, related projects, different versions of a corpus, or the publication of different sub-corpora.

2.4. Reusability - R

To comply with the final principle of reusability, data should be properly described, with the information provided being both accurate and comprehensive (R1).
Relevant and therefore necessary metadata is dependent on the specific domain and existing community standards (R1.3). However, detailed provenance is an important part of this point (R1.2). It has to be clear where the data came from and who should be acknowledged for having played a part in its creation. For CMC corpora, for example, we assume that information on the type of communication (e.g. microblog, blog, forum), the origin of the data (platform e.g. Twitter, Facebook), the year of provenance as well as the corpus creator, possible updates and version numbers are crucial for corpus reusability.
Finally, the data should have a clear and accessible usage license, so potential users know what they can and cannot do with the data (R1.1).

[7] http://search.language-archives.org

3. Assessment of FAIR data management in existing CMC corpora

3.1 Methodology

For the empirical part of this study, we investigate a list of European CMC corpora and evaluate where and to what extent they comply with FAIR.
Our selection of corpora is based on the CLARIN CMC Resource Family[8], a publicly accessible and easily findable list of corpora dedicated to computer-mediated communication. Although the list is published via the CLARIN infrastructure, it contains language resources within and outside the CLARIN community, and corpora of various sizes (from 600,000 up to 670 million tokens), sources (Twitter, Facebook, Blogs, etc.) and languages (e.g. Slovenian, Dutch, German, English, Lithuanian). Of the 24 corpora listed in the CLARIN Ressource family at the time of this study, around 50% (13) were deposited within research data repositories of the CLARIN infrastructure (12) or similar providers (these corpora are marked with an asterisk in the table). This shows a relatively high awareness of the benefits of using established infrastructures for data management. However, as depositing data in a repository does not necessarily fulfill all the requirements for FAIR, we analysed the detailed compliance with FAIR (see Section 2) for each corpus of the list. Whenever applicable, we evaluated the compliance for both metadata (abbreviated as m/M) and the data itself (abbreviated as d/D). For the evaluation of metadata characteristics, we only considered machine-actionable, structured metadata, as prescribed by FAIR and further elaborated in Mons et al. (2017), as fully compliant. Corpus websites or scientific papers dedicated to the description of the corpus, which can be considered as additional metadata (availability listed separately in the table in column *Docu*) were investigated if no other metadata was available, but would only resolve to *partially compliant*. Furthermore, we added columns to indicate the size of the corpus in tokens (Size), the openness of the data and its license (Open+Lic). We interpret the general Reusability principle (R1) as whether the – in our opinion – important information on the data provenance, author, version and year of production of texts are provided. On the other hand, we have omitted the column for the use of FAIR vocabularies in the Interoperability principle (I2) because we believe that it is not (yet) applicable to the domain of CMC corpora. We also omitted A2 (preservation of metadata after data is not available anymore) because we cannot evaluate this point.
In order to check the rather abstract principle of Findability, we queried the search engines and data repositories mentioned in section 2.1.

3.2 Results

Below we summarize the results of our investigation. The detailed evaluation for each corpus can be seen in Table 1.

3.2.1 Findability of CMC corpora

Regarding the findability of the analysed CMC corpora, we observed the expected differences between corpora that were deposited in a research data repository and those that were not. The FAIR principle F1 requires metadata and data to have a persistent identifier (PID). Although the existence

[8] https://www.clarin.eu/resource-families/cmc-corpora

of such is not always obvious, the deposited corpora all provided a PID. Similarly, machine-actionable metadata (F2) was only available for deposited corpora, while other corpora where described mainly via corpus websites or research papers dedicated to the description of the corpus. For a few corpora, neither machine-actionable nor other types of data descriptions were available. The link between metadata and data (F3) was ensured for deposited data through PIDs in the metadata. Links provided on websites or in scientific publications were in some cases outdated. Concerning the findability of corpora via search interfaces (F4) we noticed the use of a data repository greatly increased findability because most add the information to special search engines like the VLO or OLAC. To our surprise, some of the corpora did not yield any results (apart from the CMC resource family website itself) with any of these search engines.

3.2.2 Accessibility of CMC corpora

We found a similar situation for compliance with the Accessibility principle in the investigated corpora. Deposited corpora were usually more accessible in terms of the retrievability of data and metadata via standardized protocols that are open, free and universally implementable (A.1.1), and that allow for authentication and authorisation when needed (A1.2). While accessibility does not necessarily mean open or free, most deposited corpora use Creative Commons or academic licenses. For the latter, an institutional user account valid for the CLARIN infrastructure[9] (e.g. a university login) suffices to retrieve data from CLARIN repositories.

For non-deposited corpora, metadata can often only be retrieved online via the HTTP protocol, while the data is not accessible or its accessibility is not clear and standardized (e.g. mail requests). Only sometimes there is specific information on how and under which conditions the corpus can be accessed and reused.

3.2.3 Interoperability of CMC corpora

With regard to the interoperability of corpora, that is, whether they use a formal, accessible, shared, and broadly applicable language for knowledge representation and vocabulary that complies with FAIR for metadata and data, and whether meaningful cross-references are provided, the division between deposited and non-deposited corpora is not so clear.

Non-deposited corpora often do not provide metadata in a standardised format (I1), but only describe the corpus on webpages or within a research paper, having deposited the corpus in a research data repository usually includes the availability of structured metadata files. However, while CLARIN enforces the repositories to use the CMDI standard, its inherent flexibility does not ensure comprehensive and appropriate documentation. CMDI only enforces a certain way of encoding information, but there are no mandatory metadata fields, meaning that even fully compliant CMDI metadata can contain very little information. With regard to the data itself, there are no clear instructions as to the data format in which a corpus should be uploaded to CLARIN[10] or any other data repository. Hence, some of the encountered formats do not comply with the FAIR requirements of being "formal, accessible, shared, and broadly applicable".

We have also found that the vocabularies used for data and metadata (I2) are rarely standardised or even documented and therefore do not comply with FAIR.

Although the need for appropriate cross-references (I3) is a rather subjective matter, we have found some corpora that would benefit from clear cross-references to other projects, different versions, or related corpora.

3.2.4 Reusability of CMC corpora

The availability of extensive metadata is essential for the reuse of CMC corpora, this includes metadata that goes beyond the needs of the original corpus project. But since there is no clear community standard about which information has to be provided and which metadata fields have to be filled in, there is still a lot of room for improvement.

FAIR also require licensing information and information on data provenance. The deposited corpora analysed in this work were all explicitly licensed. In most of the cases, a common licensing framework like the Creative Commons licenses was used to provide clear and comprehensive licensing information. For non-deposited corpora, the licensing is less coherent. Sometimes the article describing a corpus also covers the usage license (e.g. it states that the corpus is openly available but then does not state whether it can be reused and under which conditions).

Regarding the data provenance most corpora indicated an author, however, the concrete source of the data, its year of provenance, and especially the versioning information was not always clear.

Finally, FAIR recommends using domain-relevant community standards, but there are no clear standards for CMC data that are adhered to by the majority of corpora. This regards standardised vocabularies, minimum sets of metadata as well as data formats for CMC corpora. Note that there is a TEI SIG[11], but only few corpora were actually using TEI. Moreover, although CLARIN provides a list of recommended formats[12], there are no strict rules on using them and in case a non-standard format is chosen, there is no obligation to document choices, tags or structure. This leads to relatively free data formats, that might not be well documented (e.g. custom XML formats).

4. Discussion

In general, it can be said that depositing a corpus in a data repository helps to enforce Findability and Accessibility of corpora, while non-deposited corpora, in contrast, were often less findable (e.g. they were listed in the CMC resource family but not findable via any link or paper outside of this registry) and accessible. Given the lack of PIDs and structured metadata, these corpora were generally less compliant with FAIR.

In terms of Interoperability and Reusability, however, deposited and non-deposited corpora require further steps in order to comply with FAIR. This regards especially comprehensive documentation and the use of interoperable

[9] https://www.clarin.eu/content/federated-identity

[10] CLARIN provides some guidelines on data formats (see 3.2.4) but these are very generic.

[11] https://wiki.tei-c.org/index.php?title=SIG:Computer-Mediated_Communication

[12] https://www.clarin.eu/content/standards-and-formats

and reusable vocabularies and formats for knowledge representation, which apparently have not yet been established in the community. This lack of standardised formats might be self-induced by the many different corpus tools used by the community (e.g. different formats needed for different software packages that are used in parallel, or the software might be flexible enough to use semi-standardised data structures like custom XML, JSON, or CoNLL). One could argue that the CMC community does not need such common standards because the field is very close to computational linguistics, and people are sufficiently proficient in data conversion and data handling to work with their own standards. However, this usually leads to diverging definitions of identical terms, different terms for identical concepts, or even to different underlying schemata altogether. But to achieve true *conceptual interoperability* (Chiarcos, 2012), common terms and schemata linked with a common vocabulary and embedded into an encompassing ontology are paramount.

Also the data's provenance is a critical point for reusability. Documentation and the corpus description should comprise all steps from data collection, (pre-)processing and eventual transformations and modifications. Versioning should be explicit, that is, the scope and origin of different sub-parts of a corpus and their versions must be clear and the date of any update should be indicated, especially for corpora which are being constantly refined. Furthermore, in order to be reusable CMC data also needs to provide information on the time of data collection[13], as well as on the people involved in the collection, processing, and publication of the corpus, including an up-to-date contact address.

5. Conclusion and Future Outlook

Our study analysed the data management policies for CMC corpora in Europe according to the FAIR principles introduced by Wilkinson et al. (2016). Through a detailed investigation of 24 CMC corpora listed in the CLARIN resource family, we have shown that the currently prevalent data management policies are often only partly and almost never fully compliant with FAIR principles. While depositing a corpus in repositories for data preservation (e.g. via the CLARIN infrastructure or other data repositories) helps to ensure the findability and accessibility of research data, interoperability and reusability are exclusively driven by implicit (community) standards. However, such implicit community standards are not necessarily known to everyone when creating a CMC corpus for the first time, which may lead to non-interoperable or non-reusable data. In order to promote FAIR data management for CMC corpora, we see two necessities for the future: first, (continued) interest and efforts for depositing CMC corpora at (institutional) repositories for long-term research data preservation; and second, community-driven efforts to raise awareness for all stages of FAIR research data management.

In this respect, the already ongoing efforts within the community to introduce a TEI-CMC are particularly welcome and should be supported and the creation of a CLARIN K(nowledge)-Centre[14] for CMC could formalise and centrally register already existing expertise even further. All in all, this could make research on CMC corpora truly FAIR.

[13] Note that not all data repositories provide appropriate fields for such information.

[14] https://www.clarin.eu/content/knowledge-centres

6. References

Beißwenger, M., Wigham, C. R., Etienne, C., Fišer, D., Suárez, H. G., Herzberg, L., … Zesch, T. (2017). Connecting Resources: Which Issues have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? In E. W. Stemle & C. Wigham (Eds.), Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (pp. 52–55). https://doi.org/10.5281/zenodo.1041877

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? European Journal of Human Genetics, 26(7), 931–936. https://doi.org/10.1038/s41431-018-0160-0

Branco, A., Calzolari, N., & Choukri, K. (2018). LREC 2018 workshop proceedings: 4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), 1387–1390.

Burnard, L., & Bauman, S. (Eds.). (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Chiarcos, C. (2012). Interoperability of Corpora and Annotations. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata (pp. 161–179). https://doi.org/10.1007/978-3-642-28249-2_16

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., … Hunter, L. E. (2018). Three Dimensions of Reproducibility in Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 156–165.

European Commission: Directorate-General for Research & Innovation. (2016). H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020 (No. 3).

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 1525–1531.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017).

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, 37(1), 49–56. https://doi.org/10.3233/ISU-170824
Wieling, M., Rawee, J., & van Noord, G. (2018). Reproducibility in Computational Linguistics: Are We Willing to Share? Computational Linguistics, 44(4), 641–649. https://doi.org/10.1162/coli_a_00330
Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018–160018. https://doi.org/10.1038/sdata.2016.18

APPENDIX

Corpus	Size	F1	F2	F3	F4	A1	A1.1	A1.2	I1	I3	R1	R1.1	R1.2	R1.3	Open+Lic	Docu
Corpus of contemporary blogs (cs)*	1m	y	y	y	MD	MD	MD	MD	mD	NA	AS-Y	MD	Md	mD	CC-BY-NC-ND	--
SoNaR New Media (nl)*	35m	y	y	y	MD	Md	MD	ME	MD	m	ASVY	Md	MD	MD	ACA-BY-NC-ND	WP
DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0 (de, it, en)*	600k	y	y	y	MD	MD	MD	MD	MD	NA	ASVY	MD	MD	MD	ACA-BY-NC-ND	WP
The Mixed Corpus: New Media (et)*	25m	n	n	n	md	--	--	--	MD	NA	AS-Y	md	MD	MD	on request (partly download)	W-
Suomi 24 Corpus (fi)*	2.6b	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	ACA-BY-NC	WP
CoMeRe repository (fr)*	80m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY	WP
Dortmund Chat Corpus (de)*	1m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY	WP
LITIS v.1 (lt)*	190k	y	y	y	MD	MD	MD	MD	MD	NA	ASVY	MD	MD	MD	ACA-BY-NC-ND	WP
Blog post and comment corpus Janes-Blog 1.0 (sl)*	34m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Forum corpus Janes-Forum 1.0 (sl)*	47m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
News comment corpus Janes-News 1.0 (sl)*	14m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Twitter corpus Janes-Tweet 1.0 (sl)*	139m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Wikipedia talk corpus Janes-Wiki 1.0 (sl)*	5m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Flemish Online Teenage Talk (nl)	2.9m	n	n	n	--	--	--	--	--	--	----	--	--	--	no data	--
Dereko – News and Wikipedia subcorpus (de)*	670m	y	y	y	md	Md	Md	NA	MD	m	---Y	MD	MD	MD	CC-BY-SA	WP
DWDS – Blogs (de)	102m	n	n	n	m-	--	--	m-	--	m	A---	--	--	--	only query[2]	-P
Monitor corpus of tweets f. Austrian users (de, en)	40m	n	n	n	m-	m	m	m	--	NA	AS--	--	md	-d	on request	WP
FORUMAS_INDV corpus (lt)	600k	n	n	y[1]	mD	mD	mD	D	--	m	A---	--	m-	--	download	W
INT_KOMETARAI_INDV2 corpus (lt)	4m	n	n	y[1]	mD	mD	mD	D	--	m	A---	--	m-	--	download	W
NTAP climate change blog corpus (no, en, fr)	21m	n	n	n	--	--	--	--	--	NA	---Y	--	--	--	no	P
Corpus of Highly Emotive Internet Discussions (pl)	160m	n	n	n	m-	m	m	m-	--	NA	AS-Y	--	md	--	on request	P
sms4science (de, it, fr, rm)	0.5m	n	n	n	m-	m	m	m-	--	--	ASVY	--	mD	--	only query	W
What's up, Switzerland? (de, it, fr, rm)	5m	n	n	n	m-	m	m	m-	--	NA	AS-Y	--	mD	--	no (not yet)	W
The Corpus of Welsh Language Tweets (cy)	7m	n	n	n	m-	m	m	m-	--	--	AS--	--	md	--	on request	W

Table 1: FAIR evaluation of CMC corpora.

(M) fulfilled / (m) partially fulfilled for metadata; (D) completely / (d) partially fulfilled for data; (y) yes; (n) no; (NA) not applicable
R1: (A) author information, (S) data source, (Y) year of data production/collection, (V) version information
Docu: unstructured corpus documentation: (P) scientific publication dedicated to corpus description, (W) corpus webpage
* Deposited in research data repository (e.g. CLARIN, Metashare, Zenodo)
[1] There is no structured/machine readable metadata, but the corpus website provides a link to the data
[2] Only query, web page claim CC-BY-SA

A Mixed Quantitative-Qualitative Approach to Disagreement in Online News Comments on Social Networking Sites

Louise-Amélie Cougnon[1], Jeanne Coppin[1], Violeta Gutierrez Figueroa[2]
[1] Université catholique de Louvain, ILC, MiiL ; [2] Université catholique de Louvain, ILC, Cental
E-mail: louise-amelie.cougnon@uclouvain.be, violeta.gutierrez@uclouvain.be, jeanne.coppin@gmail.com

Abstract

This paper investigates disagreement constructions on online social networking sites (SNS). It forms part of a wider project on hate and conflict speech modelling. Combining different research theories from conversational analysis and corpus linguistics, we have devised a six-class disagreement typology that we have manually tested on a 20 000-word corpus of Reddit comments on media posts. We then completed this analysis with a description of linguistic markers that pave the way towards future automated research. Finally, we present politeness strategies and repairs that maintain mutual understanding in media posts' comments. Our analysis proposes new classifications adapted to SNS. Moreover, it highlights regular forum trends, face-to-face and group threatening acts and Reddit-specific strategies to maintain or repair disagreement.

Keywords: disagreement, conversation analysis, Reddit

1. Introduction

This paper forms part of a wider project on hate and conflict speech modelling. We investigate conflict constructions in social media commentary and, in particular, in French commentary to press media posts. We have chosen to focus on Reddit productions because Reddit is a discussion forum that allows more space and freedom (although it has its own implicit rules - cf. section 3.1). To be more precise, this analysis is more about disagreement than conflict because (1) conflict is a rather subjective state of discussion and (2) the frontier between a discussion where people disagree and fight is very thin and permeable. In order to investigate disagreement in online discussions, we first manually annotate our corpus and extract basic statistics that allow a primary view of disagreement. We then qualitatively look into the corpus in order to extract conversational patterns and, specifically, disagreement and politeness strategies.

2. State of the Art and Positioning

For the past 25 years, numerous linguistics studies have focused on interactions in computer-mediated communication (CMC). Of the various approaches, we chose to follow mainly conversation analysis (CA) for its ability to show how speakers perform interpersonal actions and how these actions are organised socially by language. The framework of CA allows a better understanding of how interactions are constructed by converging to an agreement (Sacks, 1987). In this perspective, disagreement is an important part of interaction in CMC, as it is opposed to this convergence (Shum and Lee, 2013). For CA, "disagreement can be defined as the expression of a view that differs from that expressed by another speaker" (Sifianou, 2012). This definition is important because it introduces the notion of an individual's "view", which may not refer to the traditional true/untrue dichotomy. We chose to complete this definition with that of Rees-Miller (2000:1088): "a Speaker S disagrees when s/he considers untrue some Proposition Puttered or presumed to be espoused by an Addressee A and reacts with an utterance the propositional content or implicature of which is Not P". This second definition adds two nuances: "some proposition", i.e. not all (*partial agreement* is important in our typology) and "presumed to be", a nuance that includes the notion of *interpretation* in disagreement.

Conflictual and non-collaborative interactions in particular have been studied in CMC, as disagreement is likely to happen in anonymous and equal-status online communities (Shum and Lee, 2013). Some authors, such as Bou-Franch and Blitvich (2014), after an analysis of off-line taxonomies of conflictual conversations, concluded that CMC requires new models that "can tackle the affordances of digital technologies". Furthermore, each CMC subtype (social networks, forums, collaborative platforms,) seems to have its own style of communication, with explicit and implicit rules that can lead to disagreement (Poudat and Ho-Dac, 2019; Cougnon and Bouraoui, 2017). Langlotz and Locher (2012) suggested a new 5-class typology to annotate disagreement on scales related to aggression and face-threatening acts (FTAs). Aggression often comes along within disagreement studies. Two main points of view stand out in the literature: disagreement is considered either as a sign of intimacy and sociability (Simmel, 1961) or as a non-preferred response (Bilmes, 1988) and impoliteness (Sifianou, 2012). Given the anonymous status and lack of intimacy on Reddit, we will focus on the second trend. In addition to this, our Reddit corpus is made of commentaries to media posts, a context conducive to strong exchanges of views for opposition.

The concept of politeness is prevalent in situations of conflict, as disagreement can be considered inherently face-threatening (Muntigl and Turnbull, 1998). We will look into politeness strategies on Reddit, following a mixed typology based on Deng (2016) 12-classes and Maiz-Avéralo (2019) 8-classes ones. Our study also focuses on positive politeness (Brown and Levinson's (1987), notably when a speaker wants to save another's face by using politeness strategies which indicate that the FTAs produced are not intended (Deng, 2016).

3. Corpus and Methodology

3.1 Corpus Extraction

Based on its rules, Reddit falls somewhere between a social network and a discussion forum. It is the fifth most visited website in the USA, with 330 million monthly active users and more than 138,000 active communities[1]. Two main features define Reddit's specificity: (1) Posts receive *up/down* votes and either rise or drop from the top of the page depending on the users' selected factor for sorting the content (hot, controversial, new, etc.). It is therefore the community that decides which posts are more visible and popular. (2) With the *subreddit* function, posts are classified according to a theme (e.g. r/gameofthrones) or a region (e.g. r/newyorkcity).

In this study, we extracted the posts from the subreddit r/france, which mainly deals with press media content shared by the participants. We used the Reddit API[2] to retrieve posts and comments. We filtered the posts in order to eliminate those that were not sharing *news* content. Then, we kept posts that had at least 3 comments of depth-level 3 in order to make sure that the media content was debated enough and might contain disagreement. We selected a sample of 500 lines from 16 news posts with a total of 481 comments and 21,027 words. An example of the extracted content is given in Table 1. Due to a technical limitation of the API, comments from a depth > 9 were not retrieved.

Example 1 shows multilevel comments about the post "Saleté, rats, punaises de lit, embouteillages, […] Paris, ville-poubelle", published in the newspaper Marianne. The **bold** part is a PE argument (cf. section 3.2).

depth	comment
0	*Damn, I'm waiting for anti-parisian's comment. I hope s/he'll be inspired*[3]
1	*As far as I'm concerned, I'm waiting for the mandatory "What about Marseille"*
2	*Oh no, Marseille isn't in this category, otherwise they'd always win*
3	*Well **I've been to Marseille**, and in terms of cleanliness Marseille and Paris are on a par.*
4	*Usually, I'd say that they're more or less the same in terms of dirtiness, but when the garbage collectors are on strike and the mistral wind is blowing, Marseille turns into a giant dump.*

3.2 Annotation Methodology

In order to investigate the place of disagreement in online discussions, we first decided to annotate the corpus with disagreement tags. This annotation [4] focuses on the operationalisation of disagreement. It also tries to follow the 3 important CA concepts defined in section 2: subjective views (no true/untrue dichotomy), partial agreement and personal interpretation. The annotation tags are inspired by Muntigl and Turnbull (1998), Plantin (2016) and Lu, Chiu, and Law (2011).

- **IC = irrelevancy claim:** comment questioning the coherence of the speech of the interlocutor by asserting that his/her contribution has nothing or little to do with the conversation, that the argument is false or incorrect, that it lacks logic, that it is contradictory, etc. The interlocutor is sometimes invited to propose an alternative argument. Example: *You didn't hear what they were saying. Stop making up stories*.

- **CH = challenge:** comment highlighting a reluctance in relation to the discourse of the other person, materialised mostly by a marker such as the question mark or an indirect interrogative. Often followed by injunctions inviting the interlocutor to provide arguments. It can be limited to indicating that it lacks context or complexity. Example: *Are you implying that the dirtiness is due to a degradation in education and that before people didn't soil the streets?*

- **PE = personal experience:** comment advancing an argument of personal authority. The person who answers has already experienced the situation, he/she is an expert in the field, he/she knows people, etc. Presence of a clear marker of "I" (je, moi, etc.). Example 1, section 3.1.

- **EK = external knowledge:** commenting an argument with a quote from an external objective authority figure: statistics, numbers, one-time events, etc. Example: *I remember a number of twitter threads indicating that apparently Japan was not necessarily as rosy as we tend to believe in regard to cleanliness*

- **CW = common wisdom:** comment advancing an argument that refers to common sense, to the collective unconscious, to received ideas and not to studies, historical events or any other type of concrete support. Example: *The most common being the one who writes 'Long live the King' everywhere.*

- **PA = partial agreement:** comment that does not dispute entirely the other argument, agreeing to part of the statement, but disagreeing on another part. Often presented as "yes, but...". Example: *I'm not saying otherwise, only we make a big deal about this while practically ignoring good old white anti-Semitism.*

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

[2] https://www.reddit.com/dev/api/

[3] All examples are translated from French to English for readability (our translation).

[4] The inter-coder agreement (Cohen, 1960) is quite low (Kappa = 0.324 for the tagging of types and Kappa = 0.483 for tagging the presence of disagreement). This low rate can be explained by some limits we were faced with, especially the fact that irony is difficult to classify. We plan to solve this problem in future work.

3. Analyses

4.1 Quantitative Approach

4.1.1 Type Frequencies

Type	Mean/annotator (tokens)	Freq/types (%)
IC	35.7	12.7
CH	55.0	19.4
PE	35.7	13.5
EK	38.7	11.8
CW	126.3	35.8
PA	20.0	6.7

Table 1. Stats on disagreement types in the Reddit corpus.

The statistics for each disagreement type presented in Table 1 show that the most frequent types illustrated are CW and CH. CW is the most common type with a very high frequency rate (35.8) compared to second place (19.4). This can be explained by the fact that CW was considered by the annotators as being 'the catch-all type'. In addition to this, referring to common sense is the most popular way of arguing. Challenge is also a frequent type as users often want to challenge the limits of their interlocutors' argumentation. At last, we notice that these two categories form more than half of all disagreement tags (55.2), whereas, for example, nuanced argumentation (PA) only occurs in 6.7% of the cases.

4.1.2 Comment Depth

On Reddit, users have the opportunity to respond to a particular comment, thereby creating an additional news comment 'level'. We analysed the levels' depth in our corpus. Results are illustrated in Figure 1. Unsurprisingly, the frequency of comments decreases as depth increases, with a more stable rate from level 6 depth onwards.

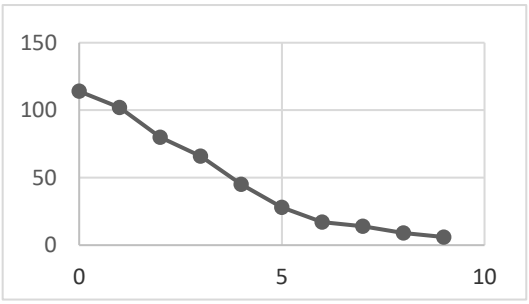

Figure 1. Stats on comment depth.

We also searched for a link between the depth of the comment and the type of disagreement, dividing the depth levels in 3 (0-2, 3-5,5+). Our hypothesis was that the more a conversation deepens, the greater the complexity of arguments (multi-type) and the more factual they become (EK). The chi-square test fails to reject H0 ($p>0.05$ $df=12$), which means that no disagreement type is more typical of a surface or of a depth argument.

4.2 Qualitative Approach

4.2.1 Linguistic Markers

For each type of disagreement, a bottom-up approach allowed us to identify linguistic particularities of each category. The linguistic markers were manually compiled to help progress manual and automatic analyses.

1. Irrelevancy claims (IC) consider other users' contributions either as false (*Stop making up stories*, *Your story is rubbish*) or inconsistent/poor (*Be a little bit coherent*, *It's ridiculous to attack him on his age*). It is a class represented by the following features:
   - lexicon (*coherent*, *true/false*, *fiction*, *fake news*, *absurd*, *nonsense*);
   - verbal modes (imperative) and clause types (exclamative). The IC user might also use an indirect interrogative to question the other participant's level of comprehension: *You have not understood a thing*, *You haven't heard* or *If you could only see*.
2. Disagreement by challenge (CH), shows the following specific features:
   - punctuation marks: "?", "!" as in "err…?";
   - quotation marks for quote back function or as regular quotes;
   - verb tenses: surprisingly, there is significant use of subjunctive, as it is the tense of indirect speech: *suggesting that*, *saying that*; imperative forms are also common and serve several functions, such as giving advice or asking for proof;
   - question words: *where, which, when, why, how*;
   - value judgement expressions: *it doesn't make any sense*, *Bravo to the hasty conclusions*, *You are the cutest*, *Thank you for being relevant;*
   - strawman technique: the special acronym "AJA" for *today I learned*[5], as in *Today I learned that one needs to be armed to justify a police intervention against himself...* /s[6].
3. The personal experience (PE) stands out with significant use of 1st person, often with repetition of pronoun *I*: *I can confirm it. I had to carry my plastic bottle through the whole of Tokyo without finding a trash can*. A predominant use of simple past can be explained by the expression of past personal experiences. Expressions of authority are common at the start and end of the sentence: *As a former smoker, I can say that, given how you feel, it's way better*.
4. The external knowledge (EK) type involves the explicit mention of an external source:
   - It often takes the form of a link embedded in a word or directly copy-pasted in the comment's

[5] *aujourd'hui j'ai appris que*

[6] Source : AJA qu'on a besoin d'être armé pour justifier une intervention des forces de l'ordre contre soit .... /s. (The special smiley(s) "/s" refers to sarcasm.)

body. It is often preceded by a brief introduction and a colon.
    - Indirect speech can be used to refer to a source that is not associated to a link (*a number of twitter threads*, *I saw a study mentioning*) or to the initial post (*One can see in the video*).
    - EK also refers to statistics and numbers (*1962: 68,7%*) and to definitions (*according to wiki* followed by the quoted content).
5. The common wisdom (CW) type calls on common sense or provides statements that are introduced as being obvious. The expression of these statements often uses reported speech, which is used to refer to what a group of people supposedly said, without providing any proof of it (e.g. *The most diligent is the one who writes everywhere 'Long live the King*').
6. Lastly, in almost all cases, partial agreements (PA) carry two main parts: the approval and the disagreement expression. PA may take a short form, in which the approval word or expression is directly followed by the disagreement: *Well done, but there are contexts and situations where it will not work*.
    - The long form incorporates the part with which the speaker agrees and then introduces the disagreement. These two sequences might form one sentence: *So yes, Lipscani and little Paris are clean, the city is nice but […] is comparable to a large and less developed European city.*
    - They can also be made of two sentences: *Yes, there are also cars […]. But what are the odds […] they hit the headstone?* Besides *yes*, we found other approval expressions: and, that, *indeed, even if*.
    - As for the alternative forms of *but*, we found: *just, well (interjection), however*.
    - A less common way of expressing partial disagreement consists of rephrasing a statement in order to nuance a part of it. In this disagreement type, the user who disagrees does not directly use the other user's words (impersonal expressions), but clearly refers to them. Example 2 hereinafter illustrates a partial disagreement with rephrasing.

depth	comment
0	*On the other side, Japan is known for not having a single trash can in town and yet being tidy. It is mainly a question of culture…*
1	*It is more a matter of (parental) education than a culture matter; we grew up in the same country and yet I do not throw away stuff anywhere.*

4.2.2 Politeness Strategies and Repairs in Reddit Comments

Two different kinds of rupture of the conversation floor have been manually observed on Reddit: (1) a rupture of mutual understanding of those interacting and (2) a rupture of politeness, often through face-threatening acts. Both fall back on politeness strategies and conversations repairs. We will focus here on Reddit-specific cases.

Mutual understanding can be disrupted because of many different phenomena, such as answer overlaps in an instant chatroom, lack of intelligibility and spelling mistakes. In order to restore it, the speakers use repairs, which can be defined as "the set of practices whereby a co-interactant interrupts the ongoing course of action to attend to possible trouble in speaking, hearing, or understanding the talk" (Kitzinger, 2013: 229). Within the structure of Reddit, the depth of the messages allows the speakers to reply to a particular message directly so that question/answer (q/a) flow is mostly respected. Sometimes a q/a pair is at the same message depth, but then one or the other user choses to which message he/she answers, and the interaction is rebalanced on the next turn. At other times, an answer is given at a 2 or more depth-level difference and the discussion is simply interrupted.

On Reddit, self-repairs are also particular as they appear with the "edit" mode on the message itself. Most of the time, the user rewrites his message by adding at the bottom "edit:" so that the other users are aware of the change. Figure 1 shows an editing part in a comment on Reddit.

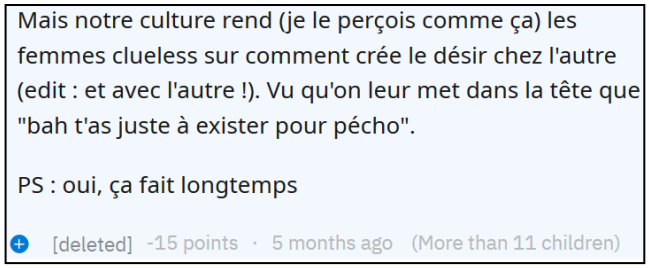

Figure 1: Capture of a Reddit comment about the post "Santé, contraception, plaisir... C'est quoi, au juste, la charge "sexuelle" ?", published on LCI.

As Muntigl and Turnbull (1998) highlighted in their work, the different disagreement types create a continuum of FTA's from least aggravating to most aggravating. PA's are on the least aggravating side because they contain an approval claim that supports the other speaker and gives them credit. Example: *You're not far from the truth, actually*. On the other hand, IC's are the most aggravating as they assume that the speaker's contribution is not relevant, often by personally attacking the speaker: *No, I think you make huge shortcuts and you have no idea how hard the RSC work in REAL life*. In this example, the speaker uses the pronoun *you* in an accusing tone. By capitalising *REAL*, he implies that the other speaker doesn't know what the reality looks like and also that his contribution is worthless. Regardless of the severity of the FTA, disagreements sometimes come along with politeness strategies that aim to reduce this act:

- Apologise for disagreeing. Example: *I'm sorry, but I disagree*. The apology *I'm sorry* softens the disagreement by giving the impression that the speaker disagrees unwillingly.
- Mitigation. Example: *I do not question at all their right to have a fair competition, but I just do not find them really serious with their approach and rather*

demagogues, crying wolf and not providing any solution. The speaker anticipates possible misunderstandings *I do not question at all…* and uses adverbs (*at all, just, really, rather*) to soften his claim.

- Acknowledgement of responsibility. Example: *So allow me to correct myself: there is some prejudice in what I said. Obviously, a raw video, well, that's enough per se. But hey, Twitter is the Internet's garbage*. The speaker acknowledges that his previous turn was subjective and contributed to the disagreement. After correcting his previous claim *Obviously… enough per se*, he reaffirms his disagreement: *But hey*.
- Common ground. Example: *We agree on most of the points*. By stating that the other speaker and himself agree on most of the points, the speaker minimises the extent of his disagreement and by extension the FTA implied.

When the first speaker decides to use one of these politeness strategies, the other speaker might decide to continue with another politeness strategy. The reaction of the second speaker seems to depend on the topic: if the topic is not too controversial it is more likely to work. On the other hand, when the topic is highly controversial or if the conflict has gone too far, the other speaker will stay indifferent to it.

5 Perspectives

This first study of our corpus opens up interesting future avenues. Our working perspectives include formalising repair strategies in order to semi-automatically find them and get an idea of their distribution, modelling linguistic features to enable disagreement type classification, and deepening the conversation analysis approach of disagreement through a qualitative analysis of its development and closing and of its consequences on social relations.

6 References

Bou-Franch, P., & Blitvich, P. (2014). Conflict management in massive polylogues: A case study from YouTube. Journal of Pragmatics, 73, pp. 19-36,

Bilmes, J. (1988). The Concept of Preference in Conversation Analysis. Language in Society, Vol. 17, 2, pp. 161-181.

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

Cougnon, L.-A. & Bouraoui, J.-L. (2017). Orality and Literacy of Telephony and SMS. In K. Bedijs & Ch. Maaß (eds). Romance Languages in the Media, Manual of Romance Linguistics, 23, De Gruyter, pp. 154-175.

Deng, J. (2016). On the Politeness Strategies in Chinese Internet Relay Chat Communication. Open Journal of Modern Linguistics, 6, pp. 293-301.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017). Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene. Proceedings of the First Workshop on Abusive Language Online, pp. 46-51.

Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C., & Sahay, S. (2017). Technology Solutions to Combat Online Harassment. Proceedings of the First Workshop on Abusive Language Online, pp. 73-77.

Kitzinger, C. (2013). "Repair". In Sidnell, J. and Stivers, T. (Eds). The Handbook of Conversation Analysis, Wiley-Blackwell.

Langlotz, A., & Locher, M. A. (2012). Ways of communicating emotional stance in online disagreements. Journal of Pragmatics, 44(12), pp. 1591-1606.

Lippi, M., & Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. ACM Trans. Internet Technol., 16(2), pp. 1–25.

Maiz-Avéralo, C. (2019). Losing Face on Facebook: Linguistic Strategies to Repair Face in a Spanish Common Interest Group. In P. Bou-Franch & P. Garcés-Conejos Blitvich P. (Eds) Analyzing Digital Discourse. Palgrave Macmillan.

Muntigl, P., & Turnbull, W. (1998). Conversational structure and facework in arguing. Journal of Pragmatics, 29(3), pp. 225-256.

Plantin, C. (2016). Dictionnaire de l'argumentation. Une introduction aux études d'argumentation, ENS Éditions, Vol. 1.

Poudat, C., & Ho-Dac, L.-M. (2019). Désaccords et conflits dans le Wikipédia francophone. CORELA - COgniton, REprésentation, LAngage, 31.

Rees-Miller, J. (2000). Power, severity, and context in disagreement. Journal of Pragmatics, 32(8), pp. 1087-1111.

Sacks, H. (1987). On the preferences for agreement and contiguity in sequences in conversation. In G. Button & J. Lee (Eds), Talk and social organization. Clevedon: Multilingual Matters.

Safi Samghabadi, N., Maharjan, S., Sprague, A., Diaz-Sprague, R., & Solorio, T. (2017). Detecting Nastiness in Social Media. Proceedings of the First Workshop on Abusive Language Online, pp. 63-72.

Shum, W., & Lee, C. (2013). (Im)politeness and disagreement in two Hong Kong Internet discussion forums. Journal of Pragmatics, 50(1), pp. 52-83.

Sifiani, M. (2012). Disagreement, Face and Politeness. Journal of Pragmatics, 44, pp. 1554-1564.

Simmel, G. (1961). The sociology of sociability. American Journal of Sociology LV (3). Reprinted in T. Parsons et al. (eds.), Theories of society. New York: Free Press.

How haters write: analysis of nonstandard language in online hate speech

Kristina Pahor de Maiti*, Darja Fišer*♦, Nikola Ljubešić♦‡

*Faculty of Arts, University of Ljubljana, Slovenia
♦Jožef Stefan Institute, Ljubljana, Slovenia
‡Faculty of Computer and Information Science, University of Ljubljana, Slovenia

E-mail: kristina.pahordemaiti@ff.uni-lj.si, darja.fiser@ff.uni-lj.si, nikola.ljubesic@ijs.si

Abstract

This paper analyzes the linguistic standardness of hateful Facebook comments in Slovene. The analysis was performed on a subset of the FRENK corpus which contains socially unacceptable discourse (SUD) towards migrants and LGBT. The nonstandard linguistic features were manually annotated using a custom-built annotation schema. The analysis showed that SUD comments are less standard than non-SUD comments and that their nonstandard features go deeper than the surface spelling deviations which are typical of CMC in general.

Keywords: hate speech, CMC, orthography, socially unacceptable discourse (SUD), standardness, Facebook

1. Introduction

Socially unacceptable discourse (SUD) is by no means a new phenomenon. However, its propagation has become more prominent with the development of social media and under the guise of anonymous/fake user profiles. Since SUD is often studied only as a societal phenomenon, its linguistic aspect is often neglected. This paper thus focuses on a set of surface linguistic features with the aim to identify the general linguistic characteristics of hateful comments in comparison to non-hateful ones.

The paper is structured as follows: Section 2 gives an overview of related work, Section 3 describes the study design, Section 4 presents the results, and Section 5 concludes the paper with suggestions for further research.

2. Related work

Hate speech can be understood as a mechanism of subordination for generating an atmosphere of fear, intimidation, harassment and discrimination (Nielsen 2002). This is done through hatred or disqualification of an individual or a group based on their race, skin colour, ethnicity, sex, disability, religion or sexual orientation (Nockleby 2000). However, Vehovar et al. (2012) point out that hate speech also includes various forms of offensive speech, such as hurtful, derogatory or obscene comments about someone.

CMC is well known for its unconventional spelling and often integrates informality and deviations from the norm also on the level of grammar and punctuation (Verheijen, 2017; Crystal, 2010). In addition, research has already proved that the personality plays an important role in CMC-communication and influences how we express ourselves (Gill et al., 2006). Despite the highly normativist culture in Slovenia, studies of Slovene tweets have shown that nonstandard writing practices are very common in informal communication on social media (Fišer et al., 2018), whereby the notion of linguistic standardness pertains to the level of author's compliance with the linguistic norm that is prescribed by normative orthographic and grammar guides (Ljubešić et al., 2018). However, Zwitter Vitez (2016) points out that negative comments actually have a more standard orthography and a more complex syntax structure than positive comments.

3. Study Design

3.1 Research questions and hypotheses

The aim of our analysis was to investigate the length, lexical richness and linguistic standardness of Facebook comments in order to establish whether any specific linguistic characteristic can be observed in hateful comments compared to the features that are typical of computer-mediated communication in general.

- Research question 1: Comment length
  - Hypothesis 1.1: On average, SUD-comments are shorter than non-SUD comments.
- Research question 2: Lexical richness
  - Hypothesis 2.1: Non-SUD comments have a richer vocabulary than SUD comments.
  - Hypothesis 2.2: Non-SUD comments contain more emoticons and emojis than SUD comments.
- Research question 3: Linguistic standardness
  - Hypothesis 3.1: Punctuation to non-punctuation ratio is higher in SUD-comments.
  - Hypothesis 3.2: SUD comments are linguistically less standard than non-SUD comments.

3.2 Dataset

In this paper, we used the FRENK corpus which contains 6,545 and 4,571 comments about migrants and LGBT respectively that were posted in response to posts on the Facebook pages of the three Slovene mainstream news media with the most visited web sites according to the Alexa service[1] (Ljubešić et al., 2019). The FRENK corpus is annotated according to a project-specific annotation

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

schema. Comments that do not include socially unacceptable discourse are marked with the "Acceptable speech" label. The rest are assigned two-dimensional labels indicating the type of socially unacceptable discourse (SUD) and its target (see Table 1). The comments targeting groups/individuals on the basis of their religion, gender, sexual orientation, ethnicity, race, etc. are annotated as "Background". If SUD is aimed at individuals due to their particular group affiliation (professional or political affiliation, etc.), then the "Other" category is selected.

Type of SUD	Background – violence (comments containing threat or call to physical violence)
	Background – offensive speech
	Other – threat (comments containing threat or call to physical violence)
	Other – offensive speech
	Inappropriate speech (comments without a specific target that contain uncivil language)
Target of SUD	Migrants/LGBT
	Related to migrants/LGBT
	Journalist/media
	Commenter
	Other (comment targeting individuals/groups that do not belong to indicated groups or are unsupportive of migrants/LGBT)

Table 1: Annotation of SUD in the FRENK project.

For our analysis, we extracted from the corpus all 520 comments containing elements of violence and threat regardless of their target for both topics (migrants: 417 comments; LGBT: 103 comments), and a randomized sample of 520 comments (respecting the same share per topic) labeled as "Acceptable speech". The dataset has been verticalized, morphosyntactically tagged and lemmatized with the ReLDI tagger (Ljubešić and Erjavec, 2016).

3.3 Typology

The annotation schema has been developed based on the guidelines for normalising CMC (Čibej et al., 2016) and the Slovene Normative Orthography Guide (Toporišič et al., 2007). Slovene has a strong prescriptive tradition that stretches far beyond orthography and also covers grammar and lexis. Our definition of standardness is based on writing conventions of language regarding spelling, lexis and grammar as set forth in the Normative Orthography Guide. The annotation schema thus consists of five categories as follows in Table 2.

Category	**Description of a class & example (*ex.*→*standard* /English/)**
Orthography (O)	Incorrect use of lower-/upper-case
	Punctuation and spacing[2]
	Typographical errors
	Regional transformations of standard lexis (*kuj*→*takoj* /immediately/)
	Character flooding (BRAVOOOOOOOOOO)
	Omission of diacritics (*ce*→*če* /if/)
Lexis (L)	Content words from dialects & slang (*lih*→*ravno* /temporal just/)
	Nonstandard abbreviations & acronyms
	Words in a foreign language
	Semantically inappropriate words (*mogli*→*morali* /could instead of should/)
Morphology (M)	Erroneous verb/noun affixes (*sprejom*→*sprejem* /with spray/)
	Grammatical gender/number/aspect
Syntax (S)	Incorrect use of grammatical cases
	Definiteness (*taglavne*→*glavne* /the main; vernacular particle "ta" added/)
	Conjunctions
	Syntactic ellipsis not justifiable by the context or clearly non-neutral
	Inappropriate parts of speech (*noben* (pronoun)→*nobeden* (noun) /no one/)
	Nonstandard collocations (*na vsake toliko kvatre*→*na vsake toliko* OR *na vsake kvatre* /from time to time; tautology/)
Word order (W)	Nonstandard/non-neutral

Table 2: Annotation schema.

3.4 Annotation

Manual annotation of the dataset was performed by one annotator who used the following reference guides: SSKJ – Dictionary of the Slovenian Standard Language; [3] Pravopis – the normative orthography guide for Slovene;[4] Slovene grammar;[5] Slovene lexicon SloLeks;[6] Janes-Norm

[2] We annotated missing, redundant, excessively repeated or incorrect punctuation markers. Spacing however was annotated only on the level of words (erroneously written together or apart, e. g. *nebo*→*ne bo* /will not/) and not on the level of punctuation markers due to the technical limitations that originate in the verticalization of the text.

[3] Dictionary of the Slovenian Standard Language: https://fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

[4] Slovenski pravopis: https://fran.si/

[5] Toporišič, J. (2004). Slovenska slovnica. Obzorja, Maribor.

[6] Slovene morphological lexicon: http://eng.slovenscina.eu/sloleks

corpus.[7] The nonstandard token is defined as every token that can be placed in at least one category of the typology. In case an emoticon/emoji[8] appeared at the end of the sentence, it was treated as a final punctuation marker. Comments written entirely in foreign language were out of scope. Even if a token could be attributed to several classes of a single category, that category has been indicated only once. In case of overlapping categories, we decided to prioritize the O-category. For example, the incorrect use of supine, which could be placed in the category S or O, was annotated with the O-label enabling us to classify a word as nonstandard simply due to its clear nonstandard graphological dimension and not by trying to guess whether the author made a grammatical mistake. Similarly, the regional variants of pronouns, possibly taking the label O or L, were again placed under the O-category, so there was no need to differentiate between typographical mistake and dialectal transformation. Since Slovene has a free word order and syntactic ellipsis can be interpreted in many ways, we annotated only clearly incorrect and non-neutral word order and omissions (e. g. missing auxiliary verb).

4. Results and discussion

4.1 Basic statistics

Our dataset comprises a total of 19,091 tokens which are equally divided between SUD and non-SUD comments. For our analysis we extracted only the relevant tokens (18,103) and removed all irrelevant tokens (988; comments written entirely in foreign language).

	SUD	Non-SUD	Total
Nonstandard	2,925 (30%)	1,842 (22%)	4,767 (26%)
Standard	6,683 (70%)	6,653 (78%)	13,336 (74%)
Total	9,608	8,495	18,103

Table 3: Structure of the dataset (number of tokens).

4.2 Quantitative analysis

Comment length. The median for the comment length was 12 tokens per comment for SUD subset and 11 for non-SUD subset. The calculation took into account all tokens in the comments. This rejects Hypothesis 1.1 that SUD comments are shorter than non-SUD comments as a result of immediate, emotional response to a newspaper article. This is in line with the findings of Zwitter Vitez (2016) that negative comments are more argumentative and complex.

Lexical richness. We observe two aspects of the lexical inventory of Facebook comments. First, we calculated the type to token ratio (TTR) for each type of discourse over 100 random draws of 1000 tokens. TTR is slightly higher for SUD comments (0.61) in comparison to non-SUD comments (0.58). Second, we calculated the content-to-function-word ratio for SUD comments which was 1.32 and thus again slightly higher compared to non-SUD comments which was 1.25. These results show that vocabulary richness is higher in SUD comments and therefore reject Hypothesis 2.1. Lexical characteristics of hateful speech are addressed in detail by Franza et al. (2019) but a quick comparison of SUD and non-SUD nouns referring to a person in our dataset shows 9 offensive nouns (*idiot, vermin, ...*) vs. 1 general (*human*).

Next, we calculated the relative frequency of emoticons and emojis which was 0.005 for SUD comments and 0.009 for non-SUD comments. We also counted the number of different emoticons and emojis. SUD comments contain 24 different emoticons and emojis while non-SUD contain 34 – 35% of which are overlapping with those found in SUD comments. To test whether the occurrence of emoticons and emojis in non-SUD comments in comparison to SUD comments is significantly higher, we ran an approximate randomization test with 1,000 iterations, obtaining a p-value of 0.0008. This means that the probability to obtain the same or greater difference between the two types of comments randomly is below 0.001. Therefore, we can safely discard the null hypothesis that there is no difference between the usage of emoticons and emojis in SUD and non-SUD comments. This confirms Hypothesis 2.2 stating that there are more emoticons and emojis in non-SUD comments.

Less frequent use of emojis in SUD could be explained by the lack of available emoticons and emojis or a more cumbersome accessibility of more specific emojis through the emoji keyboard (Bočková 2019) which can be perceived as too time-consuming during the creation of an emotionally-charged comment. In addition, the use of emojis could be influenced also by the possible communication strategy of the author with which they try to achieve emotional detachment from the content of the comment (with the absence of emoticons/emojis the emotional expressivity of the comment is lowered and the comment could be perceived as less emotional/more reasonable and thus more cogent, especially because – as argued by Micciche (in Laflen & Fiorenza 2012) – emotions can be naïvely perceived as the opposite of reason). Both subsets contain more emojis than the more traditional emoticons but the overlap of the latter is bigger between the datasets. This is not unexpected as different OS/application providers offer different sets of emojis (with greater expressiveness) in contrast to emoticons which are limited to the keyboard characters. In addition, the user needs to possess some knowledge to be able to create emoticons whereas emojis only need to be picked out of the proposed set.

Linguistic nonstandardness. By counting all punctuation markers versus all other tokens, we obtained a punctuation-to-non-punctuation ratio of 0.09 for SUD and 0.12 for non-SUD comments. This rejects Hypothesis 3.1 which

[7] Janes-Norm – manually annotated and normalized corpus of nonstandard Slovene CMC: https://www.clarin.si/repository/xmlui/handle/11356/1084?locale-attribute=en

[8] Emoticon is a representation of a facial expression with a combination of keyboard characters (e. g. :->), whereas the emoji represents facial expressions, emotions, other notions or objects in a form of a symbol, icon or a picture.

assumed more punctuation markers in SUD comments due to a possibly higher expressiveness of such comments. The result is not surprising as non-hateful comments are not necessarily all neutral. It might be useful to also take into account the sentiment of the comments in future research.

	O	L	M	S	W
SUD	2,414 (82%)	298 (10%)	44 (2%)	384 (13%)	156 (5%)
Non-SUD	1,659 (90%)	104 (6%)	17 (1%)	87 (5%)	44 (2%)

Table 4: Amount of nonstandard tokens per category in the dataset.

As Table 3 shows, a total of 4,767 nonstandard elements were identified in the dataset. The share of nonstandard tokens in non-SUD subset is 22%, whereas SUD subset contains 30% of nonstandard tokens. In Table 4, the percentage of non-standard features in SUD and non-SUD comments can be observed with regard to their type (e.g. spelling mistakes were categorized as O – Orthography). It should be noted that some tokens have been classified into more than one category (e.g. a token with incorrect spelling (O – Orthography) and grammatical case (S – Syntax)). As indicated in Table 4, by far the most prominent category in both types of discourse is Orthography with slightly over 80% of the annotations in SUD and 90% in non-SUD comments. The rest of the categories are much less frequent in both types of comments: Syntax represents 13% of the annotations in SUD comments and 5% in non-SUD, the Lexis category was assigned to 10% of SUD comments and 6% of non-SUD, Word order had a 5% share in SUD and 2% in non-SUD, and Morphology 2% in SUD and 1% in non-SUD.

The result of the chi-square test (X^2 (1, N = 18,103) = 178.4, p = 0.0001) on the independence of the variables of linguistic standardness and the social acceptability of the comment showed that we can reject the null hypothesis on the independence of the variables and accept the alternative hypothesis that these two variables are actually dependent. Based on these results we can confirm Hypothesis 3.2 that SUD comments are more nonstandard than non-SUD comments. While the prevalence of the O-category in both subsets was not unexpected as »CMC language is prototypically known for the use of unconventional, non-standard spelling« (Verheijen, 2017), it is interesting that all the other categories were twice as frequent in SUD-subset compared to non-SUD comments. A more detailed analysis would be needed, but this could indicate that nonstandard features in SUD comments are more profound and go deeper than the surface spelling deviations which are typical of CMC in general. Another interesting observation is the very low number of irrelevant comments (i.e., those written in languages other than Slovene) throughout the dataset (≤ 1%) except in non-SUD migrants-related comments where the share of irrelevant tokens was 14%. While this is not the focus of our analysis, the fact that authors of non-hateful speech convey their message in different languages could indicate their closer connection with other cultures which could also be a reason for their stance.

4.3 Qualitative analysis

Looking into the O-category of SUD subset, we noticed that nonstandard punctuation stands out as an important nonstandard feature with nonstandard punctuation markers representing 5% of all O-category labels. This mainly includes multiplication of punctuation (,,,,,,,,,,,,,,,,,,), nonstandard combinations (..!!!) and frequent use of duplicated punctuation (..) which is nonstandard despite its popular use on social media. These elements can also be found in non-SUD subset but are less frequent there.

In the S-category, we noticed a frequent use of informal syntactic structures in SUD comments (in bold):

- colloquial use of the particle »za« and preposition »od«: *za pred zid ste **za**postavit* (they should be lined up against the wall*); ni več kaj **za** razmišljat (*no need to think any further*); sranje **od** LGBT (*LGBT are a piece of shit*);*
- colloquial structure with verb »to give«: ***dajte** mi nekaj povedat (*do tell me something);
- syntactic ellipsis which can be a grammatical error (a missing auxiliary verb »to be«; now added between the slashes): *da jih **/je/ treba** ubit (*they should be killed*); **treba /je/** spustiti elektriko (*we should let the electricity flow);
- stylistically non-neutral, nonstandard structure (infinitive structure in the place of auxiliary and predicative verbs; now added between the slashes): ***/treba/** jih **/je/** kaznovati ali v zaport strpati (*they should be punished or put in jail*); mine **/je treba/** postavt (*land mines should be planted*).*

These examples are especially interesting because they are often formed as short, powerful calls to action or/and impersonal, infinitive structures which could indicate the author's (un)intentional desire to detach themselves from the content.

Lastly, we focus on the W-category and the use of emoticons and emojis. Despite the free word order in Slovene, certain placements of words are perceived by Slovene speakers as non-neutral. The analysis showed that such word order was used in more than 70% of the W-labelled tokens in both subsets. A closer look reveals that this is mainly due to the verbs that are placed at the end of the sentence. There were 45% of such cases in SUD subset and 61% of such cases in non-SUD subset (non-neutral position, i. e. final position of the verb in bold):

- *take kot si ti bi jest na garmadi **zazgal*** (I would burn people like you at the stakes*); ste **se** drugače **obnašal** (*you behaved differently*); dam si jih **peljite (***take them home*); v zemljo **se zabij (***bugger off*).*

In SUD comments, frequent non-neutral constructions of swear phrases with the noun preceding the adjective were also observed (non-neutral position of the adjective following the noun in bold):

- *golazen **pedrska (***fag vermin*); vlada **naša zablojena (***our stupid government*); golazen **necloveska zblojena***

*(*stupid inhuman vermin*); vsiljivce* ***teroristične*** *(*terrorist intruders*)*.

Taking into account that in Slovene the most important information comes last, we could suppose that by placing a verb at the end, the author puts the emphasis on the action, whereas in the case of the adjective at the end of a sentence, the author probably wants to draw the reader's attention to the quality of the headword they modify and use it as a justification of the overall meaning of the comment.

As for the use of emoticons/emojis, first a well-known phenomenon of positive emoticons/emojis being used in SUD comments has been observed. The pragmatic function of this strategy is to weaken the illocutionary force of the message (Li and Yang, 2018):

- *metek v glavo:)* – bullet to the head:)*; noter bi jih zaprli pa se naj kurijo ;)* – let them be shut inside and burn ;)*; lahko jim vržem samo ročno granato :P* – all I can do is to throw them a hand grenade :P.

Second, we analyzed the overlapping emoticons/emojis which showed to be of two categories: thematic symbols (🐷) and facial expression symbols (:D). Negative symbols prevail in SUD comments (:-(, 🤬, 😈, 😡, etc.) while there are only three such symbols in non-SUD comments (:(, 🙄, :/). Although positive symbols are used in both subsets, non-SUD comments contain many more different symbols for happiness. We also noted the following subset-specific symbols: in SUD comments we found two symbols of physical and social power: 💪 and 😎, whereas in non-SUD comments we could observe symbols of peace and love: <3, :* and ✌.

5. Conclusion

The goal of this paper was to examine basic surface linguistic features of our dataset and the level of standardness of SUD comments in relation to non-SUD comments. The quantitative analysis showed that SUD comments are longer and have richer vocabulary. Furthermore, we observed a lower frequency of emoticons/emojis and punctuation markers in SUD comments.

The results also show that SUD comments are indeed less standard than non-SUD comments even though this feature did not prove as characteristic as initially expected since the nonstandardness of non-SUD comments was also relatively high (30% vs. 22% respectively). SUD comments however exhibit a peculiar tendency towards nonstandard features (namely deviations in syntax and word order conventions) that surpass simple spelling errors (which are typical of CMC in general) and go deeper into the language structure.

As future work we envisage to examine the variety of punctuation markers used and the characteristics of final punctuation, the underlying meaning of positive emoticons in hateful comments, and the semantic role of syntactic ellipsis and impersonal structures. Valuable insights could also be gained by investigating the syntactic complexity, vocabulary and argumentation strategies in SUD comments.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project »Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society« (J7-8280, 2017–2019) and the Slovenian-Flemish bilateral basic research project »Linguistic landscape of hate speech on social media« (N06-0099, 2019–2023).

7. References

Bočková, R. (2019). The Use of Punctuation, Emoji and Emoticons in YouTube Abusive Comments. Diploma thesis. Charles University. Praga. http://hdl.handle.net/20.500.11956/107360

Crystal, D. (2010). Language and the Internet. University Press, Cambridge.

Čibej, J., Arhar Holdt, Š., Erjavec, T., Fišer, D., Zupan, K. (2016). Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje. http://nl.ijs.si/janes/wp-content/uploads/2014/09/Janes-smernice-v1.0.pdf

Fišer, D., Miličević, M., Ljubešić, N. (2018): Zapisovalne prakse v spletni slovenščini. In Fišer, D. (Ed.) (2018). Viri, orodja in metode za analizo spletne slovenščine. Ljubljana: Znanstvena založba Filozofske fakultete. pp. 124–139.

Franza, J., Ljubešić, N., Fišer, D. (2019): The lexical inventory of Slovene socially unacceptable discourse on Facebook. 7th Conference on CMC and Social Media Corpora for the Humanities. Cergy-Pontoise, France. Submitted.

Gill, A. J., Nowson, S., Oberlander, J. (2006). Language and Personality in Computer-Mediated Communication: A cross-genre comparison. Journal of Computer Mediated Communication.

Laflen, A., & Fiorenza, B. (2012). "Okay, My Rant is Over": The Language of Emotion in Computer-Mediated Communication.
https://doi.org/10.1016/j.compcom.2012.09.005

Li, L., Yang, Y. (2018). Pragmatic functions of emoji in internet-based communication---a corpus-based study. Asian-Pacific Journal of Second and Foreign Language Education 3.1 (2018).

Ljubešić, N., Fišer, D., Erjavec, T. (2019): The FRENK datasets of socially unacceptable discourse in Slovene and English. https://arxiv.org/abs/1906.02045

Ljubešić, N., Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In Nicoletta Calzolari et al. (Ed.) (2016). Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia. European Language Resources Association (ELRA). pp. 1527–1531.

Ljubešić, N., Erjavec, T., Fišer, D. (2018). Orodja za procesiranje nestandardne slovenščine. In Fišer, D. (Ed.) (2018). Viri, orodja in metode za analizo spletne slovenščine. Ljubljana: Znanstvena založba Filozofske

fakultete. pp. 74–98.
Nielsen, L. B. (2002). Subtle, Pervasive, Harmful: Racist and Sexist Remarks in Public as Hate Speech. Journal of Social Issues, 58: 265-280.
Nockleby, J. T. (2000). Hate speech. Encyclopedia of the American constitution, 3(2), 1277-9.
Toporišič et al. (2007). Slovenski pravopis. ZRC SAZU, Ljubljana.
Toporišič, J. (2004). Slovenska slovnica. Obzorja, Maribor.
Vehovar, V., Motl, A., Mihelič, L., Berčič, B., Petrovčič, A. (2012). Zaznava sovražnega govora na slovenskem spletu. Teorija in praksa, 1(49), 171–189, 233–234.
Verheijen, L. (2017). WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication. In D. Fišer, & M. Beißwenger (Eds.), Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World. Ljubljana: Ljubljana University Press. pp. 72-101.
Zwitter Vitez, A., Fišer, D. (2016). Linguistic Analysis of Emotions in Online News Comments - an Example of the Eurovision Song Contest. In Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities. Ljubljana, Slovenia. pp. 74–76.

The lexical inventory of Slovene socially unacceptable discourse on Facebook

Jasmin Franza*, Darja Fišer*+
*Faculty of Arts, University of Ljubljana, Slovenia
+Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: jasmin.franza@ff.uni-lj.si, darja.fiser@ff.uni-lj.si

Abstract

While social media enables the plurality of opinions, it is also where a lot of discrimination and hate take place. This paper presents a lexical analysis of Slovene socially unacceptable discourse in the FRENK corpus of Facebook comments on topics related to migrant and LGBT issues in order to establish the lexical footprint of SUD online. It also shows how the vocabulary of hate changes depending on the target and that words are often not hateful on their own, but that it is the context that makes them hateful.

Keywords: hate speech, lexical analysis, corpora, socially unacceptable discourse (SUD)

1. Introduction

With the spread of social media, socially unacceptable discourse (SUD), such as hate, discriminatory, offensive or threatening speech, has reached new dimensions. It is essential to analyse the newly developed practices by which SUD is spread to better understand and efficiently tackle the problem. There has been a large increase of expressions of intolerance and hatred towards migrants and LGBT in Slovenia since 2015, when the Balkan migrant wave reached the country and the referendum on same-sex marriage took place. This paper focuses on the surface-level lexical features with the goal of describing the general lexical characteristics of offensive and violent Facebook comments in Slovene targeted at migrants and LGBT from the FRENK[1] corpus. Thus, we identified the characteristic lexical footprint of SUD against the two target groups of interest.

Section 2 gives an overview of the related work, Section 3 focuses on the description of the corpus used, Section 4 describes the study design, in Section 5 results are discussed and Section 6 concludes the paper with some ideas for further research.

2. Related work

There is an increasing need of developing new linguistic resources and tools for the analysis and identification of hate in online language. One of the main sources of data in investigating hate speech has been harvesting user-generated content from comments on online platforms (Millar et al., 2017). Projects aiming to analyse and restrict hate speech online, including those that approach the problem from a linguistic point of view are active in many countries. The FRENK[2] project aims to provide a deeper understanding of SUD in Slovenia combining legal, linguistic and social perspectives (Fišer et al., 2017).

This paper presents a descriptive lexical analysis of our corpus, which has been shown to be an efficient tool for examining discourse features in previous publications. In a corpus analysis of tweets Fišer and Kalin Golob (2018) identify typical behaviours of different types of users and their adoption of evaluative adjectives using keywords. They also perform an analysis of the word classes through which they get an insight into the communicative goals of Twitter users. The corpus method has been employed for investigations of the SUD lexis as well. For example, ElSherief et al. (2018) focus on the targets of hate on social media and propose a model for distinguishing whether it is directed towards a specific person/entity or generalized towards a group of people sharing certain characteristics. Their lexical analysis highlights the most noticeable features that distinguish between directed and generalized hate speech, similar to our approach in which we focus on the distinctions between hateful and violent comments. Brindle (2016) applies a corpus-driven approach to study frequency, keywords, collocations and concordances of white supremacist language, thereby facilitating the understanding of hate speech online with the aim of providing tools to combat extremism and intolerance.

3. Corpus

The FRENK corpus was collected from Facebook pages of three mainstream news media. It covers two topics, migrants and LGBT, and was enriched with manual annotations of the comments (Ljubešić et al., 2019). The corpus contains 30 posts with 6545 comments for migrants, and 93 posts with 4571 comments for LGBT. Comments were annotated for the type (*acceptable*, *background-violence*, *background-offensive speech*, *other-threat*, *other-offensive speech* and *unacceptable*) and the target of SUD (*migrants* or *LGBT*, *related to migrants* or *LGBT*, *journalists or media*, *commenter*, *other*).

The annotation campaign was performed by 32 master students from the University of Ljubljana. They all attended an initial half-day annotator training, after which they were randomly split between the two topics (migrants or LGBT). Each comment was annotated by approximately eight annotators and after each task comments with the highest inter-annotator disagreement were manually analysed by an expert (a social scientist working at the national centre for reporting hate speech) in order to communicate the disagreement issues and improve the usbsequent annotations (for more details see Ljubešić et al., 2019). In this paper we take into account the most frequent label selected by the annotators.

[1] http://nl.ijs.si/frenk/.

[2] http://nl.ijs.si/frenk/english/.

The main distinctions between the types of SUD depend on whether SUD is aimed at the background of a person (i.e. on the basis of their religion, sexual orientation etc.), whether SUD's target are other groups or individuals (in contrast to just being unacceptable in terms of swearing) and whether there are elements of violence in SUD, amounting to the following six categories:
(1) Acceptable speech includes all the comments where there are no elements of hate or violence, they are either neutral or positive comments. E.g.: "*Slovenci, dajmo se naučit sprejet drugačnost.*" (Slovenians, let's learn to accept something different.)
(2) Background-violence is selected when there is a threat or call to physical violence, e.g.: "*Begunce poslati v koncentracijska taborišča in zapliniti, postaviti zid in postreliti vse, ki pridejo blizu."* (Refugees should be all put in concentration camps and fumigated, a wall should be built and all who come by should be shot.)

(3) Background-offensive speech is selected when there is no call to physical violence, but the comment is offensive due to background, e.g.: "*Muslič̌i so majmuni in posiljevalci, sama drhal*." (Muslims are monkeys and rapists, they're only a rabble.)

If the comment is directed against individuals on the basis of their individual characteristics or due to them belonging to other groups (doctors, journalists, firefighters etc.), it belongs to one of the *Other* categories. If there is a threat or call to physical violence, the comment was categorised as (4) *Other-threat*, e.g.: "*Tanja, ti si neumna, prišel bom k tebi domov in te pretepel*" (Tanja, you're dumb, I'll come to your house and beat you up.). If there is no call to physical violence, but the comment is still offensive, it is classified as (5) *Other-offensive speech*, e.g.: "*Tone, kar pišeš je neumno, nimaš pojma, si bedak in rit usrana*." (Tone, what you write is stupid, you have no idea, you are a shitty fool.)

If the comment is not directed against anyone specifically, yet it contains terms that are socially unacceptable, controversial, obscene or vulgar, it belongs to the category (6) *Inappropriate speech*, e.g.: "*Kakšen kretenizem!*" (What an insanity!)

In this paper, we took into account both topics (migrants and LGBT) but focus only on the comments labelled *background-violence* and *background-offensive speech*, using the *acceptable* comments as the control group. The lexical analysis was conducted in the SketchEngine [3] corpus query system (Kilgariff et al., 2014) where we compiled five subcorpora which were automatically tagged and lemmatized with TreeTagger. [4] Due to automatic preprocessing and the frequent usage of non-standard language in the comments (which are discussed in detail in Pahor de Maiti et al., 2019), there are expectedly some pre-processing errors which we manually marked as such and excluded from the rest of the analysis.
The size of each subcorpus is given in Table 1. Among the SUD subcorpora, the one with offensive comments against migrants is the largest, while the one with violent comments against LGBT the smallest. For both targets the offensive subcorpora are bigger than the ones concerning violence. It is noteworthy that the acceptable subcorpus represents more than half of the entire corpus, which is a welcome surprise, given that the topics were chosen with high expectations for SUD.

Subcorpus	No. of tokens	No. of words	No. of unique words	No. of unique lemmas
Background, offensive – LGBT	26,842	22,684	7,172	4,797
Background, offensive – migrants	57,696	49,758	12,332	8,066
Background, violence – LGBT	1,202	1,048	678	549
Background, violence – migrants	7,010	6,088	2,704	1,984
Acceptable	99,885	84,526	19,549	12,916
Total	*192,635*	*164,104*	*42,390*	*28,312*

Table 3: Size of the FRENK corpus.

It is important to keep in mind that the research carried out in this paper was done on a very small corpus using techniques usually employed for much larger amounts of data. Still, the methodology showed to be useful and provided a valid insight in the data.

4. Study design

In this study, we focused on a descriptive lexical analysis of frequency lists and keywords in order to observe the lexical tendencies in SUD, comparing the results across topics and SUD types.
Establishing the frequency of words within the collected material is one of the most basic analyses that can be carried out using a concordancing software (Page et al., 2014) and it helps to establish the regular patterns or norms in the dataset. We analysed the frequency distribution of the word classes as well as the frequencies of the nominal, verbal and adjectival lemmas. In order to enable generalization of the results, we also manually tagged the lemmas in focus with semantic fields, developed through the observation of the data. Semantic tagging means assigning semantic categories to words and it is used to overcome the lack of semantic information syntax-oriented part-of-speech tagsets usually have (Abdou et al., 2018). Several semantic categorisation schemas have been proposed (such as the Lancaster UCREL Semantic Tagset USAS[5]; Rayson et al., 2004), but we decided to develop our own because of our highly specific focus in order to better highlight the most important semantic distinctions. The semantic annotation was performed manually by two independent annotators who subsequently sought agreement on the final tag. Nouns were categorised in the following semantic fields (some examples translated into English are shown in brackets): TARGET-SPECIFIC ("migrant", "refugee", "homosexual", "lesbian"), MILITARY ("weapon", "army"), SOCIAL ROLE ("mother",

[3] https://www.sketchengine.eu/.
[4] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
[5] http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf .

"child"), ADMINISTRATION ("country", "politics"), TRANSPORT ("train", "ship"), IDEOLOGY ("god", "home"), NON TOPIC-SPECIFIC ("time", "people") and NOISE (pre-processing errors). Proper names were categorised in the most suitable semantic field as well. As a subcategory of TARGET-SPECIFIC nouns we also took into account INSULTS ("fag", "vermin") in order to investigate in detail the form in which hate is expressed against each of the targets. As far as verbs are concerned, we singled out only those that express VIOLENCE ("kill", "shot") and classified the rest as OTHER, as we were interested especially in determining how much is violence expressed directly. Adjectives were grouped on the basis of their sentiment, distinguishing between POSITIVE ("healthy", "smart"), NEGATIVE ("sick", "poor") or NEUTRAL ("similar", "Slovene"). The classification and analysis were performed out of context with a possibility to examine the concordances as well wherever required.

For each subcorpus a frequency list of word classes was compiled in order to compare the share of each word class across the subcorpora. Next, we analysed the 40 most frequent nouns, 30 verbs and 20 adjectives. As the subcorpus *violence-LGBT* is very small, it was analysed but not compared to other subcorpora. Finally, in order to formulate key concepts within a specific subcorpus which make it distinctive (Brindle, 2016), we examined keywords to have a contrastive view on the subcorpora and identify their major differences. Baker (2010, 104) defines a keyword as "a word which occurs statistically more frequently in one file or corpus, when compared against another comparable or reference corpus". We employed SketchEngine for keyword extraction which uses the simple maths method to determine the keyness score of keywords. It works with normalized (relative, per million) frequencies in the focus and the reference corpus (Kilgariff, 2009).[6]

In this paper we compared the SUD subcorpora with the *acceptable* subcorpus.

Our hypothesis is that SUD has a different lexical footprint than acceptable discourse. Moreover, we expect that hate is expressed in a distinctive way against each target. It is also presumed that some traits of SUD could be shared, therefore it is our interest to discover both unique and overlapping features of acceptable and unacceptable discourse.

5. Lexical analysis

5.1 Frequency

Word classes. From the word class analysis (Table 2), it can be seen that proper nouns represent a very small share of the vocabulary, which indicates a generic nature of the comments. No major differences among the subcorpora were observed, which could indicate that the language on Facebook is quite uniform, be it offensive/violent or not, no matter which group it targets. Still, there are some more subtle differences that are worth investigating.

First, in the v*iolence-migrants* subcorpus common nouns (+2.38%) and main verbs (+1.71%) are more frequent in comparison to the *acceptable* subcorpus. Also, both subcorpora expressing *violence* contain more adpositions (+1.74% *migrants* and +1.43% *LGBT*) than the *acceptable* one. Interestingly, especially in *violence-migrants* (-2.62%), but also in *offensive speech-migrants* (-1.13%) and *violence-LGBT* (-1.63%) there is less punctuation. It would be interesting to examine if there is any correlation with the emotional state of the author, i.e. if the commenters do not use punctuation in a rush of anger.

Word class	*Violence migrants*	*Offensive migrants*	*Violence LGBT*	*Offensive LGBT*	*Acceptable*
Noun - proper	5%	4%	5%	3%	5%
Noun - common	22%	19%	22%	21%	20%
Verb - main	13%	12%	12%	11%	11%
Verb - auxiliary	7%	7%	7%	7%	7%
Adjective	6%	6%	6%	7%	6%
Adverb	7%	7%	8%	6%	7%
Particle	4%	4%	4%	4%	5%
Interjection	0%	0%	0%	0%	0%
Pronoun	11%	12%	11%	10%	12%
Numeral	1%	1%	1%	1%	1%
Adposition	8%	7%	8%	6%	6%
Conjunction	10%	10%	10%	11%	10%
Abbreviation	0%	0%	0%	0%	0%
Punctuation	7%	8%	8%	10%	10%

Table 2: Share of the word classes per subcorpus.

Nouns. As seen from table 3, the distribution of the semantic fields is quite variable across the subcorpora with IDEOLOGY representing a quarter of the inventory in all the subcorpora except *violence-migrants*, where the most represented category is MILITARY. It is interesting to note that in both of the *offensive* subcorpora INSULTS are not among the most frequent nominal lemmas we investigated, which indicates that commenters usually do not use slurs or offence words but rather employ other means to express hate. In the *violence* subcorpora hatred is expressed through INSULTS for LGBT and MILITARY nouns for migrants.

	Violence migrants	*Offensive migrants*	*Violence LGBT*	*Offensive LGBT*	*Acceptable*
TARGET-SPECIFIC	10%	10%	/	18%	8%
TARGET-SPECIFIC: INSULTS	15%	5%	**30%**	3%	/
MILITARY	**30%**	5%	10%	/	/
SOCIAL ROLE	5%	10%	5%	**25%**	23%
ADMIN.	13%	**23%**	3%	10%	13%
TRANSPORT	5%	/	/	/	/
IDEOLOGY	5%	**23%**	23%	**25%**	**25%**
NON TOPIC-SPECIFIC	8%	20%	25%	13%	20%
NOISE	10%	5%	5%	8%	13%

[6] https://www.sketchengine.eu/documentation/simple-maths/?highlight=simple%20maths.

Table 3: Distribution of nouns across semantic fields.

The TRANSPORT category is observed only in the *violence-migrants* subcorpus where commenters describe how and by which means they want to get rid of the refugees.

Interesting differences can also be seen in the NON TOPIC-SPECIFIC category, such as the frequent occurrence of "world" and "life" LGBT subcorpora, and "nation", "home" and "money" in the migrants subcorpora. This is an important indicator of the themes surrounding and constituting hate speech towards our two targets.

Apart from the overlap of the most frequent categories, there is a substantial lexical overlap across the subcorpora as over 50% of the analysed nominal lemmas from the *acceptable* subcorpus overlap with the lemmas from the SUD subcorpora. Those are for example "family", "child", "right" for LGBT and "man", "state", "muslim", "refugee" for migrants.[7] The overlapping lemmas are neutral per se, yet the commenters make them negative and use them to hurt and feel superior.

Verbs. Verbs expressing violence are expectedly present only in the *violence* subcorpora and represent a fair amount of the frequency list (*migrants* 20%, *LGBT* 30%). However, our analysis shows that hate is expressed in different ways towards the two targets. Among the most frequent verbs in the violent comments towards LGBT, "torture", "beat", "shoot" and "kill" are used, whereas for migrants the verbs "slaughter" and "shoot" were encountered, which shows the aggressive attitude towards the two groups.

Adjectives. In all five subcorpora the most frequent adjectives do not express any connotation (~ 50%). The largest proportion of negative adjectives was found in the *violence* subcorpora (*migrants* 30% and *LGBT* 40%), although it is difficult to draw any conclusions from these results due to the small size of the dataset. The most hateful adjectives were found in the violent comments towards migrants, e.g. "killed", "dead", "barbed".

5.2 Keywords

Offensive-migrants. Unsurprisingly, the key lemmas "refugee" (keyness score 1.890) and "migrant" (1.580) stand out in the comments on the respective topics. Among the morphosyntactic categories**,** the most typical for this kind of discourse are female proper nouns in singular accusative (1.420) and locative (1.360), and main verbs in present tense in third person plural (1.420). A concordance analysis shows that this result stems from the fact that in Slovene, countries are feminine and Facebook users wish to send migrants to other countries.

Offensive-LGBT. Commenters worry about children being adopted by LGBT and therefore the lemma "child" stands out with a keyness score of 2.150. Among the top 20 most salient keywords, we expectedly find 50% of LGBT-related lemmas, while 25% are associated with family and 20% are insults.

Violence-migrants. In the 20 top-ranking key lemmas 70% of them are related to violence and military action. In the list of key morphosyntactic categories, supine main verbs rank highest (4.880), which also has an imperative meaning in the Slovene grammar.

Violence-LGBT. Among the 20 top-ranking key lemmas 45% are insults or suggest violence against LGBT, which is less than in the subcorpus of violent comments against migrants (30%). We can therefore confirm that commenters employ different strategies to show their violence against these minority groups. Examining the key morphosyntactic categories, we noted that this is the only subcorpus that has punctuation (3.320) among the most significant categories, which is surprising and gives room for further investigation.

6. Conclusions

The goal of this paper was to determine the lexical footprint of SUD against migrants and LGBT through a descriptive analysis of frequency and keyword lists. We identified the SUD footprint as consisting of frequent usage of supine, verbs in third person, common nouns and main verbs, while punctuation is often absent.

As hypothesised, the results show that Facebook users express offence and violence towards these two groups in different ways, as can be seen from the adjective and verb frequency analysis. It has also been noted that while some words are not hateful on their own, the context in which they are used makes them offensive or violent. This is an excellent starting point for an investigation on evaluative lexis which is planned as future work.

Current hate speech detection primarily puts attention on distinguishing between hate speech and non-hate speech. However, as our analysis reveals, hate speech is far more nuanced which must be taken into account to effectively tackle hate speech online.

Moreover, the different amount of insults in the keyword analysis of the *violence* subcorpora confirms as well that commenters on social media respond to migrants and LGBT in a different way. From the frequency lists it was possible to note that commenters use military lemmas mostly against migrants (33%), while this topic is far less common against LGBT (5%). On the other hand, more insults are used towards LGBT (28%) than migrants (15%).

While this work encompasses primarily a descriptive statistical analysis of the lexical inventory of the FRENK corpus, future activities will include statistical inference testing of the selected phenomena. In our future work, we will also compare the results of the Slovene FRENK corpus with its English counterpart which has recently been annotated as well as tackle the connotative meaning of the lexis used in SUD oriented towards migrants and LGBT. Furthermore, more specific analyses on the *acceptable* subcorpus will be made, for example splitting it in two separate subcorpora, one for LGBT and one for migrants, to have a deeper understanding of the target-specific language as such.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national research project »Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society« (J7-8280, 2017 – 2019) and the Slovenian-Flemish bilateral basic research project "Linguistic

[7] All the examples of the words from the FRENK corpus are direct translations from Slovenian. The POS has been preserved, yet in translation of the whole phrase some words could be paraphrased.

landscape of hate speech on social media" (N06-0099, 2019 – 2023). We would also like to thank Nikola Ljubešić for collecting and processing the data we used in this paper.

8. References

Abdou, M., Kulmizev, A., Ravishankar, V., Abzianidze, L. and Bos, J. (2018). *What can we learn from Semantic Tagging?*. https://arxiv.org/abs/1808.09716.

Baker, P. (2010). Corpus methods in linguistics. In Litosseliti, L. (Ed.), *Research Metgods in Linguistics*. London: Continuum, pp. 93—116.

Brindle, A. (2016). *The Language of Hate. A Corpus Linguistic Analysis of White Supremacist Language.* London and New York: Routledge.

ElSherief, M., Kulkarni, V., Nguyen, D., YangWang, W. and Belding, E. (2018). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*. Palo Alto, California: Stanford University, pp. 42—51.

Fišer, D. and Kalin Golob, M. (2018). Analiza tvitov slovenskih korporativnih uporabnikov. *Proceedings of the Conference on Language Technologies & Digital Humanities.* Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 69—76.

Fišer, D., Ljubešić, N. and Erjavec, T. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. *Proceedings of the 1st Workshop on Abusive Language Online, ACL 2017*. Vancouver: Canada, pp 46—51.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, pp. 7—36.

Kilgarriff. A. (2009). Simple maths for keywords. In Mahlberg, M., González-Díaz, V. and Smith, C. (Eds.), *Proceedings of Corpus Linguistics Conference CL2009.* United Kingdom: University of Liverpool.

Ljubešić, N., Fišer, D. and Erjavec, T. (2019). *The FRENK datasets of Socially Unacceptable Discourse in Slovene and English*. https://arxiv.org/abs/1906.02045.

Millar, S., Baider, F. H. and Assimakopoulos, S. (2017). Harvesting and Analysing Online Comments to News Reports. In Assimakopoulos, S., Baider, F. H. and Millar, S. (Eds.), *Online Hate Speech in the European Union. A Discourse-Analytic Perspective.* Cham: Springer International Publishing, pp. 17—21.

Page, R., Barton, D., Unger, J. W. and Zappavigna, M. (2014). *Language and Social Media. A Student Guide.* London and New York: Routledge.

Pahor de Maiti, K., Fišer, D. and Ljubešić, N. (2019). How haters write: an orthographic analysis of online hate speech. *7th Conference on CMC and Social Media Corpora for the Humanities*. Cergy-Pontoise, France. Submitted.

Rayson, P., Archer, D., Piao, S. and McEnery, A. M. (2004). The UCREL semantic analysis system. *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*. Lisbon, Portugal.

Computer-mediated versus non-computer-mediated corpora of informal French: Differences in politeness and intensification in the expression of contrast by *au contraire*

Jorina Brysbaert[1,2], Karen Lahousse[1]
[1]KU Leuven, [2]Research Foundation - Flanders (FWO)
E-mail: jorina.brysbaert@kuleuven.be, karen.lahousse@kuleuven.be

Abstract

In this paper, we will present a case study of linguistic exploitation of a computer-mediated written French corpus (*Yahoo Contrastive Corpus of Questions and Answers*) and a non-computer-mediated spoken French corpus (*Discours sur la ville : Corpus de Français Parlé Parisien des années 2000*). We will observe that, although both corpora contain similar (informal) spontaneous conversations, there is an intriguing difference between them with respect to two key features of the contrastive adverb *au contraire*. More precisely, we will show that in the computer-mediated written corpus, *au contraire* is more often used with a corrective interpretation and is more often intensified by means of an adverb such as *bien* than in the non-computer-mediated spoken corpus. This finding suggests that users of computer-mediated language correct their (anonymous) interlocutors more explicitly than speakers in non-computer-mediated direct conversations, which indicates that there are underlying differences in politeness and intensification between both corpora.

Keywords: contrastivity, informal French, politeness

1. Introduction

In this paper, we compare the use of the contrastive marker *au contraire* 'on the contrary' in a corpus of written computer-mediated communication (*Yahoo Contrastive Corpus of Questions and Answers* (De Smet, 2009)) and a corpus of spoken non-computer-mediated communication (*Discours sur la ville : Corpus de Français Parlé Parisien des années 2000* (Branca-Rosoff et al., 2011; Branca-Rosoff et al., 2012)). These corpora are similar in that they both contain (quite informal) spontaneous language in the form of question-answer conversations, but differ in that the contact is direct (CFPP2000) or computer-mediated (YCCQA). We will show that, with respect to two fundamental features of *au contraire*, there is an important difference between the two corpora, which hints at underlying differences in politeness and intensification. We will first focus on the discourse-semantic type of contrast that is typically believed to be expressed by *au contraire*, i.e. a corrective contrast. The use of *au contraire* with a corrective interpretation is more frequent in the computer-mediated written corpus than in the non-computer-mediated spoken corpus. Next, we will show that the intensification of *au contraire* by means of an adverb such as *bien*, *tout* or *même*, occurs more often in the computer-mediated written corpus than in the non-computer-mediated spoken corpus.

2. Methodology

In what follows we briefly present the two corpora: the *Yahoo Contrastive Corpus of Questions and Answers* (YCCQA), illustrating informal written computer-mediated French (2.1), and the *Discours sur la ville : Corpus de Français Parlé Parisien des années 2000* (CFPP2000), representing informal spoken non-computer-mediated French (2.2).

2.1 Yahoo Contrastive Corpus of Questions and Answers

The *Yahoo Contrastive Corpus of Questions and Answers* (YCCQA) has been compiled by Hendrik De Smet (KU Leuven) and contains language data from the website https://answers.yahoo.com/. This website consists of an online discussion forum where questions can be asked and answered. The topics of the questions are very diversified, ranging from education and politics to music and humor, which implies that different genres are represented. Importantly, we believe that the YCCQA can be placed somewhat in the middle on the continuum between spoken language (i.e. *language of immediacy* in the terms of Koch & Oesterreicher (1985, 2007)) and written language (i.e. *language of distance*)[1]. With respect to certain parameters distinguished by Koch & Oesterreicher (1985, 2007), the YCCQA can be situated towards the *language of immediacy* pole. The texts in this corpus are all of the type question-answer and can therefore be considered to form a kind of 'written dialogues'. In addition, conversation topics are more or less freely chosen and language use is quite spontaneous. With respect to other parameters however, the YCCQA clearly tends towards the *language of distance* pole: it contains conversations between anonymous interlocutors, in a setting characterized by spatio-temporal distance. All in all, although the YCCQA contains written data, the texts are to some extent similar to spontaneous, informal spoken-like language, as is illustrated in example (1)[2]:

[1] Note that according to some linguists, it is not straightforward to apply the model of Koch & Oesterreicher to computer-mediated communication (e.g. Dürscheid, 2016).

[2] The texts in the YCCQA contain a lot of spelling errors, which we did not correct.

(1) - *Pour être "une vrai femme" faut il absolument devenir mère ?*
*- Absolument pas **au contraire** je dirais même, car le fait d'être mère casse l'image même de la femme telle que l'homme la conçoit, c'est à dire une amante !!!* (YCCQA)
'- In order to be "a real woman", does one absolutely have to become a mother?
- Not at all **on the contrary** I would even say, since the fact of being a mother breaks the very image of the woman as a man conceives it, i.e. a lover !!!'

The questions and answers in the YCCQA all date from after 2006. The corpus contains about 6,1 million words and is available as a TXT-file. Examples of *au contraire* were extracted using the concordancer AntConc[3].

2.2 Corpus de Français Parlé Parisien

The *Corpus de Français Parlé Parisien des années 2000* (CFPP2000) has been compiled on the initiative of Sonia Branca-Rosoff at the university of Paris 3-Sorbonne-Nouvelle (Branca-Rosoff et al., 2011; Branca-Rosoff et al., 2012) and consists of interviews conducted with inhabitants of Paris and the surrounding suburbs. The interviewees were questioned on their relationship to the district in which they live and on their life in general. Given the interview setting with questions and answers, the language data in the CFPP2000 are to a certain extent similar to the written conversations in the YCCQA. The CFPP2000 indeed also contains quite spontaneous dialogues, with more or less freely developing conversation topics. However, in contrast to the YCCQA, the CFPP2000 consists of face-to-face interactions between interlocutors who are not completely unknown to each other. This means that, based on the parameters distinguished by Koch & Oesterreicher (1985, 2007), the CFPP2000 should be placed closer towards the *language of immediacy* pole than the YCCQA. An example of a conversation in the CFPP2000 is given in (2):

(2) *spk 1: c'est quoi ou qui ?*
spk 2: ben mon prof de de français de troisième + voilà [rire de Julie]
spk 1: qu'est-ce que c'est alors un prof charismatique ?
spk3 spk2: [1] oui [rire de Katia] [2] non c'est un prof
spk2: malade qui tapait sur son bureau mais [rire d'Amélie] non mais il était vraiment fou enfin il avait des tics des
*spk1: ah non c'est l'horreur **au contraire*** (CFPP2000)
'spk 1: what or who is it ?
spk 2: well my French teacher of the third year + so [laughter of Julie]
spk 1: what does that mean then, a charismatic teacher ?
spk3 spk2: [1] yes [laughter of Katia] [2] no it is a teacher
spk2: mad [teacher] who was typing on his desk but [laughter of Amélie] no but he was really crazy anyway he had tics
spk1: oh no that's horror **on the contrary**'

The first interviews in the CFPP2000 date from 2005, but new interviews are still regularly added. Recordings as well as transcriptions of the interviews are available online, on the site http://cfpp2000.univ-paris3.fr/. The website also offers a search engine to search through the corpus, but we downloaded the transcriptions and used the concordancer AntConc to extract examples of *au contraire*. At the moment of extraction (2017), the CFPP2000 contained about 650.000 words.

3. Use of *au contraire* in YCCQA and CFPP2000

In this section, we present an analysis of two properties that have considered to be characteristic of the contrastive adverb *au contraire* – as opposed to other contrastive adverbs such as *par contre* and *en revanche* – but that have not yet been tested on the basis of a systematic corpus analysis: *au contraire* (i) is typically used in corrective contrastive relations (3.1) and (ii) can be intensified by means of an intensifying adverb such as *bien*, *tout* or *même* (3.2). We will also establish a correlation between both properties, by showing that the intensification of *au contraire* is characteristic of corrective contexts (3.3).

3.1 *Au contraire* as a marker of correction

It has often been suggested that *au contraire* is an adverb specializing in the marking of corrective contrastive relations (Csűry, 2001; Danjou-Flaux, 1980; Danjou-Flaux, 1983; Masseron & Wiederspiel, 2003). In this type of contrastive relation, the second part of the contrast acts as a correction or replacement of the first part. In example (3) for instance, the verb *diviser* 'to divide' corrects or replaces the verb *aider* 'to help'. Formally, corrections are characterized by the presence of a negation in the first part of the contrastive relation (Csűry, 2001; Danjou-Flaux, 1980; Danjou-Flaux, 1983). This negation (underlined in the next examples) can be realized by a morpho-grammatical negation marker (3) or by a lexical element, as an adjective with negative interpretation (4):

(3) *- Par ces temps de grands froids pourquoi n'ouvre t'on pas les portes des mosquées et des églises ? pour les SDF*
*- Parce qu'ils sont tous egoistes, ils se rassurent avec leur religion mais ca n'aide personne **au contraire** ca divise les peuples.* (YCCQA)
'- With this very cold weather, why don't they open the doors of the mosques and the churches? for the homeless
- Because they are all egoists, they reassure themselves with their religion but it does not help anybody **on the contrary** it divides people.'

[3] Note that the YCCQA is in fact a multilingual corpus: it contains texts in French (France) (6,1 million words), English (United Kingdom & Ireland) (10,1 million words), German (Germany) (5,8 million words) and Spanish (Spain) (7,4 million words). For the analysis of the adverb *au contraire* presented in this paper, we obviously only used the French sub-corpus.

(4) *Les viandes maigres, poissons et crustacés, sont à consomer avec modération, car la graisse qu'il contient […] devient du gras trans donc cancérigène et inutilisable pour l'organisme,* ***au contraire****, ca lui donne du travail d'élimination.* (YCCQA)
'Lean meat, fish and shellfish have to be consumed with moderation, since the fat it contains […] becomes trans fat thus carcinogenic and unusable for the organism, **on the contrary**, it gives him elimination work.'

We used the presence of an element with a negative interpretation as a criterion to determine which occurrences of *au contraire* in our corpora are corrective. More precisely, examples in which the first part contains a negation were considered to be corrective (see (3)-(4)), whereas examples in which the first part does not contain a negation were counted as non-corrective (5):

(5) *Ben déjà j'y suis allé très tôt tout seul (mm) vers six sept ans (mm) voilà parce que ma mère travaillait tôt + et mon père lui il travaillait très tard* ***au contraire*** *donc il se (mm) réveillait pas (mm) donc euh j'y allais tout seul quoi* (CFPP2000)
'Well I was going there alone already very early (mm) around six seven years (mm) that's because my mother was working early + and my father, he was working very late **on the contrary** so he did (mm) not wake up (mm) so uh I was going there alone'

Given the hypothesis in the linguistic literature, we expected to find a high frequency of corrective *au contraire* in our two corpora. It appears from Table 1 that, in general, this prediction is borne out: as the total of both corpora shows, *au contraire* often expresses a correction (69,9%). However, our analysis also reveals an important difference between the YCCQA and the CFPP2000. In the YCCQA, *au contraire* is very often found in a correction (73,9%), whereas in the CFPP2000, *au contraire* occurs more frequently in a non-corrective context (55,9%) than in a corrective context (44,1%). The proportion of corrective uses of *au contraire* is thus greater in the computer-mediated YCCQA than in the non-computer-mediated CFPP2000, and this difference is statistically significant, χ^2 (1, N = 509) = 23.50, p < .001, with a small effect size, Cramer's V = 0.22.

	YCCQA	CFPP2000	Total
Corrective	326 (73,9%)	30 (44,1%)	356 (69,9%)
Non-corrective	115 (26,1%)	38 (55,9%)	153 (30,1%)
Total	441 (100%)	68 (100%)	509 (100%)

Table 1: Absolute and relative frequencies of *au contraire* in corrective versus non-corrective contexts in the YCCQA and the CFPP2000.

We therefore hypothesize that the degree of corrections expressed by *au contraire* depends at least partly on the medium (written computer-mediated versus spoken non-computer-mediated). In the context of an interview (CFPP2000), with a direct contact with the interlocutor, speakers could prefer not to correct their interlocutor too often and too explicitly by means of *au contraire*, for reasons of politeness. On an online discussion forum (YCCQA), where most interlocutors are anonymous, politeness might be less important and reactions might be more direct. Note however that the difference between the YCCQA and the CFPP2000 with respect to the corrective use of *au contraire* might be influenced by other properties of the corpora as well, such as the possibility to reinforce contrast prosodically in the spoken CFPP2000 versus the absence of prosody in the written YCCQA, or the fact that questions and answers on the Yahoo discussion forum are evaluated, which creates competition between the users.

3.2 Intensification of *au contraire*

In the linguistic literature, it has been pointed out that the adverb *au contraire* can be intensified by means of the intensification adverbs *bien* (6), *tout* (7) and *même* (Csűry, 2001; Flaux, 2003; Masseron & Wiederspiel, 2003):

(6) *Je vais te le répeter pour la Xème fois : les féministes ne sont pas contre les hommes, bien* ***au contraire****.* (YCCQA)
'I'll repeat it for the Xth time: feminists are not against men, quite **the contrary**.'

(7) *On pourrait même dire que de telles personnes ne possèdent pas beaucoup d'humilité. Tout* ***au contraire****, ils sont vraiment trempé dans du pur orgueil.* (YCCQA)
'One could even say that such people do not have much humility. Quite **the contrary**, they are really soaked in pure pride.'

According to Csűry (2001), these combinations are not frequent in formal French, but nothing is known about their use in informal French. Hence, we examined the frequency of intensified *au contraire* in our two informal French corpora. As becomes clear from Table 2, *au contraire* is in general more often non-intensified (82,3%) than intensified (17,8%), but again, the results for the YCCQA are different from those for the CFPP2000. The YCCQA contains 89 examples (20,2%) in which *au contraire* is intensified, most of them (i.e. 81) with the adverb *bien*, while in the CFPP2000, there is only one example (1,5%) of *bien au contraire*. The intensification of *au contraire* is thus more frequent in the informal written computer-mediated YCCQA than in the informal spoken non-computer-mediated CFPP2000, and this difference is statistically significant, χ^2 (1, N = 509) = 12.91, p < .001. The effect size of this association is rather small, Cramer's V = 0.17.

	YCCQA	CFPP2000	Total
Intensification	89 (20,2%)	1 (1,5%)	90 (17,8%)
No intensification	352 (79,8%)	67 (98,5%)	419 (82,3%)
Total	441 (100%)	68 (100%)	509 (100%)

Table 2: Absolute and relative frequencies of intensified versus non-intensified *au contraire* in the YCCQA and the

CFPP2000.

Importantly, this result also highlights the fact that some specific linguistic features are register-dependent, and that it is interesting to include CMC-corpora in analyses of such linguistic features.

3.3 Intensification of *au contraire* in corrective versus non-corrective contexts

Given that the intensification of *au contraire* has received almost no attention in the linguistic literature, it also remains unclear in which linguistic contexts this adverb is typically combined with an intensifying adverb. We therefore examined whether the intensification of *au contraire* could be related to its occurrence in a (non-)corrective context. The results are given in Tables 3 and 4. With respect to the YCCQA (see Table 3), we observe that there indeed seems to be a link between intensification and corrective context. In this corpus, the adverb *au contraire* is more often intensified in a corrective (25,8%) than in a non-corrective context (4,3%), and this difference is statistically significant, $\chi^2(1, N = 441) = 22.90$, $p < .001$, with a small effect size, Cramer's $V = 0.23$. This result suggests that in informal written computer-mediated French, language users might feel a certain need to intensify the adverb *au contraire* when expressing a correction. This correlation is not surprising, since the use of an intensifying adverb can be a way to put more emphasis on the correction and hence, on the message the user of the discussion forum wants to convey to an interlocutor who is not present in the speech context. This finding might also explain the quasi-absence of intensified *au contraire* in the CFPP2000 (see Table 2). In section 3.1, we showed that there are less corrections in the CFPP2000 than in the YCCQA. Given that the intensification of *au contraire* is especially frequent in corrections (see Tables 3 and 4) and that the CFPP2000 contains less of these corrections (see Table 1), it logically follows that *au contraire* is less often intensified in the CFPP2000. Finally, note that the association between intensification and corrective context is not statistically significant in the CFPP2000, Fisher exact test (1, $N = 68$), $p = 0.44$, which is probably due to the low overall frequency of intensified *au contraire* in this corpus (see Table 4)[4].

	Corrective	Non-corrective	Total
Intensification	84 (25,8%)	5 (4,3%)	89 (20,2%)
No intensification	242 (74,2%)	110 (95,7%)	352 (79,8%)
Total	326 (100%)	115 (100%)	441 (100%)

Table 3: Absolute and relative frequencies of intensified versus non-intensified *au contraire* in corrective versus non-corrective contexts in the YCCQA.

	Corrective	Non-corrective	Total
Intensification	1 (3,3%)	0 (0%)	1 (1,5%)
No intensification	29 (96,7%)	38 (100%)	67 (98,5%)
Total	30 (100%)	38 (100%)	68 (100%)

Table 4: Absolute and relative frequencies of intensified versus non-intensified *au contraire* in corrective versus non-corrective contexts in the CFPP2000.

4. Conclusion

In this paper, we analyzed two properties of the contrastive adverb *au contraire* in two different corpora: the YCCQA (informal written computer-mediated French) and the CFPP2000 (informal spoken non-computer-mediated French). With respect to the first property, the use of *au contraire* in a corrective context, we observed that *au contraire* occurs more frequently in a correction in the YCCQA than in the CFPP2000. As for the second property, the intensification of *au contraire* by means of an intensifying adverb such as *bien*, similar results were found: the YCCQA contains more cases of intensified *au contraire* than the CFPP2000. We also showed that these two properties are related: the intensification of *au contraire* is especially frequent in corrections. In sum, although the YCCQA and the CFPP2000 contain comparable language data (i.e. informal French in question-answer format), there are important differences with respect to the use of the adverb *au contraire*, suggesting that users of computer-mediated language often explicitly correct their (anonymous) interlocutors, which is less the case in direct (i.e. non-computer-mediated) conversations. It would of course be interesting to further test this hypothesis on other CMC- and spoken corpora, to examine whether the difference applies to all types of computer-mediated versus non-computer-mediated communication or whether it is specific to the language used on a discussion forum versus in an interview setting. More in general, since the properties of *au contraire* that we discussed especially seem to be frequent in computer-mediated language, this paper also shows the need to include CMC-corpora in analyses that consider the register-dependency of specific linguistic features.

5. References

Anthony, L. (2014). *AntConc Version 3.4.4 [Computer Software]*. Tokyo: Waseda University. http://www.laurenceanthony.net/.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., Pires, M. (2011). Constitution et exploitation d'un corpus de français parlé parisien. *Corpus*, 10, pp. 81--98.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., Pires, M. (2012). Discours sur la ville: Corpus de Français Parlé Parisien des années 2000 (CFPP2000). http://cfpp2000.univ-paris3.fr/CFPP2000.pdf.

[4] We performed a Fisher exact test instead of a chi-square test, since some of the expected values were smaller than 5.

Csűry, I. (2001). *Le champ lexical de mais: Étude lexico-grammaticale des termes d'opposition du français contemporain dans un cadre textologique.* Debrecen: Kossuth Egyetemi Kiadó.

Danjou-Flaux, N. (1980). Au contraire, par contre, en revanche: Une évaluation de la synonymie. In *Synonymies*. Lille: Presses universitaires de Lille, pp. 123--148.

Danjou-Flaux, N. (1983). Au contraire, connecteur adversatif. In *Connecteurs pragmatiques et structure du discours*. Genève: Université de Genève, pp. 275--303.

De Smet, H. (2009). *Yahoo Contrastive Corpus of Questions and Answers*. Leuven: Department of Linguistics, University of Leuven.

Dürscheid, C. (2016). Nähe, Distanz und neue Medien. In H. Feilke, & M. Hennig (Eds.), *Zur Karriere von 'Nähe und Distanz'. Rezeption und Diskussion des Koch-Oesterreicher-Modells.* Berlin: De Gruyter, pp. 357--385.

Flaux, N. (2003). Au contraire (de) et le sens de contre. In P. Péroz (Ed.), *Contre: Identité sémantique et variation catégorielle*. Metz: Université de Metz, pp. 289--309.

Koch, P., Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36, pp. 15--43.

Koch, P., Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für Germanistische Linguistik*, 35(3), pp. 346--375.

Masseron, C., Wiederspiel, B. (2003). Contrastivité adverbiale: Au contraire, contrairement à, par contre. In P. Péroz (Ed.), *Contre: Identité sémantique et variation catégorielle*. Metz: Université de Metz, pp. 311--341.

Productivity of Anglicism Bases in Hyphenated German Compounds

Steven Coats[1], **Adrien Barbaresi**[2]
[1]University of Oulu, Finland, [2]Berlin-Brandenburg Academy of Sciences, Germany
steven.coats (at) oulu.fi, barbaresi (at) bbaw.de

Abstract

This study investigates hyphenated German compounds that contain English constituents, a part of the German lexicon that exhibits great diversity. We determine the frequency of a set of English noun bases in different constituent positions in three corpora of online language, and quantify the productivity of bases in these compounds using a metric based on Shannon entropy, a measure of information content. Compound entropy values for English bases reflect their diversity of use, and are unequally distributed for left-hand, internal, and right-hand constituents. The semantics of the base types with the highest frequencies and entropy values reflect contemporary cultural and technological concerns. Differences in entropy according to constituent position may be an indication of word class conversion of anglicisms in German.

Keywords: Anglicisms, German, Compounds, Corpus linguistics, Productivity, Entropy

1. Introduction

English is the most important source language for new words in German at present, accounting for the majority of lexical borrowings (Onysko, 2007). The prevalence of English in numerous public, job-related and international contexts in German society ensures a broad diffusion of language knowledge, so that a bottom-up diffusion of English-based neologisms and compounds is possible, in a sharp contrast to previous patterns of foreign lexical assimilations, for example of Latin or French words (Krome & Roll, 2017). Anglicisms exhibit a great deal of variation in written German: They can be integrated with minimal modifications (e.g. *Discounter*), assimilated in morphology and orthography and adapted to German inflectional paradigms (e.g. *geliked* or *gelikt*, 'liked on social media') (Burmasova, 2010; Coats, 2018), or serve as the basis for calques (loan translations) and syntactic constructions. Recent research has cataloged their overall diversity (Eisenberg, 2011, 2013), investigated the extent to which they overlap in meaning with existing German lexemes (Onysko & Winter-Froemel, 2011; Winter-Froemel, Onysko, & Calude 2011), or documented their assimilation to German orthography and inflectional morphology (Coats, 2018). English elements can also be joined to German constituents with hyphens to create compounds (e.g. *Urlaubs-Feeling* 'holiday feeling', *Service-Zentrum* 'service center'), a productive process which is the focus of the present research.

The processes by which composite words in German can be formed by combining indigenous and exogenous lexical material have been described in the literature (Fleischer & Barz, 2012), and the importance of relative frequencies of bases for the recognition and processing of compounds has been shown in experiments (e.g. Lüdeling & de Jong, 2002; Baayen, Wurm, & Aycock, 2007), but the productivity of English bases in hyphenated compounds has not yet been a focus of German lexicography. Most research into productivity in German has focused on the productivity of affixes, rather than word bases (Lüdeling & Evert, 2005; Lüdeling, Evert & Heid, 2000). Recent developments have highlighted the importance of corpus-based methods as well as the need to extract relevant information and adequately describe changes or findings based on the explanatory power of statistical indicators (Hein & Engelberg, 2017). It is indeed still necessary to find a suitable methodology to study the dynamics of anglicisms in German, all the more since empirical frequency-based investigations have not always been the main research focus (Burmasova, 2010). In this study, we consider the productivity of English base constituents in hyphenated German compounds. By taking into account recent web and computer-mediated communication (CMC) corpora, we hope to capture phenomena unseen in standard written German, as these corpora consist of genres which are expected to differ from commonly accepted rules. Indeed, we do not stick to the concept of rules but rather try to derive norms from empirical data (Habert & Zweigenbaum, 2002) by way of statistical indicators which are related to information theory and as such yield a particular view on language constituency and productivity.

Building upon quantitative approaches to the measure of morphological productivity developed by Baayen (1994a, 1994b, 2001) and others (Hay & Baayen, 2003; Moscoso del Prado Martín, Kostić, & Baayen, 2004), we utilize Shannon entropy (Shannon, 1948) to measure the productivity of English constituents in different internal word positions in large corpora from the web and from Twitter. In light of findings from response latency experiments, this may be evidence that English constituents increasingly take part in productive word formation processes in German. In addition, the semantic values of the most productive English bases may shed light on broader developments in the German lexicon and in German-speaking society as a whole. The study addresses the following questions: 1) Which constituent base anglicism types are most frequent in hyphenated German compounds, and 2) What can morphological diversity measures such as Shannon entropy tell us about the dynamics of anglicisms borrowed into German compounds?

In the next section, a review of previous research in morphological productivity is provided, followed by a brief overview of German usage norms for hyphenated compounds. In Section 3, the data and methods used to calculate

Shannon entropy from the corpora are presented and German compounds briefly reviewed. Section 4 presents the results, and Section 5 closes the paper with a summary and outlook for future research.

2. Previous work

2.1. Productivity measures

Baayen (1993, 1994a, 1994b, 2001) proposed several measures of morphological productivity, including the category-conditioned degree of productivity: the ratio of *hapax legomena* (words that occur once in a text or a corpus) for an affix to the size of its morphological category, which represents that the probability a new word encountered in a text or corpus will be a type that has not yet been encountered, given that it belongs to a particular morphological class. For example, for the German deadjectival nominal suffix *-keit* (e.g. *Sparsamkeit* 'thriftiness'), the category-conditioned degree of productivity is the ratio of the sum of all *hapax* words ending in *-keit* to the sum of all words ending in *-keit*. Because *-keit* is more productive than suffixes such as *-nis* or *-tum*, it will have a higher value when using this measure.

For word bases in compounds, frequencies can be assessed by the morphological family size (the number of distinct types in which a base appears) and the cumulative family frequency (the sum of token frequencies for all types in which a base appears). Morphological family size is negatively correlated to reaction times in lexical recognition experiments (Baayen, Wurm, & Aycock, 2007). For example, because a German base like *Schrift* 'writing' may occur in a large number of compound words (e.g. *Schreibschrift* 'handwriting', *Schriftsteller*, 'writer', *Unterschrift* 'signature', etc.) with relatively high frequencies, compounds containing the base are recognized more quickly than are compounds that contain constituents with lower frequencies or smaller family sizes, such as *Schund* 'rubbish', which is a constituent in a smaller number of words (e.g. *Schundliteratur* 'trashy writing').

Like morphological family size, cumulative family frequency has been shown to correlate negatively with reaction times in lexical recognition experiments (Baayen & Hay, 2002; Baayen, Lieber, & Schreuder, 1997; De Jong, Schreuder, & Baayen, 2000; Schreuder & Baayen, 1997). Hay (2001) found that for English compounds, frequent words are processed more quickly, and thus likely to be stored in the mental lexicon as opaque single units of meaning, whereas infrequent compounds may be stored as decomposable items. For German verbs, Lüdeling and De Jong (2002) found a negative relationship between morphological family size and response latencies in an experimental task.

Moscoso del Prado Martín, Kostić, and Baayen (2004) utilized a metric based on Shannon entropy to calculate an "information residual" for a word, or the difference between the logarithm of a word's frequency to the Shannon entropy for all inflected forms of the word. They found that in regression models of word response latencies, the word information residual provides a better fit than morphological family size or cumulative family frequency, meaning that from a statistical standpoint this additional indicator yields more fine-grained information and is thus suitable to draw conclusions on lexical use.

2.2. Hyphenation in German compounds

In standard German orthography, constituent elements in composite words are typically linked without a hyphen. Hyphens can be optionally used in composite words in order to emphasize particular constituents or to enhance the legibility of long composite words with multiple constituents (Duden, 2006, p. 39; Fleischer & Barz, 2012, p. 193). Hyphenation is recommended if a composite form contains a constituent that is an abbreviation or initialism (*Fussball-WM* 'football/soccer world cup'), and is preferred if the first constituent element of a compound is a proper noun, especially a personal name (Fleischer & Barz, 2012, p. 193; e.g. *Merkel-Regierung* 'Merkel government'). Hyphenation is also used in longer composite phrasal word forms (*Pro-Kopf-Verbrauch* 'per capita use') (Duden, 2006, p. 41; Fleischer & Barz, 2012, p. 175). Fleischer and Barz note that composite word formations in German can incorporate foreign constituents as first elements, internal elements, or final elements, "without restrictions" (2012, p. 111).

3. Data and methods

A list of potential English constituents was created by combining the 3,262 most common nouns in the British National Corpus (Kilgarriff, 1997) with the 10,000 most common nouns in the 9.6b-token ENCOW16ax corpus, a corpus of English texts from the web (Schäfer, 2015; Schäfer & Bildhauer, 2012), then converting to lower case and removing types containing punctuation or shorter than 4 characters. The list thus combines attested data from a stratified reference corpus and more current utterances from a large web sample. The frequencies of these 8,313 unique types as left-hand, central, or right-hand constituents in hyphenated German words were determined in three corpora of online German: A German Twitter corpus of 534m tokens (Coats, 2018), the DECOW16bx corpus, a German web corpus of 11b tokens (Schäfer, 2015; Schäfer & Bildhauer, 2012), and a corpus of German WordPress blogs with 2.1b tokens (Barbaresi, 2016). In order to account for non-standard capitalization (common on Twitter and in other informal online genres), all words were converted to lowercase. A regular expression was used to additionally target plural and genitive forms while taking potentially unknown forms into account.

For each of the 8,313 English potential base types, we used token counts for individual compounds and the cumulative family frequency for the base (i.e. the summed frequencies of all compound types containing the base) to calculate an entropy measure. The compound Shannon entropy of a base can be calculated according to the formula

$$H(B) = -\sum_{i=1}^{n} \frac{F(x_i)}{F(B)} \cdot log_2 \frac{F(x_i)}{F(B)} \qquad (1)$$

where B represents an anglicism base, $F(x_i)$ the frequency of a particular compound type containing B, n the morphological family size for the base, and $F(B)$ the cumulative family frequency for the base. The value can range from zero to $\log_2 n$. As an example, the entropy for the English base *payment* can be calculated in a corpus of 8 tokens in which *Payment-Taste* 'payment button' occurs 5 times, *Crypto-Payment-App* 'crypto payment app' twice, and *Online-Payment* 'online payment' once. In this example, the English base 'payment' occurs as a left-hand constituent, as an internal constituent, and as a right-hand constituent. Entropy can be calculated by constituent position, or a total entropy score can be calculated that takes into account all possible configurations: In this example, the left-hand, internal, and right-hand compound entropies would be zero (because only one type occurs at each word-internal position), while the total entropy would be 1.30. In general, low entropy values indicate skewed frequency distributions, while high values indicate more uniform distributions.

4. Results and discussion

The corpora feature a large number of hyphenated compounds containing anglicisms: the Twitter corpus 619,338 unique types, the most frequent of which are *youtube-video*, *live-tracking*, and *start-up*; the DECOW16bx corpus 6,567,984 types (*online-shop*, *html-code*, and *bb-code* are the most frequent) and the WordPress corpus 808,648 distinct types (*us-dollar*, *online-shop*, and *after-sales-service* are the most frequent).

Tables 1, 2, and 3 show the base types with the highest cumulative family frequencies for the three corpora, their frequencies, left-hand, internal, right-hand and total entropy calculations for the types, and the three most frequent types containing the base elements.

Many of the most frequent bases in the three corpora denote entities related to technology or the internet (*video*, *twitter*, *facebook*, *youtube*, *blog*, *internet*, *code*, *software*). Left-hand, internal, right-hand, and total entropy values (indicated by H_L, H_I, H_R, and H_T in the tables) provide an overall indication of the diversity of the frequency distributions for compounds containing bases in that constituent position. Entropy values for the bases in different hyphenated word positions vary from low (*dollar* in right-hand position) to high (*team*, *system* in right-hand position). Lower entropy values can indicate that a base has been lexicalized in a hyphenated word (*us-dollar*), resulting in far higher frequencies of that type compared to other types in the morphological family. For the Twitter and DECOW16bx corpora, internal entropy values are the lowest, right-hand entropies intermediate, and left-hand entropies the highest. For the WordPress blogs corpus, right-hand entropy values are highest.[1] Because the base types considered in the study are primarily English nouns, this value may also provide a preliminary indication of word class conversion of borrowings (Figure 1). This possibility, however, needs further exploration, for example by comparison with English-language compounds.

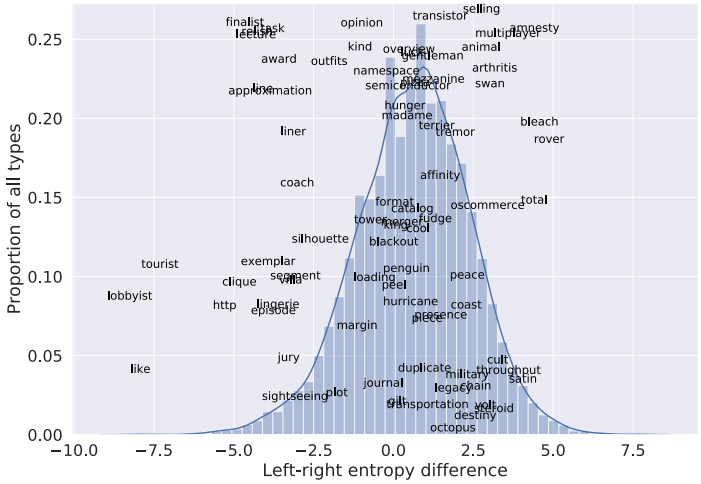

Figure 1: Left-right entropy difference distribution, DECOW16bx corpus

Figure 2 shows the total compound morphological entropy versus the morphological family size for the Twitter corpus, the DECOW16bx corpus, and the WordPress blogs corpus. In the plots, each point represents an English base, the red line represents the maximum entropy, and the magenta line represents a nonparametric locally-weighted regression.

For many of the 8,313 English-language bases analyzed in the study, entropy values are close to the maximum possible entropy curves in the three corpora, which can be seen as an indication of the relative lack of syntactic or semantic constraints on hyphenated word formation in German. Comparing the regression curves (magenta lines in the figures) to the maximum entropy curves (red lines) shows overall higher entropy values for the WordPress blogs corpus and for the Twitter corpus, compared to the DECOW16bx corpus. This is likely due to the more informal nature of written language on Twitter and in blogs, genres that exhibit relatively high rates of creative textual features such as non-standard orthography, expressive lengthening, or emoticon and emoji use (Argamon et al., 2007; Coats, 2016), and that therefore may also exhibit a greater diversity of hyphenated compound types. In contrast, many of the genres that comprise the DECOW16bx corpus, such as news articles, consist of relatively formal, conventionalized writing in which non-standard usages are uncommon.

The bases with the highest total entropy values in the Twitter corpus are the forms *team* and *chef*, followed by *media*, *twitter*, and *video*. For the web corpus, the types with the highest total entropy are *team* and *system*, followed by *forum*, *software*, and *service*. For the blogs cor-

[1] Because the anglicism constituents are nouns and in German compound nouns are almost exclusively right-headed, it is conceivable that right-hand entropies may be lower. The reason for the reversal of this pattern in the WordPress blogs corpus is unknown at this stage. Although different text processing procedures may play a role, compounding creativity may explain this behavior.

Table 1: Most frequent types, Twitter Corpus

Base	Freq.	H_L	H_I	H_R	H_T	Most freq. types
video	29412	8.72	7.98	6.91	8.67	‘youtube-video’, 7285, ‘video-interview’, 803, ‘youtube-videos’, 763
twitter	27459	8.99	8.96	7.58	9.25	‘twitter-account’, 2602, ‘twitter-zufallsstory’, 1510, ‘twitter-app’, 772
facebook	23901	8.40	8.14	5.98	8.56	‘facebook-seite’, 2740, ‘facebook-gruppe’, 751, ‘facebook-fans’, 606
team	18113	8.63	6.90	10.68	11.02	‘orga-team’, 553, ‘dfb-team’, 335, ‘social-media-team’, 305
chef	17818	7.81	4.90	9.63	9.88	‘spd-chef’, 573, ‘fdp-chef’, 452, ‘ex-chef’, 356
news	17553	6.98	7.83	5.91	7.02	‘heise-news’, 1925, ‘it-news’, 1812, ‘fake-news’, 1189
youtube	15435	4.71	6.44	5.61	4.87	‘youtube-video’, 7285, ‘youtube-kanal’, 1312, ‘youtube-videos’, 763
blog	15396	5.78	3.85	8.88	8.70	‘blog-eintrag’, 1212, ‘blog-beitrag’, 838, ‘blog-artikel’, 668
marketing	14121	8.64	8.42	4.80	8.02	‘online-marketing’, 2349, ‘content-marketing’, 1339, ‘influencer-marketing’, 414
media	14097	8.34	8.95	2.45	9.38	‘social-media’, 1051, ‘social-media-team’, 305, ‘social-media-marketing’, 286

Table 2: Most frequent types, DECOW16bx Corpus

Base	Freq.	H_L	H_I	H_R	H_T	Most freq. types
system	582231	9.91	9.68	11.97	12.22	‘it-system’, 16797, ‘erp-system’, 11115, ‘content-management-system’, 9506
internet	507073	8.90	10.62	6.33	9.15	‘internet-seite’, 36068, ‘internet-adresse’, 19172, ‘internet-auftritt’, 16666
forum	412839	8.96	10.27	10.24	10.63	‘feuerwehr-forum’, 17621, ‘fan-forum’, 10562, ‘hifi-forum’, 9812
code	398698	8.46	9.61	3.57	4.17	‘html-code’, 151710, ‘bb-code’, 132344, ‘qr-code’
team	359205	9.31	10.36	12.64	12.85	‘top-team’, 5455, ’orga-team’, 3963, ’support-team’, 3307
shop	344979	6.51	8.66	5.22	6.07	‘online-shop’, 171998, ‘internet-shop’, 9308, ‘shop-system’, 5905
version	337625	7.07	7.67	9.85	9.89	‘beta-version’, 13907, ‘pc-version’, 10173, ‘windows-version’, 9784
software	325250	8.07	10.93	10.55	10.47	‘software-entwicklung’, 9605, ‘software-update’, 7840, ‘software-lösung’, 7746
service	321222	8.12	9.81	9.65	10.46	‘full-service’, 6824, ‘it-service’, 6728, ‘service-center’, 6591
video	254431	9.32	10.80	8.53	10.27	‘youtube-video’, 15269, ‘hd-video’, 5563, ‘video-kritik’, 5517

Table 3: Most frequent types, WordPress blogs corpus

Base	Freq.	H_L	H_I	H_R	H_T	Most freq. types
blog	47324	6.86	9.07	11.86	11.84	‘blog-post’, 743, ‘satire-blog’, 555, ‘blog-event’, 522
shop	39271	6.29	8.02	4.91	5.18	‘online-shop’, 13952, ‘online-shops’, 9930, ‘web-shops’, 2969
team	31366	7.16	7.97	12.00	12.10	‘orga-team’, 659, ‘dream-team’, 239, ‘blog-team’, 224
system	26528	6.81	7.30	11.53	11.63	‘erp-system’, 481, ‘herz-kreislauf-system’, 318, ‘crm-system’, 256
video	23244	7.31	9.49	9.83	10.26	‘youtube-video’, 1478, ‘youtube-videos’, 1286, ‘video-interview’, 478
chef	21689	5.41	6.89	9.84	9.94	‘spd-chef’, 875, ‘fdp-chef’, 627, ‘ex-chef’, 520
version	20518	4.50	4.51	9.74	9.76	‘beta-version’, 1125, ‘online-version’, 564, ‘pc-version’, 564
service	19351	6.19	8.54	6.01	6.98	‘after-sales-service’, 7669, ‘euro-finanz-service’, 339, ‘shuttle-service’, 300
film	19273	6.52	9.18	10.25	10.52	‘film-reviews’, 575, ‘science-fiction-film’, 433, ‘bond-film’, 328
dollar	17667	5.48	9.43	1.07	2.36	‘us-dollar’, 14271, ‘us-dollars’, 572, ‘petro-dollar’, 112

pus, the highest-entropy types are *team* and *blog*, then *system*, *film*, and *video*. Several of the most-attested types are words used in domains of interaction that have been particularly affected by the influx of English, and may represent examples of *Bedürfnislehnwörter* (‘necessary borrowings’, Carstensen, 1965), or lexical elements whose denotation is not well-represented by existing German lexemes. This is the case for the brand names among the most frequent and highest-entropy bases (*twitter*, *facebook*, and *youtube*), and may also be true for workplace-related elements (*team*, *chef*, *service*, and *marketing*). Types with high total compound morphological entropy values represent those English-language elements that have been borrowed into the German lexicon and are the most flexible in terms of their potential productivity. Types with low values, on the other hand, are typically used only in one or a few set formulations.

It should be noted that many of the types in the base wordlist may not be anglicisms, but rather Greek- or Romance-language-derived words common to most European languages (*system*, *service*, *version*, *video*, etc.), which may have undergone borrowing from the source language directly into German, or may also have been borrowed via English mediation. In addition, some types represent borrowings that have long been established in the German lexicon (e.g. *film*, *chef*), and thus may no longer be perceived as anglicisms or borrowings.

5. Summary and future outlook

Compounding via hyphenization is a productive word formation process in German, and we found many hyphenated compound types including English elements across three different CMC and web corpora: a Twitter corpus, a corpus of diverse web texts, and a blog corpus. We measured the tendency of 8,313 English nouns to appear as elements in hyphenated German compounds and documented a tremendous diversity of types. Many of the most frequent types overall (e.g. *online-shop*, *youtube-video*) are hyphenated compounds that have been borrowed into German in

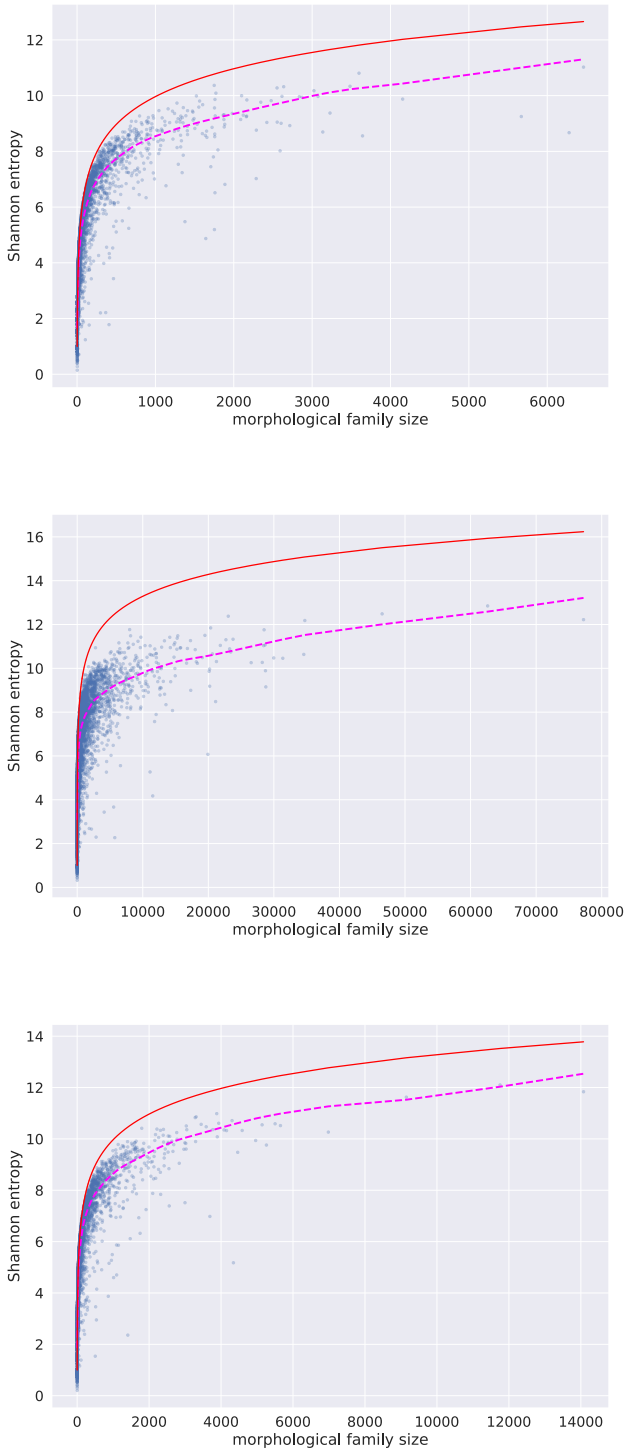

Figure 2: Shannon entropy vs. morphological family size, Twitter corpus, DECOW16bx corpus, and WordPress blogs corpus

order to denote new activities, technologies, or behaviors ("necessary borrowings"). The entropy analysis shows differences in entropy according to base constituent position within hyphenated compounds, and suggests that word formation via compound hyphenization of English bases is more productive on Twitter and on blogs, compared to other online genres. This can be interpreted as a consequence of the more spontaneous nature of text as used on blogs and on Twitter, which, as a paradigmatic form of CMC, has been suggested to have a status between speech and writing in terms of communicative typology and feature frequencies (Barbaresi & Würzner, 2014; Coats, 2016; Tagliamonte & Denis, 2008).

Future work with the data can be organized along three lines. First, a more thorough analysis of entropy according to constituent position within compounds and comparison with indigenous German lexical elements and English compound words may shed light on the dynamics of how borrowings are integrated into German morphological paradigms and whether conversion is taking place for some types. Second, many hyphenated compounds can also be written without hyphenation – a frequency analysis of the semantics of hyphenated and non-hyphenated compounds may be revealing. Finally, a frequency analysis of anglicism-containing hyphenated compounds of different structural types, according to the classification proposed by Fleischer and Barz (2012), may give further insight into the productivity of this large, chaotic, and fascinating word class.

6. References

Argamon, S. et al. (2007). "Mining the blogosphere: Age, gender, and the varieties of self-expression". In: *First Monday* 12.9. http://firstmonday.org/ojs/index.php/fm/article/view/2003.

Baayen, R. H. (1993). "On frequency, transparency, and productivity". In: *Yearbook of Morphology 1991*. Ed. by G. E. Booij and J. V. Marle. Dordrecht: Kluwer, pp. 109–149.

– (1994a). "Derivational productivity and text typology". In: *Journal of Quantitative Linguistics* 1, pp. 16–34.

– (1994b). "Productivity in language production". In: *Language and Cognitive Processes* 9, pp. 447–469.

– (2001). *Word frequency distributions*. Dordrecht: Kluwer.

– (2003). "Probabilistic approaches to morphology". In: *Probability Theory in Linguistics*. Ed. by R. Bod, J. B. Hay, and S. Jannedy. Cambridge, MA: MIT Press, pp. 229–287.

Baayen, R. H. and J. Hay (2002). *Affix productivity and base productivity*. Paper presented at ESSE 6, Strasbourg. https://pdfs.semanticscholar.org/db0d/e479b9686acd21\-e8ebc5147059c46ae0ed30.pdf.

Baayen, R. H., R. Lieber, and R. Schreuder (1997). "The morphological complexity of simplex nouns". In: *Linguistics* 35, pp. 861–877.

Baayen, R. H., L. H. Wurm, and J. Aycock (2007). "Lexical dynamics for low-frequency complex words: A regression task across tasks and modalities". In: *The Mental Lexicon* 2.3, pp. 419–463.

Barbaresi, A. (2016). "Efficient construction of metadata-enhanced web corpora". In: *Proceedings of the 10th Web as Corpus Workshop*, pp. 7–16.

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

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.

Coats, S. (2016). "Grammatical feature frequencies of English on Twitter in Finland". In: *English in computer-mediated communication: Variation, representation, and change*. Ed. by L. Squires. Berlin: de Gruyter Mouton, pp. 179–210.

– (2018). "Variation of new German verbal Anglicisms in a social media corpus". In: *Proceedings of the 6th Conference on CMC and Social Media Corpora for the Humanities*, pp. 27–32.

Duden (2006). *Die Deutsche Rechtschreibung (24th ed.)* Mannheim: 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.

Evert, S. and A. Lüdeling (2013). "Measuring morphological productivity: Is automatic preprocessing sufficient?" In: *Proceedings of the Corpus Linguistics 2001 Conference*.

Fleischer, W. and I. Barz (2012). *Wortbildung der deutschen Gegenwartssprache (4. ed.)* Berlin: de Gruyter.

Habert, B. and P. Zweigenbaum (2002). "Régler les règles". In: *TAL* 43.3, pp. 83–105.

Hay, J. B. (2001). "Lexical frequency in morphology: Is everything relative?" In: *Linguistics* 39, pp. 1041–1070.

Hay, J. B. and R. H. Baayen (2002). "Phonotactics, parsing and productivity". In: *Rivista di Linguistica* 15.1, pp. 99–130.

Hein, K. and S. Engelberg (2017). "Morphological variation: the case of productivity in German compound formation". In: *Mediterranean Morphology Meetings* 11, pp. 36–50.

Jong, N. H. de et al. (2002). "The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects". In: *Brain and Language* 81, pp. 555–567.

Krome, S. and B. Roll (2017). "Anglizismen und andere fremdsprachige Neologismen als Indizien für Sprach- und Schreibwandel: Empirische Analysen zum Schreibusus auf der Basis von Textkorpora professioneller und informeller Schreiber". In: *Studia Germanistica* 19, pp. 53–91.

Lüdeling, A. and S. Evert (2005). "The emergence of productive non-medical –itis: Corpus evidence and qualitative analysis". In: *Linguistic evidence: Empirical, theoretical, and computational perspectives*. Ed. by S. Kepser and M. Reis. Berlin: Mouton de Gruyter, pp. 91–95.

Lüdeling, A., S. Evert, and U. Heid (2000). "On measuring morphological productivity". In: *Proceedings of KONVENS 2000*, pp. 57–61.

Lüdeling, A. and N. H. de Jong (2002). "German particle verbs and word-formation". In: *Verb-particle Explorations*. Ed. by N. Dehé et al. Berlin: Mouton de Gruyter, pp. 315–334.

Moscoso del Prado Martín, F., A. Kostić, and R. H. Baayen. (2004). "Putting the bits together: an information theoretical perspective on morphological processing". In: *Cognition* 94, pp. 1–18.

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.

Schäfer, R. (2015). "Processing and querying large web corpora with the COW14 architecture". In: *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)*, pp. 28–34.

Schäfer, R. and F. Bildhauer (2012). "Building large corpora from the web using a new efficient tool chain". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 486–493.

Shannon, C. E. (1948). " A mathematical theory of communication". In: *Bell System Technical Journal* 27, pp. 379–423; 623–656.

Tagliamonte, S. and D. Denis (2008). "Linguistic ruin? Lol! Instant messaging and teen language". In: *American Speech* 83.1, pp. 3–34.

Winter-Froemel, E., A. Onysko, and A. Calude (2014). "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.

Errors Outside the Lab: The Interaction of a Psycholinguistic and a Sociolinguistic Variable in the Production of Verb Spelling Errors in Informal Computer-Mediated Communication

Hanne Surkyn, Reinhild Vandekerckhove, Dominiek Sandra
CLiPS, University of Antwerp
E-mail: hanne.surkyn@uantwerp.be, reinhild.vandekerckhove@uantwerp.be, dominiek.sandra@uantwerp.be

Abstract

The present study examines unintentional errors on regular verb homophones in informal online writing produced by Flemish high school students. While it reveals a high overall error rate, some consistent patterns can be discerned with respect to the social and psycholinguistic variables operationalized in the present research design: boys produce significantly more errors than girls, but in both gender groups the same effect of homophone dominance can be found. When misspelling a homophonous form, both boys and girls make more errors on the lower-frequency form (intrusion of the higher-frequency form) than on the higher-frequency one. Yet girls seem to exhibit greater error awareness or norm sensitivity. This case study intends to demonstrate that the interaction of social and mental processes in informal CMC offers a promising new research line and that CMC offers an ideal testing-ground for examining the ecological validity of conclusions from psycholinguistic experiments on these spelling errors.

Keywords: social media writing, spelling errors, gender, homophone dominance

1. Introduction

CMC-research often has an interdisciplinary orientation. Yet, the study of CMC at the intersection of the disciplines under discussion in the present paper, i.e. sociolinguistics and psycholinguistics, is largely uncharted territory. We intend to demonstrate that research into the interaction of social and mental processes in informal CMC offers a promising new research line.

The object of investigation concerns unintentional spelling errors on regular verb homophones in informal Dutch CMC (e.g., *wordt* for the target form *word*, both pronounced as [wort] < infinitive: *worden* 'to become'). These so-called homophone intrusions are highly persistent errors in all varieties of written Dutch that typically occur when working-memory runs out of resources, leaving insufficient time for applying the grammatically based spelling rule. Most errors occur on the lower-frequency homophone, which is known as the effect of homophone dominance. This preference for the more accessible higher-frequency form is the signature of long-term memory. Hence, persistent errors on fully regular forms reflect the interplay between working-memory limitations and long-term memory (Sandra, Frisson, & Daems, 1999).

The study of spelling deviations in informal CMC is not new. Quite a lot of the linguistically-oriented research on informal CMC focuses on the deviation from spelling conventions, more particularly, on typical markers of informal online writing (e.g., De Decker & Vandekerckhove, 2017). However, there is a crucial difference between the typical approaches to spelling deviations in studies of CMC and our approach: whereas the prototypical spelling deviations that characterize CMC tend to be deliberate choices on the part of the writer, spelling errors on verb homophones can be assumed to be largely unintentional. CMC contexts offer an ideal testing-ground for testing the 'ecological validity' of the assumption that they are triggered by the interplay between working-memory and long-term memory retrieval. Thus far this hypothesis has mainly been corroborated in the artificial context of spelling experiments.

2. Research Hypothesis

We examine errors on regular verb homophones in informal online writing of Flemish high school students. More precisely, we investigate whether gender (our sociolinguistic variable) affects the number and the pattern of these spelling errors. The term 'pattern' refers to the effect of homophone dominance (our psycholinguistic variable). Our hypothesis is that gender will affect the number but not the pattern of errors.

As homophone intrusions result from a failure to apply the rule in time, gender differences in rule application should only determine *how often* such a failure occurs but should not change the basic pattern *when* it occurs, i.e., the most common intruder will remain the higher-frequency form.

As for the expected gender differences, one of the most consistent findings in western sociolinguistics is that "women conform more closely than men to sociolinguistic norms that are overtly prescribed" (Labov, 2001: 293). Spelling errors on regular Dutch verb forms are highly stigmatized. Therefore, even though the CMC-context is a context of "pluralization of written language norms" (Androutsopoulos, 2011), we predict that high school girls will consciously attempt to avoid these particular errors more than boys. Hence, we expect them to allocate more attention to verb spelling rules and to consequently produce fewer errors.

3. Corpus and Target Items

Our corpus consists of informal and private online chat conversations produced in 2015 and 2016 via Facebook Messenger and WhatsApp. It consists of 434,537 posts and 2,531,354 tokens. All messages were produced by Flemish adolescents aged between 13 and 20 years old. We focused on homophonous verbs with stem-final *d*. These verb homophones involved the finite forms of the 1st, 2nd and 3rd person singular present tense. These forms end in the stem-final *–d* or *–dt*, the *t* being the marker of the 2nd and 3rd person. Table 1 presents the gender distribution of the data.

We signal that girls are overrepresented in the corpus (tokens, posts, target forms). However, even for boys there are more than 1600 target forms. Note that the statistical analysis takes this discrepancy in gender distribution into account.

	Boys	Girls
Participants	667 (48.19%)	717 (51.81%)
Posts	151,597 (34.89%)	282,940 (65.11%)
Tokens	834,837 (32.98%)	1,696,517 (67.02%)
***d-dt* target forms**	1,698 (28.25%)	4,313 (71.75%)

Table 1: Gender distribution in the corpus.

4. Research Variables and Data Processing

Spelling performance is the dependent variable in this study. The spelling of (the ending of) the homophonous verbs was manually encoded as correct or incorrect. The social factor *Gender* is the first independent variable. It is operationalized as a binary variable in terms of the biological sex of the adolescents (male vs. female). The second independent variable, the *Frequency Dominance* of the homophones, is the major psycholinguistic variable in the present research design. This frequency-related variable is based on the ratio between the frequency of the d-form over the frequency of the dt-form. These frequencies were extracted from SUBTLEX-NL, a database of Dutch word frequencies based on more than 40 million words from television and film subtitles (Keuleers, Brysbaert, & New, 2010). On the basis of these ratios, every target form was encoded as a d-dominant or dt-dominant verb. Finally, each target form was encoded with respect to the independent variable *Expected Spelling* (*d* or *dt*). An effect of homophone dominance will occur if higher error proportions are found on the dt-form of d-dominant verbs and on the d-form of dt-dominant verbs.

5. Results

The homophones accounted for 6,011 target forms. These were produced by 1,029 chatters and were distributed across 86 verbs. The number of observations varied considerably across verbs and chatters, which is to be expected in a CMC-context.

The number of correct spellings was significantly larger than the number of incorrect ones: *p*s<.0001 in Wilcoxon signed rank tests on the correct and incorrect responses, both when using error rates for lemmas or chatters as the unit of analysis. Still, the error rates were quite high. As Table 2 shows, these high-school students misspelled more than 28% of all analysed target forms. Such an error rate is considerable: more than 1 out of 4 forms were misspelled, even though their spelling is fully predictable by a few descriptively simple rules. This confirms a well-known fact: homophonous verb form in Dutch cause many spelling errors.

	Correct	Incorrect	Total
Boys	1,072 (63.13%)	626 (36.87%)	1,698
Girls	3,234 (74.98%)	1,079 (25.02%)	4,313
Total	4,306 (71.64%)	1,705 (28.36%)	6,011

Table 2: Numbers of correct and incorrect responses in the set of homophones with a stem-final *d*.

It turned out that only five verbs had high occurrence frequencies in the corpus and together accounted for 92% of the data. For two of these verbs, the d-form was the dominant homophone. For the other three, the dt-form was dominant. We analysed this set separately (restricted set). To check whether the effects were sufficiently generalizable across verbs, we extended this set with 11 more verbs. In this extended set, five had a dominant dt-form, whereas 11 had a dominant d-form. These 16 verbs were the only ones that yielded observations in the four critical cells that are used to test the effect of homophone dominance, i.e., the cells yielded by the orthogonal combination of Frequency Dominance and Expected Spelling.

We performed generalized linear mixed effects models to analyse the binary responses (correct/incorrect) in the restricted and extended sets. Mixed models make it possible to predict the outcome variable on the basis of fixed (independent) factors and one or more random factors. We used the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in the R statistical software package (R Core Team, 2014; see Baayen, Davidson, & Bates, 2008 for a review on linear mixed effects analyses). The particular distribution of verb forms across verbs and chatters made it impossible to include both random factors simultaneously. Hence, we used each random factor in separate analyses of the restricted and extended data sets.

An effect of homophone dominance is reflected in an interaction effect between Frequency Dominance and Expected Spelling. Because we were specifically interested in the question whether such an interaction would be comparable in boys and girls, the third-order interaction between Gender, Frequency Dominance, and Expected Spelling was investigated. This effect was non-significant, both with verbs and chatters as the random variable (*p*s>.20). This was the case in the restricted and in the extended set. When including Gender as a fixed factor, besides the second-order interaction between Frequency Dominance and Expected Spelling, we found (a) a significant interaction effect, i.e., an effect of homophone dominance, and (b) a significant effect of Gender. These effects emerged in the analyses of the restricted and extended sets, whether verbs or chatters were used as the random variable (all p*s*<.0001). The effect of homophone dominance was more outspoken when the dt-form was the expected spelling. This is not surprising, as considerably fewer errors were made on the d-form than on the dt-form (for the extended set, boys: 8.10% vs. 60.70%, respectively;

girls: 5.52%, vs. 51.78%, respectively, ignoring the factor Dominance). One possible explanation is that the CMC context makes it likely that the d-form (1st person) immediately follows the subject, which minimizes the error risk (Sandra et al., 1999). Another is that CMC may induce a stronger tendency to spell the (shorter) d-form than other writing contexts.

Importantly, all significant and non-significant effects remained when we removed the 51 chatters that provided many times more data to this data set than the others (i.e., 5% of the whole group; they provided about 40% of all verb homophones ending in -d or -dt).

Figures 1 to 4 visualize the homophone effects. When the correct spelling (of the ending) of the verb was *dt*, significantly more errors were made on d-dominant verbs ($p < .001$). On the other hand, when the correct ending was *d*, significantly more errors were made on dt-dominant verbs ($p < .001$). The effect of Gender remained a robust effect when the entire data set was analyzed, i.e., across all homophone tokens in the entire set of 86 verbs ($p < .0001$).

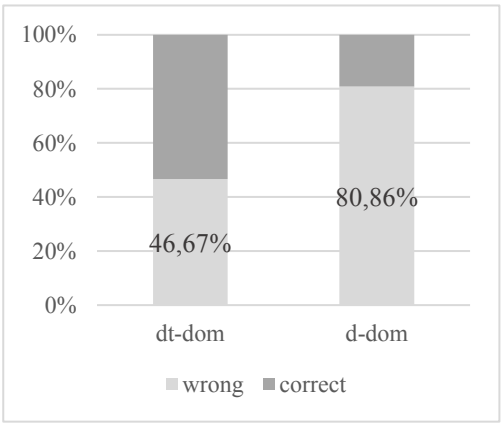

Figure 1: Distribution of correct and incorrect responses as a function of homophone dominance (boys, dt-ending).

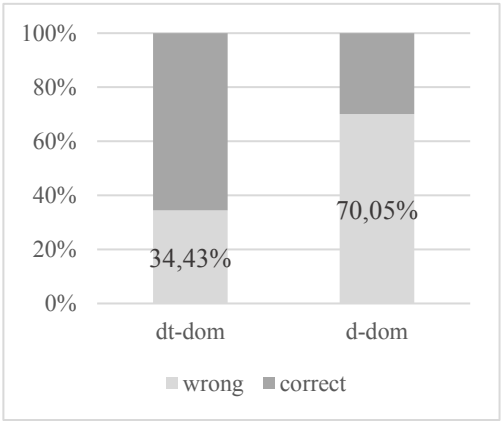

Figure 2: Distribution of correct and incorrect responses as a function of homophone dominance (girls, dt-ending).

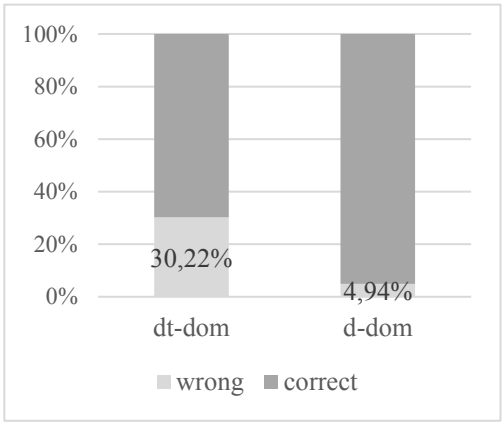

Figure 3: Distribution of correct and incorrect responses as a function of homophone dominance (boys, d-ending).

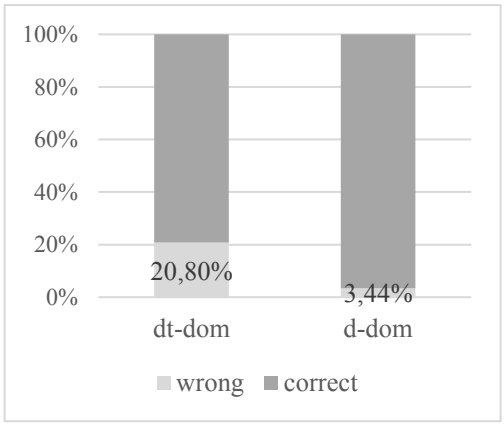

Figure 4: Distribution of correct and incorrect responses as a function of homophone dominance (girls, d-ending).

6. Discussion

Gender affects the number but not the pattern of the errors. Boys produced more errors than girls, but an effect of homophone dominance was found in both gender groups. These findings are consistent with our hypotheses about the effect of our sociolinguistic variable and its interaction with the psycholinguistic one.

We do not expect rule knowledge to play a role in the effect of gender. There is no reason to assume that boys know the verb spelling rules less well than girls. It seems more likely that they attach less importance to (a correct) rule application. In a follow-up study however, we will test for rule knowledge of high-school boys and girls in a basic verb spelling test. Thus, we will be able to exclude that the difference between gender groups is caused by a difference in knowledge of verb spelling rules rather than by a difference in attention and norm sensitivity.

Given earlier findings that women are more sensitive to social norms (Labov, 2001), it seems likely that our effect of gender is related to a higher error awareness and norm sensitivity in the female chatters in our corpus. As a result, girls consciously try to avoid these mistakes to a larger extent than boys. They seem to allocate more attention to verb spelling rules, which leads to fewer homophone errors. A quite different account of the Gender effect, which does

not make reference to the concept of norm awareness, would treat it as an effect of message length in disguise. Indeed, previous research has shown that boys write shorter messages than girls (Hilte, 2019: 159). This might indicate that boys focus more strongly on speed. Note that an account of the Gender effect in terms of a hidden variable, i.e., time pressure, is fully compatible with the view of persistent spelling errors on Dutch verb homophones that we adopted above (Sandra, Frisson, & Daems, 1999; Sandra, 2010). Higher time pressure causes a temporary overload in working memory, which creates the ideal trigger for spellers to rely on the higher-frequency form in the mental lexicon. This, in turn, causes many homophone intrusions when the lower-frequency form must be spelled. Hence, if boys' shorter messages indeed result from their stronger focus on speed, they will be more prone to homophone intrusion errors. This will manifest itself as the Gender effect that we observed.

The second important finding is that the error data for girls shows the same pattern as that for boys, despite the strong gender effect. This indicates that the gender effect is only a quantitative one, not a qualitative one. Whenever a homophonous verb form escapes the attention of writers, they may fail to apply the rule in time, and can fall prone to a homophone intrusion. This is in line with the account of Sandra et al. (1999, see also Sandra, 2010; Sandra & Van Abbenyen, 2009), in which an attentional mechanism and an automatic process of retrieval from long-term memory (i.e., the mental lexicon) jointly determine the error rate. Whereas the former determines the number of errors, the latter is responsible for the nature of the errors: a higher probability for the intrusion of the higher-frequency form on its lower-frequency homophone.

It is important to note that auto-correction cannot affect the production of spelling errors on verb homophones, and, hence, cannot have affected our error data. Auto-correction only corrects non-existing words, without taking their grammatical context into account. Since both homophones of a verb are existing forms, auto-correction cannot detect spelling errors on these forms.

7. Conclusion

In this paper we studied spelling errors on verb homophones in the informal Dutch CMC of Flemish adolescents, by combining a sociolinguistic and psycholinguistic perspective. The spelling of these forms is fully rule-governed. We investigated the effect of gender, a social variable, and homophone dominance, a psycholinguistic variable. Girls outperformed boys in the (correct) spelling of these verb homophones. However, when misspelling a homophonous verb form, both gender groups made more intrusions on the lower-frequency form (intrusion of the higher-frequency form) than on the higher-frequency one, irrespective of the form's ending (*d* or *dt*). Further research will have to show whether other social factors such as students' age and educational track co-determine the error pattern on these verbs in the CMC of youngsters.

By studying the production of verb spelling errors in a CMC context rather than in the artificial context of spelling experiments, we tested the 'ecological validity' of the conclusion from psycholinguistic experiment that these spelling errors are triggered by the interplay between working-memory and long-term memory retrieval (see the work by Sandra and colleagues). The present study revealed that the typical pattern of homophone intrusions that has been observed in spelling experiments, also occurs in spontaneous online chat conversations. Furthermore, this study showed that gender not only affects the production of typical chatspeak features (e.g., Hilte et al., 2018), but also the production of unintentional verb spelling errors in informal online writing. Finally, girls' increased attention for this error risk demonstrates that their increased norm sensitivity is also reflected in the context of CMC.

8. References

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. In T. Kristiansen & N. Coupland (Eds.), *Standard languages and language standards in a changing Europe*. Oslo: Novus, pp. 145--161.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), pp. 390--412.

De Decker, B. and Vandekerckhove, R. (2017). Global features of online communication in local Flemish: social and medium-related determinants. *Folia Linguistica*, 51, pp. 253--281.

Bates, D. Maechler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), pp. 1--48.

Hilte, L. (2019). *The social in social media writing: The impact of age, gender and social class indicators on adolescents' informal online writing practices*. Antwerp: University of Antwerp (doctoral dissertation).

Hilte, L., Vandekerckhove, R. and Daelemans, W. (2018). Expressive markers in online teenage talk. A correlational analysis. *Nederlandse Taalkunde*, 23(3), pp. 293--323.

Keuleers, E., Brysbaert, M. and New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods*, 42(3), pp. 643--650.

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Malden, MA: Blackwell Publishers Inc, pp. 261--293.

R Core Team. (2014). R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-Project.org/.*

Sandra, D. (2010). Homophone dominance at the whole-word and sub-word levels: Spelling errors suggest full-form storage of regularly inflected verb forms. *Language and Speech*, 53(3), pp. 405--444.

Sandra, D., Frisson, S., and Daems, F. (1999). Why simple verb forms can be so difficult to spell: the influence of homophone frequency and distance in Dutch. *Brain and Language*, 68, pp. 277--283.

Sandra, D. and Van Abbenyen, L. (2009). Frequency and analogical effects in the spelling of full-form and sublexical homophonous patterns by 12-year old children. *The Mental Lexicon*, 4, pp. 239

Preparing the ground for critical feedback in online discussions: A look at mitigation strategies[1]

Mario Cal-Varela, Francisco Javier Fernández-Polo
Universidade de Santiago de Compostela (Spain)
E-mail: mario.cal@usc.es, xabier.fernandez@usc.es

Abstract

As we become used to the increasingly important role of computer-mediated communication in tertiary education, our attention is also drawn to the particular relevance of positive interpersonal relationships for successful peer interaction and learning through this medium. This is particularly the case when critical feedback is exchanged between students. In this paper, we examine 240 forum discussion posts sampled from the Spanish component of the SUNCODAC corpus of online discussions and look at the students' choice of mitigation strategies in a specific rhetorical move. Four different types of strategies are identified and tracked over the timespan of a term. Quantitative analysis suggests that the presence/absence of the move and the number and combination of strategies used in it is a key aspect of the development of the genre over time. On the other hand, different subgroups of students may be playing distinct roles in the process, with non-native students seemingly acting as role models.

Keywords: discussion forum, critical feedback, mitigation strategies

1. Introduction

Managing interpersonal relationships has always been a major concern for participants in CMC and a major research topic for specialists (Rourke et al. 2001). This issue has been of particular interest for specialists in CMC in academic settings, where online interaction among teachers and students plays an increasingly important role (Swan & Shih, 2005). Building positive interpersonal relationships has been found to be crucial in securing effective interaction in these contexts and participants deploy a variety of strategies to, for instance, instil social presence to fulfil this goal.

2. Study aims

Politeness issues occupy a central position in the agenda of CMC research specialists (Locher 2010, Herring 2012). In a learning context, in particular, as Schallert & al. (citing Yang & al. 2006) state, "politeness strategies […] can foster a sense of community among participants by creating a comfort zone in which to exchange ideas as well as motivating students' participation in the learning process." (2009, p. 715) The concept of "face" (Goffman 1967), or public image, is central to Politeness Theory (Brown & Levinson 1987). Participants in an interaction will try to preserve both one's face and that of the interlocutor, particularly when face is under threat. As Schallert & al. recall (2009, p. 715), "as one interacts in CMD, face can be threatened by acts such as disagreements, criticisms, requests for information or help, and requests for clarification of a prior message." In this paper, we investigate how students participating in a discussion forum show sensitiveness towards their peers, in particular, by using language that anticipates and seeks to mitigate the presumably negative effect of criticism (cfr. Vandegriff 2013).

The aim of the forum is to work collectively towards the production of an optimal translation. One of the students (moderator) makes an initial translation proposal, which is followed by peer feedback in the form of criticisms and suggestions for improvement. Both criticism and suggestions are potential face-threatening acts for the moderator: criticism threatens the positive face of the addressee, their desire to be appreciated by others, while suggestions threaten the negative face of the student directly addressed and that of the other classmates, their desire to be unimpeded, by imposing the author's opinion on them and by constituting an implicit demand for credit and recognition.

Pre-emptive strategies used by students in a specific section of their messages were analysed in a carefully balanced sample of 240 posts from the Spanish component of SUNCODAC, a corpus of academic discussion forums compiled at the University of Santiago de Compostela (Spain). The context is a blended-learning undergraduate English into Spanish translation course, with the forum providing a natural complement to the face-to-face teaching. The group of students consists of both native and non-native Spanish speakers of different nationalities, with an intermediate to advanced level of Spanish as a second language. Regarding the non-native component in our sample, 32 posts are by English native speakers and the remaining 88 by speakers of other nationalities, among which there is a clear prevalence of Chinese speakers. Although posters' names are clearly visible to every participant, given the large number of students enrolled in the course (around 150), chances are that most participants might not be personally acquainted. Unfortunately, information on this and other aspects of the interactions, such as the participants' degree of familiarity with the genre and previous experience with forum discussions, was not systematically collected and was, therefore, not included in

[1] Financial support for this research has been provided by the Regional Government of Galicia (Directorate General for Universities); grant number: ED431B 2018/05; and by the Spanish Ministry of Science, Innovation and Universities; research project "English in the new genre of digital communication: Native and non native contexts (CMC)"; grant number: PGC2018-093622-B-I00.

the corpus metadata.

The analysis focuses on the section of the posts where the poster heralds the coming criticism (*preview of criticism*). About 60% of the 240-post sample contain this move (Fernández Polo & Cal Varela 2018), which is frequently formulated in highly mitigated terms, as in examples 1 and 2 (in bold):

(1) *Hola 15BPM! Pienso que tu traducción está muy bien y* ***sólo voy a comentar un par de detalles que yo he puesto de otra forma****.*
[Hello 15BPM! I think that your translation is very good and ***I'm only going to mention a couple of details that I have put differently!]***

(2) *Hola buenas:*
Me parece que la traducción está bien hecha en general, pero ***creo que algunos puntos pueden ser mejorados para que la versión final sea perfecta.***
[Hello there: in my opinion, your translation is very well done in general, but ***I think that some points may be improved so that the final version is perfect]***

In particular, our interest focused on how these strategies may change over time (beginning, middle and end of term), as the learning community works collectively to adapt the genre to suit their communicative needs. Our sample is so designed as to also allow us to explore the effect on the mitigation of criticism (the dependent variable) of other contextual variables like speaker's nativeness (native vs. non-native speaker use) and gender (male/female), as well as on the possible combined effect of all three variables.

3. Background

One defining feature of the current "wave" of research on CMC is a focus on variability (Locher 2010, p. 2), on exploring the effects of various contextual factors on CMC language, including purpose of the exchange, age, gender and status of participants, etc.

Regarding purpose, while there has been a lot of research on CMC in education, little is known about the online behaviour of students in multicultural settings like those represented in SUNCODAC, where a majority of local students interact online with a minority of exchange students from different linguacultural backgrounds using a lingua franca.

As for the effect of gender, Herring (1996) found that the characteristic male/female differences in communication styles observed in face-to-face interaction tend to be reproduced in academic forums, with males showing a more assertive style and one that focuses on propositional content, while females put a greater emphasis on expressing alignment and on the creation of rapport (for similar findings, see Guiller & Durndell 2006). Interestingly, however, Herring also finds that the relative weight of the genders in a forum may result in the minority gender adapting to the communicative style of the majority group, with women behaving more assertively in male-dominated forums and vice versa.

One question that remains unanswered in Herring's study is whether these "adaptations" occur from the very beginning of the participation in the forums or evolve through time, as the minority group develops an awareness of the majority practice and converges. Language is instrumental in the construction of a discourse community and at the same time reflects the evolution in the goals, priorities and relationships between its members. The effect of time in the communicative practices of online learning communities has been studied, for instance, by Swan (2002). She finds that the use of immediacy indicators, which is crucial in the constitution of online discourse communities, evolves over time, with some of these indicators becoming redundant as the community becomes consolidated, while others, like the expression of acknowledgement and approval, remain important throughout the term.

4. Method

A pilot study was conducted (90 posts) to explore the data, resulting in a 4-category typology of general attenuating tactics which draws heavily from existing research on mitigation (Caffi 1999), attenuation (Albelda & Cestero 2011, Briz & Albelda 2013) and politeness in general (Brown & Levinson 1987). All four strategies are illustrated in examples (1) and (2) above:

A. mitigating the illocutionary force of the critical comments: *creo que* (*I think that*), *yo traduciría* (*I would translate*);
B. minimizing the number or the gravity of the changes needed: *algunos detalles* (*some details*), *un par de cosas* (*a couple of things*);
C. defocusing the responsibility of the writer and the addressee e.g. by resorting to impersonal reference: *algunas partes necesitan cambiar* (*some parts need to change*), *se podrían mejorar* (*they might be improved*);
D. expressing interest in the addressee and/or in the writer's and the addressee's shared goals regarding the task at hand: *espero que te sirvan de ayuda* (*I hope that they will help you*), *para que la version final sea perfecta* (*so that the final version may be perfect*).

Random subsamples of messages were extracted from the SUNCODAC corpus to produce a larger 240-post sample containing equal numbers of texts for each level of the three variables of interest: period, sex and language background. 80 posts were thus selected from each of the three periods – beginning, middle and end – into which the term was divided (P1, P2 and P3). As shown in Table 1, the subgroups of participants as defined by the combination of

the levels of the other two independent variables (native males, native females, non-native males and non-native females) each contributed 20 messages per period to the final sample.

	NNS			NS		
	P1	P2	P3	P1	P2	P3
Female	20	20	20	20	20	20
Male	20	20	20	20	20	20

Table 4 Distribution of posts in the sample, by sex, nativeness and term period.

As some subgroups are comparatively underrepresented in the corpus (particularly males and non-native speakers), some of individual participants contributed more than one message to the sample. On the whole, however, a clear majority of 65 out of the 98 participants contributed only one (43 students) or two messages (22 students). The number of different individuals represented in each of the cells in Table 1 ranged between 10 and 19.

The above typology attenuation tactics (A, B, C and D) was used to identify and classify the mitigation strategies found in all the instances of the preview of criticism move in the final 240-post sample. In order to ensure maximum reliability, all the analyses were carried out independently by the two authors, the results compared and all contentious instances discussed until a final consensus was reached.

The participants' relative effort (or intensity) of attenuation in this part of the posts was then measured by paying attention to the following three dimensions:

a) the presence or absence of the *preview of criticism* move in the posts;
b) the degree of reiteration/repetition of the same attenuation strategy in a single move (only for strategies A and B, the only ones where reiteration or repetition of the same strategy was systematically observed);
c) the degree of complexity or combination of several strategies (A, B, C, D) in a single move.

The procedure can be illustrated with example (3), where three different strategies were identified:

(3) Hay varias cosas que creo que podrían mejorarse. [There are several things which, I think, might be improved]

Type A strategies are implemented through the combined use of *creo que* and the conditional form (*podrían*), while type B is present in the choice of generic and neutral *cosas* (*things*) rather than, for example, an explicit mention of *problems*, *errors* or similar, and the impersonal *mejorarse* illustrates type C. The example was consequently coded in the database as 2A, 1B, 1C.

5. Results

In terms of the relative presence/absence of the *preview of criticism* move (dimension a), a clear tendency is observed for participants to incorporate this component into their posts as the course evolves. The increase in the overall percentage of messages containing the move is particularly sharp from P1 to P2 and seems to level out towards the end of the term: 40%→63.75%→65%. A closer look at the data, however, reveals that the pattern is not equally strong across all participant groups. This trend was only statistically significant for the NS group (Chi-square = 15.5499, p-value = .000). As shown in Table 2, while NNSs start off at higher levels of use of the move, which remain more or less constant throughout the course, NSs start at rather low levels in P1, but show a rapid increase, surpassing their NNS classmates in periods 2 and 3.

	NNS			NS		
	P1	P2	P3	P1	P2	P3
Move absent	21	16	17	27	13	11
Move present	19	24	23	13	27	29

Table 5 Number of post with *preview of criticism* move absent/present, by language background (Non-native vs. native speakers of Spanish).

While both male and female participants contribute to the increase in the frequency of the move, as shown in Table 3, differences across periods only reach statistical significance in the case of male students (Chi-square = 8.3807, p-value = .015).

	F			M		
	P1	P2	P3	P1	P2	P3
Move absent	23	14	15	25	15	13
Move present	17	26	25	15	25	27

Table 6 Number of posts with *preview of criticism* move, absent/present, by sex.

Intensity of mitigation as measured by the accumulation of instances of the same strategy in a single message (dimension b) also shows a significant increase across periods (Table 4; Chi-square = 7.3030, p-value = 0.026). Seemingly, after a few weeks, students become aware of the relevance of doing interpersonal facework in this section of their posts and consequently increase their "mitigating effort" over the timespan of the course.

Strategies A or B	P1	P2	P3
Single	34	43	35
Multiple	26	47	64

Table 7 Number of single / multiple instances of either strategy (A or B), by period.

The pattern of distribution does not differ significantly

across genders. However, a comparison by language background of participants reveals clear and significant differences between NS and NNS students, with the former showing a much higher incidence of multiple (or reiterated) attenuators in this section of their posts and a much lower proportion of single (non-reiterated) attenuators, as shown in Table 5 (Chi-square = 11.8074, p-value = 0.001).

Strategies A or B	NNS	NS
Single	67	45
Multiple	52	85

Table 8 Number of single / multiple instances of either strategy (A or B), by language background.

When we look at the relative intensity of the move in terms of the combination of different attenuation strategies (dimension c), we observe a strong tendency for the move to become increasingly more complex from P1 to P3. Instances like (4) or (5), where two or more mitigation strategies are combined, are more frequent in P2 and P3 as compared to P1.

(4) Hola, tu traducción me parece muy acertada, pero hay un par de cosas que modificaría.
[Hi, your translation seems very right to me, but there are a couple of things that I would change.]

(5) [P]rimero que todo felicitarte por esta gran traducción porque el fragmento es uno de los más complicados de lo que llevamos hecho hasta ahora y los solventastes bastante bien. Aún así me gustaría sugerirte algunos cambios para mejorar tu trabajo.
[First of all, congratulations on this great translation because the excerpt is one of the most difficult ones we have done so far and you have dealt with it quite well. Still,I would like to suggest some changes to improve your work.]

As shown in Table 6, posts in P2 and P3 register a much lower incidence of zero-attenuation and contain a significantly higher number of examples with a combination of two or more different attenuation strategies in a single move (Chi-square = 18.4957, p-value = .005).

	P1	P2	P3
Zero (i.e. no move)	48	29	28
Simple (A, B)	2	7	4
Double (AB, AC, BC…)	24	34	30
Triple (ABC, ABD)	6	10	18

Table 9 Distribution of posts over time by complexity of mitigation strategy.

When we factor in the *nativeness* variable (native vs. non-native speakers), we can see that it is mostly native speakers who are responsible for the observed increase in the complexity of the attenuation in this move over time.

NNS			NS		
	P1	P2	P3	P1	P2	P3
zero	21	16	17	27	13	11
simple	1	3	2	1	4	2
double	13	16	16	11	18	14
triple	5	5	5	1	5	13

Table 10 Distribution of posts over time by complexity of mitigation strategy and language background of poster.

While attenuation does not become more complex over time in the NNS's posts (Chi-square = 2.1778, p-value = .903), it does show a significant level of increasing complexity, as measured by the combination of two or more strategies in a single move, in the posts written by NSs (Chi-square = 24.4516, p-value = .000). In their posts, the number of zero-attenuation moves decreases by over half from P1 to P3, whereas the frequency of moves with a combination of two or more attenuation strategies increases in a similar proportion over the same period. This tendency, while somewhat more marked in males, is statistically significant for both male (Chi-square = 16.0500, p-value = .013) and female (Chi-square = 14.6280, p-value = .023) native speakers.

6. Discussion and conclusions

The quantitative analysis of the combined effect of time, gender and nativeness on the attenuating strategies employed by the SUNCODAC forum participants in the *preview of criticism* move yields the following conclusions:

1. There seems to be a clear evolution in the writing of this *preview-of-criticism* move over the timespan of the course. In particular, the students tend to incorporate the move in their posts more and more frequently as the course develops. They also show a tendency to make this move more complex and more intensely attenuated, illustrating an increasing awareness of the importance of paying attention to politeness issues in their interventions and, more generally, revealing a process of co-construction of the genre by the group.

2. The analysis also reveals the existence of differences between native and non-native speakers in the way they choose to attenuate the negative impact of their criticism, and, more importantly, the possible existence of power relationships within the group, with some of the participants behaving as leading agents in the observed changes in the group's writing practices. In particular, although a minority in the group, NNSs seem to enjoy a relative prestige. Not only do they show a comparable stability in their language patterns over time, with little evolution over the term, but they also, and more importantly, seem to be showing the way in some of the changes observed in the NSs' posts. In comparison, NSs seem to start the term rather tentatively, without a clear idea of how to elaborate their posts and incorporating new rhetorical moves, apparently imitating NNSs and eventually surpassing

their models by the end of the period. The opposite tendency, with NNSs converging to NS patterns over time, is not observed in the data. This is a very interesting finding in terms of the role of these two groups in multicultural online student communities.

3. As regards gender differences, some of the findings contradict traditional assumptions that females tend to use more attenuated speech forms: males end up by incorporating this move more than their female classmates towards the end of the term. Their moves also become more intensely attenuated than their female classmates', as measured by the presence of multiple attenuators. As a matter of fact, males' attenuation patterns evolve more markedly over the period to conform to NNS patterns. In general, males seem to be more willing to observe and imitate other ways of doing, in other words, to make changes in their way of writing this part of their posts, assuming the majority pattern, in line with Herring's (1996) findings.

4. The effects observed for the three variables analysed in this exploratory study point to the desirability of enlarging the text sample in order to explore in much deeper detail the interaction of these factors. On the other hand, it would certainly be worth singling out some of the most active participants in order to track their individual progress over the course of the term to check the extent to which overall trends impact individual behaviour.

7. References

Albelda, M., & Cestero, A. M. (2011). De nuevo, sobre los procedimientos de atenuación lingüística. *Español Actual: Revista de Español Vivo*, 96, 9–40.

Briz, A., & Albelda, M. (2013). Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español y portugués. La base de un proyecto en común. *Onomázein. Revista Semestral de Lingüística, Filología y Traducción*, 28, 288–319.

Brown, P., & Levinson, S. C. (1987).*Politeness: Some universals in language usage*. Cambridge University Press.

Caffi, C. (1999). On mitigation. *Journal of Pragmatics*, 31, 881–909.

Fernández Polo, F. J. & Cal Varela, M. (2018). A structural analysis of student online forum discussions. In F. J. Díaz Pérez & M. Á. Moreno Moreno (Eds.), *Languages at the crossroads. Languages, accreditation and context of use* (pp. 189–200). Jaén: Universidad de Jaén.

Goffman, E. (Ed.). (1967). *Interaction ritual: Essays on face-to-face behavior*. Garden City, New York: Anchor Books.

Guiller, J., & Durndell, A. (2006). "I totally agree with you": Gender interactions in educational online discussion groups. *Journal of Computer Assisted Learning*, 22, 368–381.

Herring, S. C. (1996). Two variants of an electronic message schema. In S. C. Herring (Ed.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives* (pp. 81–106). Amsterdam: John Benjamins.

Herring, S. C., Stein, D., & Virtanen, T. (Eds.). (n.d.). *Pragmatics of Computer-Mediated Communication.* Berlin-Boston: Walter de Gruyter GmbH.

Locher, M. A. (2010). Introduction: Politeness and impoliteness in computer-mediated communication. *Journal of Politeness Research,* 6(1), 1–5.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Assessing social presence in asynchronous text-based computer conferencing. *Journal of Distance Education*, 14(2), 51–70.

Schallert, D. L., Chiang, Y. hui V., Park, Y., Jordan, M. E., Lee, H., Janne Cheng, A. C., Song, K. (2009). Being polite while fulfilling different discourse functions in online classroom discussions. *Computers and Education*, 53(3), 713–725.

Swan, K. (2002). Building learning communities in online courses: The importance of interaction. *Education, Communication & Information*, 2(1), 23–49.

Swan, K., & Shih, L. F. (2005). On the nature and development of social presence in online course discussions. *Journal of Asynchronous Learning Networks*, 9(3), 115–136.

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

Yang, M., Chen, Y., Kim, M., Chang, Y., Cheng, A., Park, Y., et al. (2006). Facilitating or limiting? The role of politeness in how students participate in an online classroom discussion. *Yearbook of the National Reading Conference*, 55, 341–356.

The Paralinguistic Function of Emojis in Twitter Communication

Yasmin Tantawi, Mary Beth Rosson
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, USA
ytantawi@hotmail.com, mrosson@psu.edu

Abstract

In response to the dearth of information about emoji use for different purposes in different settings, this paper investigates the paralinguistic function of emojis within Twitter communication in the United States. The Twitter feeds from 16 population centers spread throughout the United States were collected. In addition to a statistical analysis of emoji usage, a topic analysis was conducted using the IBM Watson API Natural Language Understanding. Further, a manual content analysis was conducted to ascertain the paralinguistic and emotional features of the emojis used. We present our characterization of emoji usage in Twitter and discuss implications for the design of Twitter and other text-based communication tools. In particular, we found a prevalence for expression of attitudes or the injection of a gesture to complement a tweet text. We conclude that there is definitely a paralinguistic substratum within the usage of emojis on text-based online communication platforms.

Keywords: Computer-mediated communication, content analysis, emoji, paralinguistics, Twitter

1. Introduction

The present day has seen an increase in the popularity of social networking sites such as Facebook, Twitter, and LinkedIn. Hundreds of such sites are currently in use daily by millions of people. Importantly, many of these social media channels rely heavily on text communication, which is very different from real world face-to-face (FTF) conversation. FTF exchanges normally include not only spoken words but also contain information about voice intonation, timing as well as a wide range of body language. This nonverbal communication plays a *paralinguistic* role in communication; it operates in tandem with the words that are uttered, to convey important emotional and social aspects of a conversation. Such information is analyzed in an almost instantaneous fashion by the individuals taking part in a conversation (Meeren et al., 2005) and significantly enriches the exchanged information.

Because simple text-based communications do not contain body language and other paralinguistic features, we propose that emojis are being recruited to play this role. *Emojis* are abstract representations of facial expressions and body language, among other things; as such they may inject paralinguistic content into text-based communications (Gkoni et al., 2017). Therefore, it can be said that the use of emojis in text-based online communications enhances the emotional content and context of a conversation. Even if emotion is expressed using text, the text itself is an abstraction of the emotional content that is typically conveyed via paralinguistic body language and facial expressions. In the work reported here, we examined the nature of emoji use in text-based communication, with a focus on their paralinguistic role.

1.1 Emojis as a Communication Medium

Emojis, which became popular during the mid-1990s, emerged from an earlier form of keyboard-based emotional expressions known as *emoticons*, which consist of a series of keystrokes that represent a facial expression. Some examples of emoticons are: ‘:)’, which indicates a smiling face; and ‘;-)’, which indicates a wink. However, unlike emoticons, which must be assembled keystroke by keystroke by the user and may vary significantly in form, emojis are predefined and not modifiable; they are selected from an emoji keyboard. Note that current emoji keyboards contain more than facial expressions. Many objects, including everyday objects like a sun or a tree, have been rendered as emojis. In addition, many activities, animals, and symbols are also represented by various emojis (Alshenqeeti, 2016), thus expanding the potential roles of emojis as part of text-based communication. In general, use of emojis appears to be increasing, with keyboards available on many electronic communication devices (Pohl et al., 2017).

1.2 Twitter as a Text-Based Communication Channel

Twitter is a microblogging platform that was launched in October, 2006. *Microblogging* can be defined as a manner by which one quickly updates one’s friends and online followers in real time with regards to the mundane activities which one is engaging in, or to tell one’s social network about current affairs in the world. Because this method of communication must be short, the Twitter platform places a length limit for each microblog, or tweet, to 280 characters, as of 2017 (Rosen, 2017). Twitter use has grown exponentially from the time of its founding in October, 2006. By April 2007, Twitter users numbered 94,000 (Java et al., 2007). At the time of writing, the number of users of this social networking site are 326 million worldwide (Cooper, 2019).

The topics discussed on the Twitter social networking site are very diverse. However, in general, Twitter users discuss the same topics that are circulating within the news media at the time that these tweets are created by individual Twitter users (Kwak et al., 2010). As a result, Twitter is seen as an effective way to quickly sample “what is happening” in the world, or at least in the world defined by

the users a Twitter reader chooses to follow.

1.3 Paralinguistic Content in Text-Based Communication

As a result of the abstractness and lack of imagery in pure text-based conversation, it is impossible to convey the wide range of paralinguistic functions that body language and facial expressions bring to FTF exchanges. The paralinguistic features of spoken language are primarily auditory and visual, unlike typed text. The auditory paralinguistic features of language include such aspects as fluctuations in vocal pitch and energy fluctuations in the speaker's voice (Scherer et al., 1973). The visual paralinguistic features of language include such aspects as moving one's eyebrows or nodding one's head (Duncan & Fiske, 1979). Because the conversations that occur on social networking sites such as Twitter are textual in nature and contain no real-time viewing of the individuals who are conversing, these traditional paralinguistic functions will be absent during the course of such a conversation, leading to potential ambiguity or misunderstanding of emotional content.

1.4 The Paralinguistic Potential of Emojis

There is a sound case for the claim that emojis are used as a substitute for the body language inherently present in face-to-face conversation, at least partly because usage of these emotional abstractions is currently increasing. (Durscheid & Siever, 2017). As the use of emojis has increased, the presence of emoticons (a series of keystrokes arranged to visually convey a facial expression) has decreased (Pavalanathan & Eisenstein, 2016). This finding suggests that the paralinguistic role offered by emojis is superior to that of other forms of digitally-represented paralinguistic cues, such as emoticons, video extracts, memes, and other popular expressive media available for use on Twitter. This finding is corroborated by the fact that the standardization of emojis has given rise to a more exact conveyance of emotional content than the unstandardized nature of emoticons, which may be interpreted in a myriad of ways. However, despite this standardization, even emojis can be interpreted quite differently, depending on, for instance, the interlocuter's cultural background (Freedman, 2018).

Previous research on emojis has revealed that, in general, emojis are used to convey emotions and simple body language in text-based conversations; however, most of these studies were conducted to ascertain usage differences across different national and gender groups. Thus, the research described in this paper is an extension of previous research that has studied differences in body language and gesture usage in real-world, in-person conversations, and by individuals of different nationalities (Algharabali & Taqi, 2018; Herring & Dainas, 2018; Ljubesic & Fiser, 2016).

With regards to gender differences in the usage of emojis, one study that investigated these differences found that females are more likely to use emojis to make sure that the emotional intent of a message is not misinterpreted; in contrast, males are more likely to use emojis to insert humor into a text-based conversation (Algharabali & Taqi, 2018). Also, females tend to use more emojis in their online communications on social networking sites than males (Herring & Dainas, 2018).

Looking only at Twitter communications, researchers have found that the populations of different countries use emojis to differing degrees. The country with the highest percentage of tweets containing emojis is Indonesia, whose Twitter users insert emojis in 46.5% of their tweets. Paraguay holds the second place in terms of tweets containing emojis, with 37.6% of tweets containing emojis, followed by the Philippines with 34.6% of tweets containing emojis. Algeria and Qatar rank fourth and fifth, with 33.5% and 32.6% of tweets emerging from those two countries, respectively, containing emojis. 10% of tweets emanating from the United States contain emojis (Ljubesic & Fiser, 2016).

We turn now to our empirical investigation of the paralinguistic role played by emojis in the United States on Twitter. The United States was chosen as the location of this analysis in order to ascertain the usage of emojis on Twitter by Americans. By having such a directed scope of analysis, a more nuanced understanding of the usage of emojis on this platform by this population can be gathered. This is in contrast to many previous studies in this domain which focused on a number of nations for their analysis of emoji usage on the Twitter platform. In section 2, we describe the methodology for sampling and analyzing the Twitter messages. Section 3 describes the results of this study; and Section 4 is a discussion of our results, as well as limitations and future work.

2. Research Methods

2.1 Data Collection

We chose sixteen population centers of the United States to sample with respect to the Twitter messages emanating from people who live there. These population centers were carefully chosen to represent multiple geographic regions within the United States as well as to include both an urban area and a rural town located just outside the urban city's border. This sampling method was chosen so as to increase the generalizability of this study. Scripts for collecting and analyzing the data were written in Python, and use various packages as discussed below.

Data collection took place on January 14, 2019, between the hours of 11:45 a.m. and 3:55 p.m. During this time, the Tweet collection program, which uses the Twitter API package (developer.twitter.com), was run on the 16 population centers in the following order. The specific time that tweets began to be collected from the population centers are indicated in parentheses next to the population center in question: New York City (11:45 a.m.); Somers, NY (11:47 a.m.); Miami (12:03 p.m.); Southwest Ranches, FL (12:07 p.m.); Chicago (12:13 p.m.); Channahon, IL (12:24 p.m.); Houston (12:48 p.m.); Kenefick, TX (12:51 p.m.); Denver (12:58 p.m.); Watkins, CO (1:07 p.m.); Phoenix (1:51 p.m.); Superior, AZ (1:58 p.m.); San Francisco (2:34 p.m.); Diablo, CA (2:38 p.m.); Seattle (2:51 p.m.); and Index, WA (2:58 p.m.). The ordering from East to West was intended to adjust informally for time zone differences, such that most tweets were collected around a given location's lunch period.

2.2 Data Analysis

After data collection, the dataset was analyzed and processed with the help of custom Python scripts. We used the scripts to count the total number of tweets; the number of tweets containing emojis (these tweets were extracted for further processing); the number of emojis; the ratio of the tweets containing emojis to the total number of tweets; and the average number of emojis per tweet.

Once the raw tweets had been counted and characterized with respect to the presence of emojis, we began to conduct a content analysis on the data. Content analysis is a method wherein the overall subject matter or other characteristics of a text is ascertained via a coding scheme.

Our general content analysis approach was multi-phased. We first sought to analyze the context in which an emoji was used. Because this study seeks to determine the paralinguistic usage of emojis in tweets, we first needed to gain a sense of what the main goal of a message was so as to then assess the possible paralinguistic role of the emoji. We used the IBM Watson API Natural Language Understanding module to classify each emoji-bearing tweet by topic categories (www.ibm.com/watson/services/natural-language-understanding/). This module draws from a set of 23 categories. After the classification process was complete, we randomly selected 5% of the classified tweets and checked them manually to ensure that this first phase of content analysis was sensible.

After classifying the tweet topics to provide context, two more phases of content analysis were conducted. First, we assessed the paralinguistic function of the emojis used in a tweet. We used a coding scheme from an earlier research paper investigating the use of emojis in tweets (Na'aman et al., 2016). These categories were: topic, attitude, gesture, and unknown. We used these assignment rules: If an emoji was used to clarify the topic of the tweet in which it appears, it was coded into the topic category. If an emoji was used to display the attitude of the writer with respect to the content of the tweet, it was coded into the attitude category. If an emoji was used to convey a gesture that the writer might otherwise have expressed using nonverbal communication, it was coded into the gesture category. Finally, if the paralinguistic role of the emoji was unclear, the emoji was placed into unclear.

After this manual coding of the paralinguistic function of emojis, a secondary content analysis examined the specific emotion conveyed by the emojis that had been classified as either an attitude or gesture. The coding scheme we applied was again taken from a previous research paper focused on the use of emojis for emotional purposes. This coding scheme consists of 11 categories: joy, surprise, praise, pride, love, anger, confusion, anxiety, disapproval, boredom, and playfulness (Sun et al., 2019). Despite the fact that some previous research has focused on the potential for emojis to be used as indicators of irony in text-based communication (Weissman & Tanner, 2018), this study did not investigate irony in a paralinguistic sense as a result of the usage of the aforementioned coding scheme, which did not include irony.

3. Results

Overall, 1,600 tweets were collected. These included 269 tweets that had at least one emoji (16.8%). There was a total of 628 emojis in this sample, for an average of 2.33 emojis per tweet.

Table 1 lists the names of the emojis observed in this sample of tweets, along with their corresponding images and the percent of the total number of emojis represented by each one; tweets with fewer than five occurrences were not included. As can be seen in the ordered list, the top 5 emojis used by this sample of users were: the face-with-tears-of-joy emoji (70 instances of use); the loudly-crying-face emoji (39 instances of use); the rolling-on-the-floor-laughing emoji (22 instances of use); the red-heart emoji (21 instances of use); and the medium-dark-skin-tone emoji (20 instances of use).

3.1 Topic Analysis

The most common Tweet topic was Society, accounting for 99 tweets. Technology and Computing was the second most common topic (78 tweets). Art and Entertainment and Sports were the topic in 71 tweets each. It should be noted that each tweet could be assigned only one topic. Following are the remaining topics observed in this sample, with the number of tweets pertaining to each topic in parenthesis next to the topic in question: Law, Government, and Politics (40); Business and Industrial (32); Family and Parenting (27); Food and Drink (23); Health and Fitness (20); Finance (15); Education (14); Pets (11); Automotive and Vehicles (10); Travel (9); Careers (8); Religion and Spirituality (7); Real Estate (6); Science (3); News (3); Style and Fashion (2); Shopping (2). 43 tweets had a topic which was unknown to the Watson API Natural Language Understanding Module.

Text	Emoji	Percentage of Overall Tweets
face-with-tears-of-joy		11.1%
loudly-crying-face		6.2%
rolling-on-the-floor-laughing		3.5%
red-heart		3.3%
medium-dark-skin-tone		3.2%
smiling-face-with-heart-eyes		3.2%
medium-light-skin-tone		3.0%
female-sign		2.7%
fire		2.2%
face-blowing-a-kiss		1.8%
kiss-mark		1.6%
person-facepalming		1.4%
weary-face		1.4%
thinking-face		1.6%
clapping-hands		1.3%
dark-skin-tone		1.3%
male-sign		1.3%
woman-dancing		1.3%
raising-hands		1.3%
person-shrugging		1.1%
purple-heart		1.1%
face-with-rolling-eyes		1.0%
light-skin-tone		1.0%
drooling-face		0.8%
sparkles		0.8%
waving-hand		0.8%

Table 1: List of emojis with corresponding images and

percentages

3.2 Paralinguistic Analysis

The tweets we analyzed contained three paralinguistic features: attitude, gesture, and topic. In this phase of our content analysis, we coded each emoji for its function in cases where a tweet had multiple emojis. We found that 274 emoji-bearing tweets (43.6%) had the paralinguistic feature of attitude; 247 tweets (39.3%) had the paralinguistic feature of gesture; and 102 tweets (16.2%) had the paralinguistic feature of topic. These general findings suggest that the primary communication purpose of emojis is to convey nonverbal information such as emotion or attitude.

Within the emojis classified as attitude or gesture, we classified the emotion reflected by the emoji. Overall, we observed the occurrence of 11 emotional features: playfulness, praise, confusion, boredom, surprise, joy, pride, disapproval, anger, love, and unknown. The most common emotional feature was joy and the least common emotional feature was praise. We observed all 11 emotional features of this coding scheme except anxiety, which was not expressed by any of the tweets which were analyzed.

The top feature of a joy-expressing emoji appeared in 186 tweets; other common emotional attributes were love (71), playfulness (63), surprise (58), disapproval (43), anger (40), pride (28) and confusion (19).

4. Discussion

With regards to evolution of the Twitter social networking platform, we observed that society was the most common topic, suggesting that Twitter may be evolving from a platform to discuss current events to one where users "hold forth" on current societal issues. This is the case since, despite the fact that the topic of Society can be seen as encompassing the topic of current events; it is the case that the topic of current societal issues is more concerned with this topic than the topic of current events since current events include impersonal topics such as newsworthy events which have little to do with society as a whole. If so, the platform providers might consider ways to invite and support such discussions (as well as art, entertainment and sports).

With regards to paralinguistic uses of emojis, we found a prevalence for attitudes and gestures. This may suggest emojis used to convey facial expressions should be more realistic, so as to closely emulate the nonverbal behavior of FTF conversation. It was assumed that the paralinguistic intent of the emojis used was attributed only to the user who wrote the emojis in question. As an example, consider the example of the smiling-face-with-heart-eyes emoji. This emoji was used to convey a positive attitude on the part of the user with regards to the tweet topic. The specific emotional message of this emoji was typically that of love or joy, interpretable based on the text surrounding the emoji. However, in the real world, it is impossible for an individual to smile and have their eyes transform into hearts. Perhaps it would be better if this emoji was revised to be a smile with tender eyes; or a smiling face with a blush, which is a more realistic image of the emotions of love and joy (Shaver et al., 1996). Despite this, it was found that the smiling-face-with-heart-eyes emoji was the top fifth emoji used, which leads to the conclusion that, despite the presence of a blushing face emoji, users are more likely to use this rather unrealistic emoji to express love or joy. We can also make an argument for keeping the smiling-face-with-heart-eyes emoji since previous research has revealed that familiarity of icons is more conducive to optimal performance in a computer setting than the concreteness of the icon itself (Isherwood et al., 2007). Therefore, since Twitter users are already familiar with the smiling-face-with-heart-eyes emoji, it may not be wise to change it to be more realistic.

The American population has been shown to use emojis to convey positive affect. This is the case since this population mostly uses emojis for the emotional feature of joy, and secondly, love.

In conclusion, we observed an emoji-based paralinguistic substratum for Twitter communication. Most of the emojis were used to convey the user's attitude towards the topic of the tweet or to supplement the text of the tweet with an artificial gesture. Emotionally expressive emojis, such as the face-with-tears-of-joy and the loudly-crying-face emojis, are used widely by Twitter users located in the United States.

We recognized that our findings must be qualified by the sampling process we used. We collected just 1,600 tweets (because of the demands of manual coding). Even though we carefully sampled different residential contexts in the United States, we recommend that further research be conducted on a larger dataset which contains Twitter users from different parts of the world, or even a larger sample from the United States. In the future, it could be the case that sentiment analysis, still in its infancy in the present day, would become more advanced and reliable, which would support the coding necessary for large-scale data collection and analysis (Lin et al., 2018).

5. References

Algharabali, N. A. & Taqi, H. A. (2018). Taming the Sting: The Use of Evaluative Emojis by College Students in Kuwait. International Journal of Linguistics and Communication, 6(1), pp. 46–60.

Alshenqeeti, H. (2016). Are Emojis Creating a New or Old Visual Language for New Generations? A Socio-semiotic Study. Advances in Language and Literary Studies, 7(6), pp. 56-69.

Cooper, P. (2019). 28 Twitter Statistics All Marketers Need to Know in 2019. Retrieved from https://blog.hootsuite.com/twitter-statistics/

Duncan, S. & Fiske, D. (1979). Dynamic Patterning in Conversation: Language, paralinguistic sounds, intonation, facial expressions, and gestures combine to form the detailed structure and strategy of face-to-face interactions. American Scientist, 67(1), pp. 90-98.

Dürscheid, C. & Siever, C. (2017). Jenseits des Alphabets – Kommunikation mit Emojis. Zeitschrift für germanistische Linguistik, 45(2), pp. 256-285.

Freedman, A. (2018). Cultural literacy in the empire of emoji signs: Who is crying with joy? First Monday,

23(9).
Gkoni, N., Druiventak, E., Bollen, Y. & Ecott, S. (2017). Snapchat Fams as a Subculture: How Influencers Use Emojis for Commodifying Cross-Platform Engagement. Retrieved from https://mastersofmedia.hum.uva.nl/blog/2017/10/25/snapchat-fams-as-a-subculture-how-influencers-use-emojis-for-commodifying-cross-platform-engagement/.
Herring, S. C. & Dainas, A. R. (2018). Receiver interpretations of emoji functions: A gender perspective. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018). Stanford, CA*.
Isherwood, S., Mcdougall, S. & Curry, M. B. (2007). Icon Identification in Context: The Changing Role of Icon Characteristics with User Experience. Human Factors, 49(3), pp. 465–476.
Java, A., Song, X., Finin, T. & Tseng, B. (2007). Why we twitter. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis - WebKDD/SNA-KDD 07.
Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is Twitter, a Social Network or a News Media? International World Wide Web Conference Committee, pp. 591-600.
Lin, B., Zampetti, F., Bavota, G., Penta, M. D., Lanza, M. & Oliveto, R. (2018). Sentiment analysis for software engineering. Proceedings of the 40th International Conference on Software Engineering - ICSE 18.
Ljubešić, N. & Fišer, D. (2016). A Global Analysis of Emoji Usage. Proceedings of the 10th Web as Corpus Workshop, pp. 82–89.
Meeren, H. K. M., van Heijnsbergen, C. C. R. J. & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. Proceedings of the National Academy of Sciences, 102(45), pp. 16518–16523.
Na'aman, N., Provenza, H. & Montoya, O. (2017). Varying Linguistic Purposes of Emoji in (Twitter) Context. Proceedings of ACL 2017, Student Research Workshop, pp. 136–141.
Pavalanathan, U. & Eisenstein, J. (2016). More emojis, less :) The competition for paralinguistic function in microblog writing. First Monday, 21(11).
Pohl, H., Domin, C. & Rohs, M. (2017). Beyond Just Text. ACM Transactions on Computer-Human Interaction, 24(1), pp. 1-42.
Rosen, A. (2017). Tweeting Made Easier. Retrieved June 29, 2019, from https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html
Scherer, K. R., London, H. & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. Journal of Research in Personality, 7(1), pp. 31-44.
Shaver, P. R., Morgan, H. J. & Wu, S. (1996), Is love a "basic" emotion? Personal Relationships, 3, pp. 81-96.
Sun, N., Lavoue, E., Aritajati, C., Tabard, A. & Rosson, M. B. (2019). Using and Perceiving Emoji in Design Peer Feedback. 13th international conference on Computer Supported Collaborative Learning (CSCL 2019), pp.296-303.
Weissman, B. & Tanner, D. (2018). A strong wink between verbal and emoji-based irony: How the brain processes ironic emojis during language comprehension. Plos One, 13(8).

Collecting and Analyzing a Corpus of WhatsApp Interactions Using the MoCoDa² Web Interfaces

Michael Beißwenger (University of Duisburg-Essen), Marcel Fladrich (University of Hamburg), Wolfgang Imo (University of Hamburg), Evelyn Ziegler (University of Duisburg-Essen)

The poster presents recent developments in the corpus project *MoCoDa²* (*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*. Since early 2018 MoCoDa² is up and running. In June 2019 the database consisted of 323 chats with 1,164 participants which comprise 26,484 user posts and 221,006 tokens. The corpus can be accessed and queried via the web frontend after a registration.

MoCoDa² adopts a donation-based collection strategy. 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, meta-data on 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 (Beißwenger et al. 2019).

In this year's contribution to the cmccorpora conference, we will put a focus on new components of the collection, editing and query interface which have been integrated since autumn 2018. We will report first results of an evaluation to what extent donators actually add relevant metadata and textual descriptions to their donations using the web-based editing interface. In addition we will report examples of how the corpus has currently been used in linguistic research on CMC in order to illustrate what users can already do with MoCoDa². Use cases are (1) analyses on the relation of emoji use and gender, (2) analyses on pragmatic functions of emojis and (3) analyses on the use of regional dialect features in messaging interactions. All use cases will be published in late 2019/early 2020 and presented on the poster with research question, data snippets and findings.

Reference:

Beißwenger, Michael/Fladrich, Marcel/Imo, Wolfgang/Ziegler, Evelyn (2019, accepted): https://www.mocoda2.de: a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions. In: European Journal for Applied Linguistics (EuJAL).

cmc-core: A basic schema for encoding CMC corpora in TEI

Michael Beißwenger[1], Laura Herzberg[2], Harald Lüngen[3], Ciara R. Wigham[4]
[1] University of Duisburg-Essen, Germany [2] University of Mannheim, Germany [3] Leibniz-Institute for the German Language, Mannheim, Germany [4] University Clermont-Auvergne, France
E-mail: michael.beisswenger@uni-due.de, herzberg@uni-mannheim.de, luengen@ids-mannheim.de, ciara.wigham@uca.fr

Abstract

Since 2013 representatives of several French and German CMC corpus projects have developed three customizations of the TEI-P5 standard for text encoding in order to adapt the encoding schema and models provided by the TEI to the structural peculiarities of CMC discourse. Based on the three schema versions, a 4th version has been created which takes into account the experiences from encoding our corpora and which is specifically designed for the submission of a feature request to the TEI council. On our poster we would present the structure of this schema and its relations (commonalities and differences) to the previous schemas.

Keywords: CMC, cmc corpora, standard, TEI

Poster abstract

In close interconnection with the activities of the CMC corpora community, since 2013 representatives of several CMC corpus projects have been developing customizations of the TEI P5 standard for text encoding in order to adapt the encoding schema and models provided by the TEI to the structural peculiarities of CMC discourse. Since the TEI-P5 standard does not offer any specific models for the representation of CMC discourse the goal of the group - which could install a special interest group (SIG) on CMC as part of the TEI community - was twofold:

(1) *short-term goal:* provide encoding schemas which people could use for representing CMC corpora in a way which is compatible with the general structure of TEI documents ('TEI customizations') even though the TEI standards does not include models of CMC.
(2) *long-term goal:* gather and evaluate experience from different corpus projects using these schemas; develop the schema further and transform it into a 'feature request' to make an official proposal for an extension of the TEI standard with specific models for CMC.

The SIG started from a 1st schema draft (Beißwenger et al. 2012, 'DeRiK schema') which formed the basis for the creation of an extended schema by the French CoMeRe group (Chanier et al. 2014, 'CoMeRe schema') which was used for the encoding of 14 French CMC corpora. The latter was further developed in the German ChatCorpus2CLARIN project 2015/16 and adopted for encoding German chat, Wikipedia and Usenet corpora (Lüngen et al. 2016, Beißwenger 2018, 'CLARIN-D schema').
Based on the three schema versions, a 4th version has been created in 2018 which takes into account the experiences from encoding the abovementioned French and German CMC corpora and which is specifically designed for the submission of a feature request to the TEI council. On our poster we would present the structure of this schema and its relations (commonalities and differences) to the previous schemas.
The goal of the new schema, dubbed *cmc-core*, is to reduce the previous schema drafts "to the max" and provide an essential architecture of concepts which are needed for the representation of documents which typically form the basis of every CMC corpus. *Cmc-core* provides <post> as a basic model to describe the peculiarities of user contributions to CMC interactions which are - even in the case of spoken "audio posts" in WhatsApp sequences - characterized by a temporal rupture between production and transmission which makes them different from turns in spoken interactions. Instances of posts are constituents of 'CMC macrostructures' (logfiles or threads) which are represented using the element from the TEI standard. Posts can be subclassified by several attributes:

- For the distinction of spoken vs. written posts, we introduced the @mode attribute with its two possible values "written" and "spoken".
- For encoding a technical back reference from one post to one previous post, and the indentation level of wiki talk contributions, we use the attributes @replyTo and @indentLevel which were already included in the previous schema drafts.
- For classifying content according to different types of creators ("human", "template", "system", "bot", "unspecified") we use the attribute @creation (a further development of the attribute @auto from the previous schema).

The new cmc-core schema will be made available together with sample encodings of chat, twitter, wiki talk and transcribed 2nd life interactions on the CMC-SIG pages in the TEI wiki in August 2019 (https://wiki.tei-c.org/index.php?title=SIG:CMC). After publication in the TEI wiki members of the TEI-CMC-SIG and colleagues from the CMC corpora community will be invited (via their mailing lists) for critical review and comments. It is planned to submit the feature request to the TEI community by October 2019.

```
<post key="1043796093786566656"
      mode="written"
      creation="human"
      type="tweet"
      subtype="retweet"
      who="#u2"
      synch="#tweetsbcrn18.t003"
      xml:lang="de"
      xml:id="p5">
    <time creation="system"> 14:35 </time> Immer wieder gerne. Kann ich mich schon für nächstes
    Jahr als Empfangs- <ref type="hashtag" target="https://twitter.com/hashtag/Engel?src=hash"
        >#Engel</ref> für das nächste BarCamp bewerben <figure type="emoji" generation="templat
            ><desc type="meaning">zany face</desc>
        <desc type="unicode">U+1F92A</desc></figure>
    <ref type="hashtag" target="https://twitter.com/hashtag/bcrn18?src=hash">#bcrn18</ref>
    <trailer>
        <!-- The following is CoMeRe Style  -->
        <fs>
            <f name="favoritecount">
                <numeric value="4" />
            </f>
        </fs>
    </trailer>
</post>
```

Figure 1: Encoding of a tweet according to *cmc-core*

References

Beißwenger, M. (2018): Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI. In: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hg.): Digitale Infrastrukturen für die germanistische Forschung. (= Germanistische Sprachwissenschaft um 2020 6). Berlin u.a., pp. 307-349.

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

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J. & Seddah, D. (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics. *JLCL (Journal of Language Technology and Computational Linguistics),* Issue 2, pp. 1-31

Lüngen, H., Beißwenger, M., Ehrhardt, E., Herold, A. & Storrer, A. (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016).* Ruhr-Universität Bochum.

CMC Text-based messages environment types

Erika Lombart

UCL, ILC, Cental

Erika.lombart@uclouvain.be

1. Web 2.0 and TBCMC

Web 2.0, also known as *participative web* or *social web* (Mangenot and Soubrié, 2015 : 3), is characterized by digital contents easy to generate thanks to new technologies everybody can easily have access to, the key place of social activities (Levin et Bryan, 2008) and computer mediated communication (CMC, December, 1997). Within CMC stands text-based CMC (TBCMC) which includes written messages generally short-lived, quickly prepared, rarely or briefly reviewed and defined by their technological environment (Paveau, 2013). Although this list might quickly become obsolete considering technology moves fast, we currently have 9 TBCMC environments: blog, forum, social network, wiki, comments section on informative or commercial sites, SMS, instant messaging, chat and email.

The purpose of this contribution is to classify each environment according to its communication specificities upon 3 criteria: communication function, communication type and type of responses.

2. Criteria selection

2.1 Communication purpose

The approach here is based on the structuralism field with Jakobson's communication functions (1960) highlighting the 6 factors of an effective verbal communication: the sender (emotive), the context (referential), the message (poetic), the channel (phatic), the code (metalingual) and the receiver (conative). If TBCMC environments can host messages with different functions and are flexible, they also present particularities that allow the production of certain content types.

2.2 Number of participants

Some messages are part of mass communication and don't have any specific receiver. They belong to media communication (Kerbrat-Orecchioni, 2003). Other ones involve 2 p or more. They are part of mediated communication.

2.3 Synchronicity

TBCMC messages can stand as comments or be part of a dialogical exchange. As comments, messages are just the expression of a feeling or a point of view. No response is expected, although sometimes an unexpected one is received. As part of a dialog, messages can be asynchronous or quasi-synchronous. The first ones ask for a delayed answer and the second ones are part of an exchange in real time but don't give the receiver access to their production as a vocal message would do (Garcia and Jacobs, 1999).

3. Categorization

Message Type	Communication Function	Communication Type	Response Type: Comment/ Quasi-synchronous Dialogue/Asynchronous Dialogue
Public Blog	Referential /Emotive /Conative	Mass Communication	Comment
Discussion Forum (Free access)	Referential/ Emotive /Conative /Phatic	Mass Communication/More than 2 interlocutors	Asynchronous Dialogue
Social Network	Conative /Emotive /Referential	More than 2 interlocutors	Comment
Comment Section	Emotive /Conative /Referential	Mass Communication/More than 2 interlocutors	Asynchronous Dialogue
SMS	Emotive /Conative /Referential /Phatic	2 interlocutors and more	Asynchronous Dialogue
Instant Messaging	Emotive /Conative /Referential /Phatic	2 interlocutors and more	Quasi-Synchronous Dialogue
Chat	Emotive /Conative /Referential /Phatic	More than 2 interlocutors	Quasi-Synchronous Dialogue
Email	Referential /Emotive /Conative	2 interlocutors and more	Asynchronous Dialogue
Wiki	Referential	Mass Communication	Comment

Table 1. TBCMC environments overview table

4. References

December, J. (1997). Notes on Defining of Computer-Mediated Communication. In *CMC Magazine*. https://www.december.com/cmc/mag/1997/jan/december.html

Garcia, A. C., & Jacobs, J. B. (1999). The Eyes of the Beholder: Understanding the Turn-Taking System in Quasi-Synchronous Computer-Mediated Communication. *Research on Language and Social Interaction*, 32(4), pp. 337--367. https://doi.org/10.1207/S15327973rls3204_2

Jakobson, R. (1960). *Essais de Linguistique Générale*. Paris: Editions de Minuit

Kerbrat-Orecchioni, C. (2003). La communication médiatisée par ordinateur : problèmes de genres et de typologie. *Les genres de l'oral*. Présenté à Université Lumière -Lyon 2. http://gric.univ-lyon2.fr/Equipe1/actes/journees_genre.htm

Levin, A., & Bryan, A. (2008). Web 2.0 Storytelling: Emergence of a New Genre. In *Educause Review*, 43(6), pp. 40-56.

Mangenot, F., & Soubrié, T. (2014). Le web social au service de tâches d'écriture. In *Recherches*, (60), pp. 89--105.

Paveau, M.-A. (2013). Genre de discours et technologie discursive. Tweet, twittécriture et twittérature. Pratiques*, Théories et pratiques des genres* (157-158), pp. 7--30.

Linguistic accommodation in online writing: Pilot study

Lisa Hilte, Reinhild Vandekerckhove, Walter Daelemans

CLiPS, University of Antwerp

lisa.hilte@uantwerpen.be
reinhild.vandekerckhove@uantwerpen.be
walter.daelemans@uantwerpen.be

Linguistic accommodation, i.e. adapting one's language use to one's conversation partner(s), is underresearched in the context of online writing. In this pilot study, we investigate whether and how teenage boys adapt their online writing style to teenage girls and vice versa.

We examine a corpus of more than 400,000 instant messages (> 2.5 million words) produced in Dutch on Facebook Messenger and WhatsApp by 1384 Flemish teenagers. We use generalized linear mixed models (GLMMs) to analyze the occurrences of non-standard chatspeak markers such as expressive typographic markers (e.g. emoji), regional language features and abbreviations. Two conversational settings are compared: same-gender conversations (including only girls or only boys) and mixed-gender conversations (including at least one girl and one boy). We want to find out whether in these mixed-gender settings, boys (resp. girls) will adapt their online writing style to a more prototypically female (resp. male) style.

Our findings reveal significant linguistic differences between girls' and boys' online writing in same-gender settings: in girl-girl conversations, girls use significantly more non-standard features than boys do in boy-boy conversations. However, in mixed-gender settings, there is no significant gender difference. We observe a pattern of *asymmetric gender convergence*, as teenage boys adapt their online writing style much more strongly to girls than vice versa.

In further research, other socio-demographic variables should be included (e.g. age, education). In addition, the non-standard linguistic markers should be de-clustered, as different patterns may emerge for e.g. expressive features and regional language use. Furthermore, we hypothesize that the observed stronger convergence by boys is mainly caused by a stronger increase in the use of *expressive* chatspeak features, which are typically favored by girls, and which may be inserted for purposes such as flirting (see e.g. Hilte, Vandekerckhove & Daelemans 2018).

References

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans (2018). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23(3):293-323.