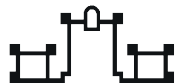


# Proceedings

of the 10<sup>th</sup> International  
Conference on CMC and  
Social Media Corpora for  
the Humanities



September 14 -15, 2023  
University of Mannheim  
Germany



UNIVERSITY  
OF MANNHEIM  
School of Humanities

IDS

LEIBNIZ-INSTITUT FÜR  
DEUTSCHE SPRACHE

Funded by

**DFG**

Deutsche  
Forschungsgemeinschaft  
German Research Foundation

Louis Cotgrove  
Laura Herzberg  
Harald Längen  
Ines Pisetta (eds.)

Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities

14–15 September 2023, University of Mannheim, Germany

Editors: Louis Cotgrove, Laura Herzberg, Harald Längen, Ines Pisetta

Published by: Leibniz-Institut für Deutsche Sprache  
Mannheim, 2023

DOI: <https://doi.org/10.14618/1z5k-pb25>

ISBN: 978-3-937241-95-1

This work is licensed under a Creative Commons “Attribution 4.0. International” license.

Conference website: <https://www.uni-mannheim.de/cmc-corpora2023>

The CMC-Corpora 2023 conference is funded by *Deutsche Forschungsgemeinschaft* under the project number 524949653.



## Preface

Following the successes of the ninth conference in 2022 held in the wonderful Santiago de Compostela, Spain, we are pleased to present the proceedings of the 10th edition of International Conference on CMC and Social Media Corpora for the Humanities (CMC-2023). The focal point of the conference is to investigate the collection, annotation, processing, and analysis of corpora of computer-mediated communication (CMC) and social media.

Our goal is to serve as the meeting place for a wide variety of language-oriented investigations into CMC and social media from the fields of linguistics, philology, communication sciences, media studies, and social sciences, as well as corpus and computational linguistics, language technology, textual technology, and machine learning.

This year's event is the largest so far with 45 accepted submissions: 32 papers and 13 poster presentations, each of which were reviewed by members of our ever-growing scientific committee. The contributions were presented in five sessions of two or three streams, and a single poster session. The talks in these proceedings cover a wide range of topics, including the corpora construction, digital identities, digital knowledge-building, digitally-mediated interaction, features of digitally-mediated communication, and multimodality in digital spaces.

As part of the conference, we were delighted to include two invited talks: an international keynote speech by Unn Røyneland from the University of Oslo, Norway, on the practices and perceptions of researching dialect writing in social media, and a national keynote speech by Tatjana Scheffler from the Ruhr-University of Bochum on analysing individual linguistic variability in social media and constructing corpora from this data. Additionally, participants could take part in a workshop on processing audio data for corpus linguistic analysis. This volume contains abstracts of the invited talks, short papers of oral presentations, and abstracts of posters presented at the conference.

We wish to thank all colleagues who contributed to the conference and proceedings this year for their fascinating and varied presentations, posters, and keynote talks. We would also like to thank the members of the international scientific committee for their support and help in reviewing the many submissions this year. Thanks also go to the Leibniz-Institute for the German Language and the University of Mannheim for providing administrative support and the wonderful locations for the conference this year, and a big thank you to the German Research Foundation (DFG) for their financial contribution to the conference.

We hope that the tenth edition of the conference series can build on the successes of the previous editions and we are looking forward to the next decade of CMC-Corpora conferences!

Mannheim, September 14 2023

On behalf of the organising committee

Jutta Bopp,  
Louis Cotgrove,  
Laura Herzberg,  
Harald Lungen, and  
Andreas Witt





# Committees

## Local Organizing Committee in Mannheim

Jutta Bopp	IDS Mannheim
Louis Cotgrove	IDS Mannheim
Laura Herzberg	University of Mannheim
Harald Lünge	IDS Mannheim
Andreas Witt	University of Mannheim & IDS Mannheim

## International Steering Committee of the Conference series

Steven Coats	University of Oulu
Julien Longhi	Cergy-Pontoise Université
Lieke Verheijen	Radboud University
Reinhild Vandekerckhove	University of Antwerp

## Scientific Committee

Paul Baker	Lancaster University
Adrien Barbaresi	Berlin-Brandenburgische Akademie der Wissenschaften
Michael Beißwenger	University of Duisburg-Essen
Mario Cal Varela	Universidade de Santiago de Compostela
Steven Coats	University of Oulu
Luna DeBruyne	Ghent University
Orphée DeClercq	Ghent University
Francisco Javier Fernández Polo	University of Santiago de Compostela
Jenny Frey	EURAC Research Bolzano
Alexandra Georgakopoulou-Nunes	King's College London
Klaus Geyer	University of Southern Denmark
Aivars Glaznieks	EURAC Research Bolzano
Jan Gorisch	IDS Mannheim
Claire Hardaker	Lancaster University
Iris Hendrickx	Radboud University Nijmegen
Axel Herold	Berlin-Brandenburgische Akademie der Wissenschaften
Lisa Hilde	University of Antwerp
Mai Hodac	Université Toulouse
Wolfgang Imo	University of Hamburg
Paweł Kamocki	IDS Mannheim
Erik-Tjong Kim-Sang	Netherlands eScience Center
Alexander Koenig	CLARIN ERIC
Florian Kunneman	Vrije Universiteit Amsterdam
Marc Kupietz	IDS Mannheim
Els Lefever	Ghent University
Julien Longhi	Cergy-Pontoise Université
Maja Miličević-Petrović	University of Bologna
Nelleke Oostdijk	Radboud University

Céline Poudat  
Thomas Proisl  
Ines Rehbein  
Sebastian Reimann  
Unn Røyneland  
Jan Oliver Rüdiger  
Müge Satar  
Tatjana Scheffler  
Stefania Spina  
Egon Stemle  
Caroline Tagg  
Simone Ueberwasser  
Reinhild Vandekerckhove  
Lieke Verheijen

Université Côte d'Azur  
Friedrich-Alexander-Universität Erlangen-Nürnberg  
University of Mannheim  
Ruhr-Universität Bochum  
University of Oslo  
IDS Mannheim  
Newcastle University  
Ruhr-Universität Bochum  
Università per Stranieri di Perugia  
EURAC Research Bolzano  
The Open University  
University of Zurich  
University of Antwerp  
Radboud University

# Contents

## Keynotes

Unn Røyneland: Eye dialect in social media – practices and perceptions .....	1
------------------------------------------------------------------------------------	---

Tatjana Scheffler: Individual linguistic variability in social media .....	2
----------------------------------------------------------------------------------	---

## Posters

Aminat Babayode, Laurens Bosman, Nicole Chan, Katharina Ehret, Ivan Fong, Noelle Harris, Alissa Hewton, Danica Reid, Maite Taboada and Rebekah Wong: Structural properties of podcasts as an emerging register of computer-mediated communication .....	3
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---

Marie-Louise Bartsch and Irina Mostovaia: Ellipsis of the subject pronoun ich ('I') in German WhatsApp chats: A usage-based approach .....	7
--------------------------------------------------------------------------------------------------------------------------------------------------	---

Elizaveta Kibisova: Building corpora of Russian fake and genuine news for linguistic analysis .....	9
-----------------------------------------------------------------------------------------------------------	---

Aenne Knierim: Tracing Perceptions of Black History by Comparison of Two Corpora .....	10
----------------------------------------------------------------------------------------------	----

Katharina Pabst, Aida Alanzi, Johanna Aminoff, Raisa Tayib and Derek Denis: Zooming in on emerging norms: Preliminary findings from a cross-linguistic investigation of videoconferencing .....	13
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

Sebastian Reimann, Lina Rodenhausen, Tatjana Scheffler and Frederik Elwert: ChrisTof: A Novel Corpus of Christian Online Forums .....	15
---------------------------------------------------------------------------------------------------------------------------------------------	----

Hannah J. Seemann, Sara Shahmohammadi, Tatjana Scheffler and Manfred Stede: Building a Parallel Discourse-annotated Multimedia Corpus .....	17
---------------------------------------------------------------------------------------------------------------------------------------------------	----

Sarah Steinsiek: Negotiating knowledge in cooperative learning scenarios: a multimodal approach to practices of computer-mediated and face-to-face communication in the university classroom .....	18
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

Jenia Yudytska: "Linguistic features, device affordances, and contextual factors: A mixed-methods, two-corpora approach""	20
Yinglei Zang: Deontic Authority in Computer-mediated Communication Between University Teachers and Students: A Comparative Study of German and Chinese	21
<b>Talks</b>	
Selenia Anastasi, Tim Fischer, Florian Schneider and Chris Biemann: IDA - Incel Data Archive: a multimodal comparable corpus for exploring extremist dynamics in online interaction.	23
Tianyi Bai: The Reply Function in WhatsApp Chat Communication	29
Michael Beißwenger, Eva Gredel, Lena Rebhan and Sarah Steinsiek: Ellipsis Points in Messaging Interactions and on Wikipedia Talk Pages	33
Laura Bothe: The representation of the 'Jew' as enemy in French public Telegram channels within the identitarian-conspiratorial milieu	39
Bruno Machado Carneiro, Michele Linardi and Julien Longhi: "Studying Socially Unacceptable Discourse Classification (SUD) through different eyes: ""Are we on the same page ?"""	45
Steven Coats: A Pipeline for the Large-Scale Acoustic Analysis of Streamed Content	51
Louis Cotgrove: "megageil, mega geil, and voll mega: Intensification in YouTube comments""	55
Selcen Erten: Exploring register variation in Turkish web corpus	60

Annamaria Fabian and Igor Trost,:	
"Digital Corpus Linguistic Analysis of the Language of Inclusion, Discrimination and Exclusion of people with disability in social media – in a German corpus of 214.926 Tweets on #disability and #inclusion between 2007-2023""	65
Anne Ferger, Andre Frank Krause and Karola Pitsch:	
Workflows and Methods for Creating Structured Corpora of Multimodal Interaction	73
Francisco Javier Fernández Polo:	
Balancing expert and peer-student identities in online discussion forums	78
Shir Finkelstein and Hadar Netz:	
"Hebrew level: Bibist.": Online Hebrew language corrections as a tool for "civilized" bashing	83
Carolina Flinz, Eva Gredel and Laura Herzberg:	
A Corpus study on the negotiation of pronominal address on talk pages of the German, French, and Italian Wikipedia	86
Florian Frenken:	
A Multivariate Register Perspective on Reddit: Exploring Lexicogrammatical Variation in Online Communities	91
Prakhar Gupta, Lliana Doudot, Romain Loup and Aris Xanthos:	
Collecting and de-identifying half a million WhatsApp messages	96
Teemu Helenius,:	
Acquiring, Analyzing, and Understanding Multimodal TikTok Short Video Data: The Case of Online Sex Worker Visibility Management	102
Sangwan Jeon:	
MigrTwit Corpora. (Im)Migration Tweets of French Politics.	108
André Frank Krause, Anne Ferger and Karola Pitsch:	
Anonymization of Persons in Videos of Authentic Social Interaction: Machine Learning Model Selection and Parameter Optimization.	112
Lothar Lemnitzer and Antonia Hamdi:	
"Also ehrlich" – From adjectival use to interactive discourse marker	118

Rosa Lorés: The recontextualization of expert knowledge: intertextual patterns in digital science dissemination	124
Harald Lungen and Laura Herzberg: Studying the distribution of reply relations in Wikipedia talk pages	131
Martti Mäkinen: MMWAH! Compiling a Corpus of Multilingual / Multimodal WhatsApp Discussions by Swedish-speaking Young Adults in Finland	136
Rachel McCullough, Daniel Drylie, Mindi Barta and Daniel Smith: CoDEC-M: the multi-lingual Manosphere subcorpus of the Corpus of Digital Extremism and Conspiracies	140
Iliia Moshnikov and Eugenia Rykova: Little Big Data: Karelian Twitter Corpus	142
Anastasiia Piroh: Multimodal Intertextual Practices in Video Film Reviews	148
Ana Eugenia Sancho-Ortiz: Scientific communication on social media: Analysing Twitter for knowledge recontextualisation	154
Ulrike Schneider and Oliver Watteler: Can I Publish my Social Media Corpus? Legal Considerations for Data Publication	160
Laurel Stvan: Collecting Health Memes for a Subcorpus of Peer Health Discourse	166
Ludovic Tanguy, Céline Poudat and Lydia-Mai Ho-Dac: Specific behaviours in Wikipedia talk pages: some insights from extreme cases	171
Ralia Thoma: “Don’t be afraid of Greeklish”: Adolescent students’ transliteration practices	176
Eva Triebel: Not an expert, but not a fan either. A corpus-based study of negative self-identification as epistemic index in web forum interaction.	182

Reinhild Vandekerckhove, Sarah Bernolet, Astrid De Wit and Tanja Mortelmans: Towards a more inclusive approach of digital literacy: social media writing at an older age .....	187
Jiayi Zhou: Phonetic Metaphor of Chinese Emojis: An Approach of Neologism Formation .....	190
<b>Author index</b> .....	194
<b>Keyword index</b> .....	195



# Eye dialect in social media – practices and perceptions

**Unn Røyneland**

University of Oslo

E-mail: [unn.royneland@iln.uio.no](mailto:unn.royneland@iln.uio.no)

## Abstract

Studies of linguistic practices in social media show that people make use of a broad linguistic repertoire in their digital communication (e.g. Androutsopoulos 2021; Deumert 2014; Cutler & Røyneland 2018; Thurlow & Mroczek 2011). New digital technologies have made possible playful, creative, reflexive, self-conscious, and non-standard ways of using language, where people deploy their linguistic resources to project local as well as trans-local orientations and affiliations, negotiate identities, take stances, mark attitudes and ideological convictions. The use of abbreviations, emojis, deliberate misspellings, initialisms, rebus spellings, and dialect features, for instance, are very common and even expected in some digital spaces and on some platforms, whereas standard orthography and spelling would be expected in others. Up until now relatively few studies have focused specifically on the use of dialect or dialect features in the digital sphere. This will be the main focus of my talk. I will discuss different methodologies in collecting and analyzing experimental, survey, and authentic written SoMe data, using studies of adolescents in Norway as a case. Questions we ask in these studies are how dialect is used in digital writing, which features, when and by whom, and last but not least for what purposes. In addition, we ask what counts as ‘dialect writing’ and whether standard and non-standard samples are perceived and processed differently. In my presentation I will discuss some of the findings while also considering to what extent variationist theory and methodology may be useful in handling highly linguistically mixed SoMe data.

**Keywords:** digital identity, dialect, social media, youth language

## References

- Androutsopoulos, J. (ed.) 2021. Digital language practices: media, awareness, pedagogy. Special Issue, *Linguistics & Education* 62.
- Cutler, C. & U. Røyneland 2018. (eds.) *Multilingual Youth Practices in Computer Mediated Communication*. Cambridge University Press.
- Deumert, A. 2014. *Sociolinguistics and Mobile Communication*. Edinburgh University Press.
- Thurlow, C. & K. Mroczek 2013. (eds.). *Digital discourse: Language in the new media*. Oxford University Press.

# Individual linguistic variability in social media

**Tatjana Scheffler**

Ruhr-Universität Bochum

E-mail: [tatjana.scheffler@rub.de](mailto:tatjana.scheffler@rub.de)

## Abstract

Computer-mediated language has become a popular source of data for analyses in linguistics and social science, aided by convenient access to large-scale ad-hoc corpora. In this talk I will present several case studies to support two main points. First, social media present a varied source of informal, spontaneous, situated discourse that can inform linguistic theory (amongst other lines of research). Second, it is important to study linguistic behavior by actors across media and domains, since both "computer-mediated communication" as a whole as well as specific social media show considerable within-corpus variation that is partially due to intra-speaker variability. Finally, I will discuss how to construct corpora that support this kind of research, from practical, ethical, and sustainability perspectives.

**Keywords:** CMC, discourse analysis, variation, corpus-building

# Structural linguistic characteristics of podcasts as an emerging register of computer-mediated communication

Aminat Babayode<sup>1</sup>, Laurens Bosman<sup>1</sup>, Nicole Chan<sup>1</sup>, Katharina Ehret<sup>1,2</sup>, Ivan Fong<sup>1</sup>,  
Noelle Harris<sup>1</sup>, Alissa Hewton<sup>1</sup>, Danica Reid<sup>1</sup>, Maite Taboada<sup>1</sup>, Rebekah Wong<sup>1</sup>

<sup>1</sup>Department of Linguistics, Simon Fraser University, Canada

<sup>2</sup>Department of English, University of Freiburg, Germany

Corresponding authors: kehret@sfu.ca, mtaboada@sfu.ca

## Abstract

Podcasts, a relatively recent audio medium, have risen in popularity since their initial appearance in the mid-2000s. Yet, little is known about their structural linguistic characteristics and their relation to other registers. Addressing this gap in the literature, we apply Biber-style multidimensional analysis (MDA) to a representative sample of Spotify podcast transcripts and compare their structural linguistic characteristics to those of selected computer-mediated registers (e.g., informational blog, interview) as well as traditional spoken registers (e.g., broadcast, conversation). Our results reveal that, while podcasts share some linguistic characteristics with traditional spoken registers such as broadcast discussion and unscripted speech, they are unlike any of the analysed registers. In fact, they exhibit unique structural characteristics combining features of involved spoken language with some features typical of informational production and narration. In short, we show that podcasts are a newly emerging register of computer-mediated communication.

**Keywords:** computer-mediated communication, register analysis, corpus linguistics, multidimensional analysis, podcasts

## 1. Podcasts as a new medium

Podcasts are a new audio-based medium, similar to radio, television, and other traditional media facilitating the sharing and broadcasting of content to large audiences (Levinson, 2013). Originally, podcasts were intended to convey information and act as a source of entertainment (Nurekeshova, 2016). Due to their relevance to diverse contexts, podcasts are a versatile form of media, with podcasts that are similar to traditional interview shows, news and politics, audiobooks, music, games, plays, and educational shows. Their broad appeal, however, has resulted in an emerging set of practices that may differ from traditional radio (Berry, 2016). Despite their popularity, little research on the characteristic linguistic features of podcasts exists. Addressing this gap, we apply Multidimensional Analysis (MDA) (Biber, 1988) to provide insight into how linguistic features associated with the conversational style of podcasts differentiate them from other types of broadcasts and other emerging registers of computer-mediated communication.

## 2. Register variation and MDA

Biber and Conrad (2001) define register as the result of linguistic variation in the lexical and grammatical choices that language users make as appropriate to the context of usage. The tool of choice for analysing register variation is Multidimensional Analysis. MDA is a multivariate statistical technique based on the frequency and co-occurrence of lexico-grammatical features, and examines how the co-occurrence patterns correlate with particular registers (Biber, 1988; Biber and Egbert, 2018). Through dimensionality reduction, MDA allows us to abstractly interpret linguistic features as representing the underlying communicative functions of the texts analysed.

## 3. Podcast transcripts as corpus

Podcasts are typically available as audio files, but also in transcript format. The data used in this study is a subset

of the English part of the Spotify Podcasts Dataset (Clifton et al., 2020). We classified the podcasts by topic as well as length and selected the top 20 categories by number of podcasts. These include topics such as Arts, Business, Comedy, History, Religion & Spirituality, Science, Sports, or True Crime. We then divided sorted each topic into 4 bins by length (up to 15 minutes in length; 15-30 minutes; 30-60 minutes; >60 minutes), from which we sampled 10% from each bin across the top 20 topics. Our final corpus counts 9,789 podcast transcripts and 64,239,291 words.

MDA involves exploring the features of the register in question in comparison to other registers to situate the register of interest in a space of linguistic variation. For this comparison, we draw on different corpora. First, we use selected registers of the Corpus of Online Registers of English (CORE) (Biber and Egbert, 2018)), to compare podcasts to other registers of computer-mediated communication (e.g., interactive discussion, informational blog). Second, we use the British National Corpus (BNC) (Aston and Burnard, 2020)) as a source for traditional spoken registers (e.g., conversation, broadcast discussion), the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al., 2000)) for conversation in a North American context, and the English Pear Stories<sup>1</sup> as a source of oral narratives.

In total, we analyse 9 different traditional registers and 10 registers of computer-mediated communication totalling over 27 million words, comparing them to the 64 million words in the podcasts (see Appendix, Table 1).

## 4. Podcasts as an emerging register

### 4.1. Podcasts and computer-mediated registers

In the MDA of podcasts and computer-mediated registers in CORE two well-defined dimensions emerge; the third dimension was extracted for statistical reasons but is not

<sup>1</sup><http://www.pearstories.org/english/english.htm>

linguistically interpretable (see Appendix, Table 2). The first dimension, “Involved vs. informational discourse” has two poles. On the positive pole cluster features typical of spontaneous spoken and involved language such as present tense verbs, contractions, and private verbs. The negative pole is defined by only a handful of features, all of which indicate an informational style: nouns, prepositions and perfect aspect. Dimension 2 comprises only positive features, namely, nominalisations, average word length, and predicative adjectives. These features are typical of abstract-informational language and can together with secondary features such as THAT-relative clauses and complements be interpreted as representing “Abstract-informational elaboration”.

Looking at the distribution of registers on these two dimensions, we find that all the written registers and interactive discussion are positioned on the negative pole of Dimension 1, i.e., they are representative of informational discourse. On the positive, involved pole, we find the spoken registers podcasts, formal speech, spoken, and interview. As a matter of fact, podcasts emerge as the most involved register in this dataset. On Dimension 2, podcasts are located somewhere in between, along with interview, while formal speech and informational blog are most representative of abstract-informational elaboration. Thus, podcasts clearly emerge as a spoken register and one unlike all the other computer-mediated registers (Appendix, Figure 1). They are strongly characterised by features of spontaneous spoken and involved language and, to some extent, features of abstract-informational elaboration.

#### 4.2. Podcasts and traditional spoken registers

The MDA comparison of podcasts and traditional spoken registers comprises three variational dimensions (see Appendix, Table 3). The first dimension, “Involved vs. informational production” is defined by features typical of spontaneous spoken and involved language such as contractions, emphatics, and first personal pronouns on the positive pole. On the negative pole, it is defined by the co-occurrence of average word length, nouns, attributive adjectives, and other features indicative of information-focused production. The second dimension which we label “Narrative” is largely defined by positive features typically associated with narration: past tense, third person pronouns and perfect aspect. Dimension 3 “Abstract elaboration” consists only of a positive pole which comprises indicators of elaboration and abstract description such as THAT-verb complements, THAT-relatives, and predicative adjectives.

The overall distribution of registers (Appendix, Figure 2) confirms this interpretation of the dimensions, for instance, broadcast news is highly informational while conversation is involved; interviews and oral narrative load high on the narrative dimension and broadcast documentary is information-elaborate. Where, then, are podcasts positioned? Interestingly, none of the three dimensions represents our podcast data very well and it is located in the middle. In terms of other registers it resembles oral narratives, interviews, (unscripted) speech and broadcast discussion across the three dimensions. Hence, podcasts do share some features with these registers but are also clearly unlike any of them. Rather, they uniquely combine features of nar-

ration, spontaneous speech and informational production.

### 5. Conclusion

This paper presented a multidimensional analysis of podcasts as an emerging register of computer-mediated communication. Comparing podcasts to a set of written and spoken computer-mediated registers from the Corpus of Online Registers of English and well-known corpora of broadcasts, conversations, and narratives, we show that podcasts are firmly a spoken register, yet, unlike all the other computer-mediated registers. Our analysis of podcasts and traditional spoken registers confirms this finding: podcasts are clearly a newly emerging register. Precisely, podcasts exhibit some similarities with a range of spoken registers, i.e. interviews, (unscripted) speech, oral narratives and broadcast discussion. Hence, they are characterised by a unique set of features and combine features of on-line spontaneous production, narration, and informational production. This characterisation dovetails with the intended purpose of podcasts as a source of both information and entertainment (Nurekeshova, 2016). It is this versatility of podcasts that makes them unique but probably also makes them a register with a comparatively large degree of internal variability (like the registers broadcast and letters). The natural next step, then, is to explore the extent of register-internal variability and further detail the lexico-grammatical features of the emerging podcast (sub)register(s). Last but not least, despite the fact that our data samples English-language podcasts only, our findings constitute a first step towards understanding and describing podcasts in general.

### 6. Acknowledgements

We are grateful for helpful comments and suggestions by two anonymous reviewers.

### Bibliographical references

- Aston, G. and Burnard, L. (2020). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.
- Berry, R. (2016). Podcasting: Considering the evolution of the medium and its association with the word ‘radio’. *Radio Journal: International Studies in Broadcast & Audio Media*, 14(1):7–22, April.
- Biber, D. and Conrad, S. (2001). Register variation: A corpus approach. In Deborah Schiffrin, et al., editors, *The Handbook of Discourse Analysis*, pages 175–196. Blackwell.
- Biber, D. and Egbert, J. (2018). *Register Variation Online*. Cambridge University Press.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.
- Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G. J., Karlgren, J., Carterette, B., and Jones, R. (2020). 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917.
- Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., and Martey, N. (2000). Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium*.
- Levinson, P. (2013). *New New Media*. Pearson, 2nd edition.
- Nurekeshova, G. R. (2016). Podcasting as a technical way of interactive communication of XXI century. *European Journal of Natural History*, pages 112–116.

## Appendix

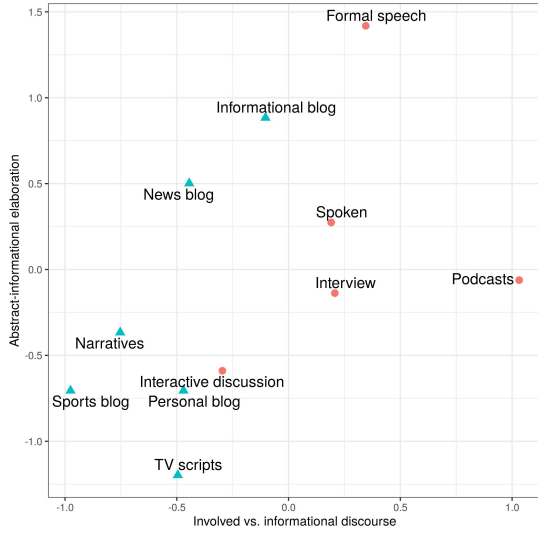


Figure 1: Podcasts compared to other registers of computer-mediated communication in the Corpus of Online Registers of English. Positive values on Dimension 1 indicate involved discourse; negative values on Dimension 1 indicate narrative description. Red dots index spoken, green triangles index written registers.

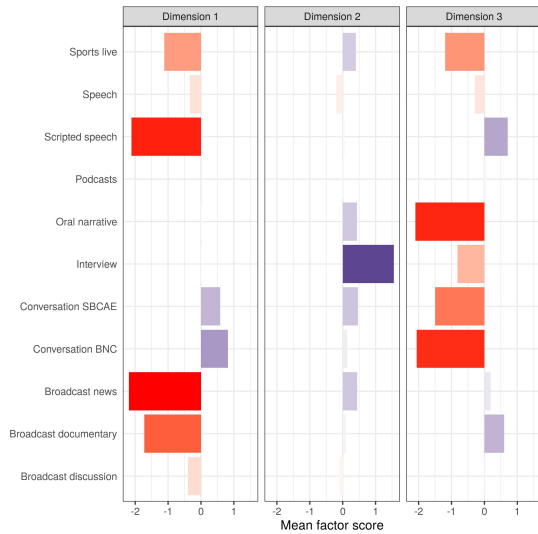


Figure 2: Distribution of podcasts and traditional spoken registers on the three dimensions. Colour intensity indicates strength of mean factor scores. Red bars indicate negative values; blue bars indicate positive values.

Register	Mode	Corpus	Words
Broadcast discussion	spoken	BNC	666,098
Broadcast documentary	spoken	BNC	37,496
Broadcast news	spoken	BNC	225,024
Conversation	spoken	BNC	3,836,745
Interview	spoken	BNC	111,155
Scripted speech	spoken	BNC	164,244
Unscripted speech	spoken	BNC	410,690
Sportslive	spoken	BNC	29,957
Formal speech	spoken	CORE	80,109
Informational blog	written	CORE	2,141,271
Interactive discussion	spoken	CORE	3,099,725
Interview	spoken	CORE	451,593
Narrative	written	CORE	424,614
News report/blog	written	CORE	9,806,239
Personal blog	written	CORE	3,264,463
Spoken	spoken	CORE	224,703
Sports report	written	CORE	2,729,925
TV scripts	written	CORE	32,502
Conversation	spoken	SBCE	209,308
Oral narratives	spoken	Pear Stories	16,149
TOTAL			27,962,010

Table 1: Registers by corpus, mode (written vs. spoken) and word count.

Dimension 1: Involved vs. informational discourse	
Present tense verbs	0.857
Contractions	0.761
Demonstrative pronouns	0.752
Private verbs	0.749
Causal subordinators	0.663
Emphatics	0.638
BE as main verb	0.607
Demonstratives	0.6
Second person pronouns	0.588
Analytic negation	0.573
Hedges	0.558
Adverbs	0.554
First person pronouns	0.52
Pronoun IT	0.533
THAT deletion	0.446
Pro-verb DO	0.42
Predicative adjectives	0.413
WH-clauses	0.392
Conditional subordinators	0.308
—	—
Nouns	-0.831
Prepositions	-0.637
Average word length	-0.391
Present participle clauses	-0.334
Dimension 2: Abstract-informational elaboration	
Nominalizations	0.783
Average word length	0.711
Attributive adjectives	0.381
Phrasal coordination	0.316

Table 2: Dimensions and features with significant loadings  $\geq |0.3|$  for the podcast and CORE data. Positive loadings indicate co-occurrence of the features; negative loadings indicate complementary distribution. Crossloading features with the same polarity and uninterpretable dimensions are excluded.

Dimension 1: Involved vs. informational production	
Contractions	0.846
First person pronouns	0.699
Analytic negation	0.661
Private verbs	0.625
Present tense verbs	0.615
THAT deletion	0.595
Pronoun IT	0.592
Demonstrative pronouns	0.579
Emphatics	0.547
Causal subordinators	0.544
BE as main verb	0.54
Pro verb DO	0.462
WH-clauses	0.412
Hedges	0.4399
Adverbs	0.387
Predicative adjectives	0.309
—	—
Average word length	-0.935
Nouns	-0.826
Prepositions	-0.779
Attributive adjectives	-0.705
Nominalizations	-0.655
Phrasal coordination	-0.612
Present participle WHIZ deletion	-0.467
Past participle WHIZ deletion	-0.464
BY-passives	-0.424
Passives	-0.362
Conjunctions	-0.359
Gerunds	-0.342
Dimension 2: Narrative	
Past tense verbs	0.956
Third person pronouns	0.567
Perfect aspect	0.406
Dimension 3: Abstract elaboration	
THAT verb complements	0.487
Nominalisations	0.472
THAT-relatives (obj.)	0.457
Demonstrative pronouns	0.447
Average word length	0.37
Split auxiliaries	0.325
THAT-relatives (subj.)	0.322
Predicative adjectives	0.314
TO infinitives	0.301

Table 3: Dimensions and features with significant loadings  $\geq |0.3|$  for the podcast and traditional spoken data. Positive loadings indicate co-occurrence of the features; negative loadings indicate complementary distribution. Crossloading features with the same polarity and uninterpretable dimensions are excluded.



# Ellipsis of the subject pronoun *ich* ('I') in German WhatsApp chats: A usage-based approach

Marie-Louise Bartsch, Irina Mostovaia

Department of Language, Literature and Media (SLM I),

Institute of German Studies, University of Hamburg

E-mail: [marie-louise.bartsch@uni-hamburg.de](mailto:marie-louise.bartsch@uni-hamburg.de), [irina.mostovaia@uni-hamburg.de](mailto:irina.mostovaia@uni-hamburg.de)

## Abstract

Previous research into ellipsis in “keyboard-to-screen communication” (Jucker/Dürscheid 2012) shows that the first-person singular subject pronoun is frequently omitted in SMS text messages: for example, the subject omission rate reported for this pronoun in previous studies based on German (Androutsopoulos/Schmidt 2002: 69) and Swiss German text messages (Frick 2017: 88-89) is 60% and 59%, respectively.

In their study on argument drop in (Swiss) German WhatsApp messages drawn from the corpus *What's up, Switzerland?*, Stark/Meier (2017) also observe a tendency for the first-person singular subject pronoun to be omitted more frequently than other subject pronouns: “[t]he majority of omissions (73.9%) occur with 1st pers. sg.” (Stark/Meier 2017: 241). However, the total subject omission rate of 18% provided by Stark/Meier (2017: 240) reveals that subject pronouns seem to be omitted less frequently in WhatsApp messages than in SMS text messages (18% compared to 53% or 54 % in Swiss German and German data respectively, cf. Frick 2017: 88; Androutsopoulos/Schmidt 2002: 69). This, however, raises the question of whether this difference between frequencies of subject omissions is linked to the affordances of WhatsApp chats (cf. Androutsopoulos 2023) or to other factors, e.g., to “strong interferences with local Swiss German dialects” (Stark/Meier 2017: 226) which Stark and Meier have observed in their data.

Built on previous research into ellipsis in spoken and written interactions, this contribution presents the results of a study on the omission of the singular first-person pronoun *ich* ('I') based on 706 German WhatsApp chats with 31,525 messages (243,549 tokens) drawn from the *Mobile Communication Database 2* (<https://db.mocoda2.de>). Using the *MAXQDA* software, 1,000 occurrences of the omitted first-person singular subject pronoun – as well as 1,000 occurrences of *ich* – identified in this corpus have been manually annotated with a range of formal and functional features. In this way we examine whether particular extralinguistic (e.g., message length, cf. Imo [2015a] and Dürscheid [2016: 456] for the tendency to shorten WhatsApp messages), syntactic (e.g., (non)occurrence of subject omissions in main and subordinate clauses and their position in the clause, cf. Haegeman 2013 and Stark/Meier 2017) and pragmatic (e.g., (non)occurrence of subject omissions in particular actions such as assessments, responses or in serious utterances, cf. Auer [1993: 207] and Androutsopoulos/Schmidt [2002: 69-70]) factors play a role for subject omissions in German WhatsApp messages.

These annotations serve as a basis for the analysis with the statistical software *R* in order to examine whether omissions of subject pronouns are formally and/or functionally motivated and whether some structures including omitted subject pronouns can be interpreted as constructions in the sense of Interactional Construction Grammar (cf. Deppermann 2006; Imo 2015b).

The study was conducted within the context of the research unit *Practices for referring to persons: usage-based approaches to personal, indefinite, and demonstrative pronouns* (FOR 5317) funded by the German Research Foundation.

## References

- Androutsopoulos, J. (2023). Kontextualisierung digital: Repertoires und Affordanzen in der schriftbasierten Interaktion. In S. Meier-Vieracker, L. Bülow, K. Marx & R. Mroczynski (Eds.), *Digitale Pragmatik. Digitale Linguistik*, vol 1. Berlin, Heidelberg: J.B. Metzler, pp. 13–38. [https://doi.org/10.1007/978-3-662-65373-9\\_2](https://doi.org/10.1007/978-3-662-65373-9_2).
- Androutsopoulos, J., Schmidt, G. (2002). SMS-Kommunikation. Ethnografische Gattungsanalyse am Beispiel einer Kleingruppe. *Zeitschrift für Angewandte Linguistik*, 36, pp. 49–80.
- Auer, P. (1993). Zur Verbspitzenstellung im gesprochenen Deutsch. *Deutsche Sprache*, 3, pp. 193–222.
- Deppermann, A. (2006). Construction Grammar – Eine Grammatik für die Interaktion? In A. Deppermann, R. Fiehler & T. Spranz-Fogasy (Eds.), *Grammatik und Interaktion*. Radolfzell: Verlag für Gesprächsforschung, pp. 43–65.
- Dürscheid, Ch. (2016). Neue Dialoge – alte Konzepte? Die schriftliche Kommunikation via Smartphone. *Zeitschrift für germanistische Linguistik*, 44(3), pp. 437–468.
- Frick, K. (2017). *Elliptische Strukturen in SMS. Eine korpusbasierte Untersuchung des Schweizerdeutschen*. Berlin: De Gruyter.
- Haegeman, L. (2013). The syntax of registers: diary subject omission and the privilege of the root. *Lingua*, 130, pp. 88–110.
- Imo, W. (2015a). Vom Happen zum Häppchen... Die Präferenz für inkrementelle Äußerungsproduktion in internetbasierten Messengerdiensten. *Networx*, 69. <https://doi.org/10.15488/2960>.
- Imo, W. (2015b). Interactional Construction Grammar. *Linguistics Vanguard*, 1(1), pp. 1–9.
- Jucker, A. H., Dürscheid, Ch. (2012). The linguistics of keyboard-to-screen communication. A new terminological framework. *Linguistik Online*, 56(6), pp.

39–64. <https://doi.org/10.13092/lo.56.255>.

Stark, E., Meier, P. (2017). Argument Drop in Swiss WhatsApp Messages. A Pilot Study on French and (Swiss) German. *Zeitschrift für französische Sprache und Literatur*, 127(3), pp. 224–252.



# Building corpora of Russian fake and genuine news for linguistic analysis

Elizaveta Kibisova

University of Oslo

E-mail: [lisa.kibisova@gmail.com](mailto:lisa.kibisova@gmail.com)

## Abstract

The speed with which fake news spreads online and people's resistance to change their minds continues to be a growing problem (Mosleh et al. 2021). Addressing stylistic and grammatical features of fake news is one of the promising lines of research (Grieve and Woodfield 2023). In Russia, intertwined with the limitations imposed on the freedom of speech, the issue has become particularly pressing during the Covid-19 pandemic and the invasion of Ukraine.

This work describes the challenges of building a corpus of Russian fake news and the matching reference corpus of genuine news for the purpose of linguistic analyses, including comparing patterns of grammatical variation based on the multidimensional register analysis (Biber 1988). The primary aim of creating the corpora is to investigate the language and style of fake news in Russian. A secondary goal is to use the datasets for the improvement of the fake news detection through the automation of the defining linguistic features.

The building of the Russian fake news corpora is a unique process with its challenges and implications. Unlike similar English corpora, the Russian fake news datasets consist mostly of social media texts, predominantly from Telegram and Facebook, and not of well-known news outlets. This is due to the tendency of the news outlets to reproduce false claims by citing or paraphrasing other sources to avoid legal responsibility. Investigations, performed by fact-checkers, usually lead to original messages in smaller and mostly anonymous SM-based outlets.

To date, the fake news corpus is a dataset of over 140 000 tokens, claims veracity confirmed by carefully chosen fact-checking agencies. The compilation of the reference (genuine news) dataset is in early stages. The need to control for source, register, size, authorship, and other variables makes it a demanding but rewarding task, with the unique well-balanced and ready for exploration datasets as a result.

**Keywords:** fake news, Russian, social media

## References

- Biber, D. (1988) *Variation across Speech and Writing*, Cambridge: Cambridge University Press.
- Grieve, J., & Woodfield, H. (2023). *The Language of Fake News (Elements in Forensic Linguistics)*. Cambridge: Cambridge University Press.
- Mosleh, M., Martel, C., Eckles, D. and Rand, D. (2021) "Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment," in *Proceedings of CHI '21*, pp. 1–13, Yokohama, Japan: Association for Computing Machinery.

# Tracing Perceptions of Black History by Comparison of Two Corpora

Aenne Knierim

Universität Hildesheim  
Universitätsplatz 1, 31141 Hildesheim  
{surname}@uni-hildesheim.de

## Abstract

In this study, modern perspectives on Black history are contrasted with historical documents through close and distant reading. A Twitter corpus was created using the hashtag #BlackHistoryMonth. It was then examined through topic modeling using BERTopic. Based on the results, thematically matching historical sub-corpora were assembled using documents from the "BWAT- Black Writing and Thought Collection" at the University of Chicago. To ensure the linguistic comparability of the corpora, linguistic measures such as the type-token ratio are applied. The results show that the topics of #BlackHistoryMonth discourse span almost all areas of life and often involve historical figures, indicating that Black memory culture is associated with Black individuals rather than historical events. In contrast to the historical narratives, the tweets show that African Americans follow the white American doctrine of national heroism. The tweets also include criticism of recent critical race theory legislation and call for alternative methods of teaching black history, such as visiting memorials.

**Keywords:** Black History Month, Mixed-Methods, Black History, Corpus Analysis, Topic Modeling

## 1. Introduction

Social media provides new ways to research minority culture. The accessibility of social media platforms as well as structuring objects like hashtags simplify public opinion giving and allow for a broad audience. Today, Twitter has become a space for the African American community, resulting in what scholars call Black Twitter. Moreover, Critical Race Theory found that racism is socially constructed and mediated, among other things, by tweets, blog posts and social media (Delgado et al., 2017). In this paper, the author comparatively analyses Black perceptions of African American history, structured by computationally generated topics drawn from Twitter and contrasts them to a corpus of historical documents written by Black authors. The author compiled a corpus of tweets with the hashtag #BlackHistoryMonth. Tweets with the hashtag are the digital celebration of Black History Month, an annual celebration of achievements by African Americans throughout history. To the best of the author's knowledge, no research exists about #BlackHistoryMonth which is why an explorative approach on the subject is adopted.

## 2. Methodology

### 2.1. Corpus Creation

By use of the SocialScrapr tool<sup>1</sup>, the author created a corpus of 100,000 tweets with the hashtag #BlackHistoryMonth spanning the duration of Black History Month, February 1st until February 28th, 2022. What "we certainly don't have from the past are detailed and large-scale automatic recordings of cultural behavior in large numbers" (Manovich, 2020) which is why the historical corpora include merely 22,000 documents. The Black Writing and Thought Collection (BWAT-collection) contains texts which date from the early 1700s to the 2000s<sup>2</sup>. It brings together "several

collections of works by Black authors, including corpora of dramatic writing, fiction and folktales, and non-fiction works such as interviews, journal articles, speeches, essays, pamphlets, and letters"<sup>3</sup>. The historical corpora were compiled based on the results of topic extraction from the Twitter corpus. The methodology is detailed in figure 1. As every topic consisted of five topic words, these terms were used as search terms to find results in the BWAT-collection. All search results are authored by African Americans.

### 2.2. Pre-Processing

Deviant linguistic forms that are typical for social media language make it important to clean the data. Stopwords, punctuation, numbers, whitespace, and leftover markup were removed. Additionally, floats were eliminated and all words were lowercased.

### 2.3. Statistical Measures for Text Comparison

Statistical measures are applied in order to permit comparison between all corpora; examples are keyword frequencies, type-token-density, the average sentence length per corpus and the number of distinct word types per corpus.

### 2.4. Creation of topic words using BERTopic

A challenge for topic modeling is the length of tweets which is limited to 280 characters. The author chose to employ BERTopic for the topic modeling due to its ability to handle sparse data. The author extracted the 20 most frequent topics using BERTopic (Grootendorst, 2022) to understand which subjects were mostly discussed during #BlackHistoryMonth 2022.

A weakness of the topic representation is that it is generated from a bag-of-words (Grootendorst, 2022). Therefore, words are likely to be related which reduces topic diversity. This shows in our results, as words with the same stem appear, as for instance in topic 9 that contains the topic words

<sup>1</sup><https://socialscrapr.io/information>, last accessed March 21st, 2023

<sup>2</sup>[textual-optics-lab.uchicago.edu/black\\_writing](https://textual-optics-lab.uchicago.edu/black_writing), last accessed on November 14th, 2022

<sup>3</sup>[textual-optics-lab.uchicago.edu/black\\_writing](https://textual-optics-lab.uchicago.edu/black_writing), last accessed on November 14th, 2022

*slavery, enslaved, slave*. Of course, one could argue that the topic words still show some semantic breadth, as they describe the person Frederick Douglass, slavery as an institution, as a person (slave) and as a state of being (enslaved).

### 3. Results

The mixed-methods approach employed in this research makes use of the two levels of analysis available: computational distant reading and hermeneutic analysis with the help of Critical Race Theory. Topic modeling is used to find the twenty most prevalent topics within the #BlackHistoryMonth corpus. Since Black Twitter did not exist in the past, other forms of media that have evolved from or have been constructed by Black networks in the past, such as pamphlets, essays, or newspaper articles have been used to document the period between 1700 and the 2000s. The methodological approach is detailed in Figure 1 in the appendix.

#### 3.1. Statistical Measures

Compared to the historic subcorpora, tweets on average have the shortest sentence length. In contrast, tweets on average have the largest number of different word types. In comparison, the twitter corpus also has a relatively high type-token ratio. Almost all documents from the historical corpora are written by male authors. Thus, this paper mostly considers the male historical perspective. Due to Twitter's privacy policy, there was no data about the tweet authors' gender. This limits the comparability between the historical corpora and the Twitter corpus.

#### 3.2. Topic Modeling

This year's theme of official Black History Month was Black Health and Wellness. Five out of the nineteen interpretable topics name a famous African American woman which results in 26 percent. For example, discourse revolved around Ketanji Jackson, he first Black woman to hold office for the Supreme Court in 7th of April, 2022<sup>4</sup>. War is also a prevalent subject on Black Twitter during #BlackHistoryMonth. Three topics show discussions around equality and inclusion at the workplace, as well as African American leadership are. Next to this, the three abolitionists Frederick Douglass, Martin Luther King, and Malcolm X were subject to the discourse. There are three cultural topics. Interestingly, two of the cultural topics featured the Black poets Maya Angelou and Langston Hughes. Topic 7 pictures baseball and its player Jackie Robinson who was the first Black sportsman to be signed in Major League Baseball. Some topics had to be excluded from the analysis, as some modern topics had no "counterpart" in the historical corpora.

### 4. Close Reading Analysis

In this chapter, selected #BlackHistoryMonth topics are compared to historical documents of the same topic with close reading.

<sup>4</sup>[nytimes.com/spotlight/ketanji-brown-jackson](https://www.nytimes.com/spotlight/ketanji-brown-jackson), last accessed on November 30th, 2022

In the #BlackHistoryMonth corpus, one third of tweets contain the tokens "firsts" or "1st" and among them, 24.4% of tweets contained the word "woman". Moreover, tweets about people who were born enslaved often tell stories about "firsts". This shows that one third of the 100,000 tweets concentrate on success stories. Additionally, within the topic corpus of slavery, tweets show a striking amount of formulations like "born into". By pattern matching the phrase, it was possible to extract tweets about individuals who were born into slavery and achieved the American dream by taking impressive job roles. In fact, the historical corpus yields no results when searching for the pattern "born into". This unveils that success stories shared on Twitter are a modern perspective people employ. In contrast, historical paragraphs comprise observations on what needs to change.

The topic of "war" mainly encompasses perspectives on the Tuskegee Airmen (WWII) and the Civil War. For the latter, tweet authors seem to hold a romanticized view of Black participation, writing about Black soldiers who fought for their liberty and "earned" equality. Little attention is given to the regiments' segregation (Fleche, 2014). Concerning World War II, the image of the "Good War" prevails, and the remembrance of the soldiers is marked by the display of national heroism (Däwes and Gessner, 2015). Following Däwis & Gessner, the U.S. national narrative of unity and moral self-confidence, however, is counterpointed by the experiences – both within and after the war – of African Americans (Däwes and Gessner, 2015).

Within the topic "work", many tweets were shared by corporate companies and contained declarations of intent for greater inclusion and diversity. A common phrase used was "diversity, equity and inclusion at the workplace". Usually, no concrete initiatives were named. For example, the term "scholarship" appeared only twice in the corpus and 76 tweets mentioned mentorship programs.

Twitter users conceive Martin Luther King Jr. and Malcolm X as bearers of hope. Many tweet authors drew connections to the new Critical Race Theory Laws in school. User @CosplayForJe, for instance, posted: "It's #BlackHistoryMonth and HB7/SB 148 are getting pushed through Florida. How are kids supposed to learn about the bravery of #Rosa Parks #RubyBridges and #MLK if it is illegal to teach about segregation because it makes white parents uncomfortable? #DontBanRosaParks" (2022).

Within the topic "Black health and wellness", the number of tweets that highlight the correlation between mental health and racism amounts to 19.8%. Studies support these correlations ((Bor et al., 2018; Halloran, 2019)). This points to concrete negative consequences which the minority group endures due to racism. One tweet, for example, highlights the psychological impacts of segregation (2022). Other users point to the importance of social justice to prevent suicide (@OvernightWalk, 2022).

### 5. Discussion

Based on the results, it is possible to present five discoveries: 1) #BlackHistoryMonth discourse is both conducted by African Americans to empower other African Americans and to educate whites. 2) Partly, African Americans follow

the white American doctrine of national heroism in remembering Black soldiers, not reminding of racially segregated troops and discrimination.<sup>5</sup> Future research should investigate if war is a unifying element between white Americans and African Americans. 3) #BlackHistoryMonth is not only anti-racist but feminist discourse. 4) Black memory culture on social media is tied to Black individuals rather than to historic events. 5) The education of African American history dislodges from education in schools and shifts to memorial sites. Future research could investigate if this is partly a consequence of oral history.

## 6. Literature

- Bor, J., Venkataramani, A. S., Williams, D. R., and Tsai, A. C. (2018). Police killings and their spillover effects on the mental health of black Americans: a population-based, quasi-experimental study. *Lancet (London, England)*, 392(10144):302–310.
- Däwes, B. and Gessner, I. (2015). Commemorating world war ii at 70: Ethnic and transnational perspectives. *American Studies Journal*, 59(1).
- Delgado, R., Stefancic, J., and Harris, A. P. (2017). *Critical race theory: An introduction*. Critical America. New York University Press, third edition edition.
- Fleche, A. M. (2014). This distracted and anarchical people: New answers for old questions about the civil war-era north. *Journal of Southern History*, 80(3):721–723.
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Halloran, M. J. (2019). African American health and post-traumatic slave syndrome: A terror management theory account. *Journal of Black Studies*, 50(1):45–65.
- Manovich, L. (2020). *Cultural Analytics*. MIT Press.

## Appendix

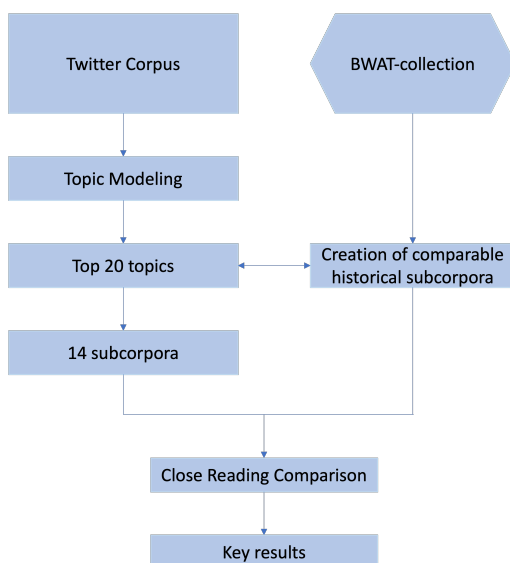


Figure 1: Corpus creation and methodology

<sup>5</sup>Däwis & Gessner state that national heroism and the notion of the "Good War" prevails.

# Zooming in on emergent norms: Preliminary findings from a cross-linguistic investigation of videoconferencing

Katharina Pabst<sup>1</sup>, Aida Alanzi<sup>2</sup>, Johanna Aminoff<sup>2</sup>, Raisa Tayib<sup>2</sup>, Derek Denis<sup>2</sup>

<sup>1</sup>Radboud University, <sup>2</sup>University of Toronto Mississauga

E-mail: [katharina.pabst@ru.nl](mailto:katharina.pabst@ru.nl), [aydah.alanzi@mail.utoronto.ca](mailto:aydah.alanzi@mail.utoronto.ca), [johanna.aminoff@mail.utoronto.ca](mailto:johanna.aminoff@mail.utoronto.ca), [raisa.tayib@mail.utoronto.ca](mailto:raisa.tayib@mail.utoronto.ca), [derek.denis@utoronto.ca](mailto:derek.denis@utoronto.ca)

## Abstract

In this pilot study, we examine conversational norms in the multimodal register of Zoom in several typologically distinct language varieties. Specifically, we combine methods from variationist and interactional linguistics to investigate backchanneling, i.e., the use of minimal responses like *uh-huh* and *mhm* to signal interlocutor engagement (Oreström, 1983; Eiswirth, 2020). We analyze 13 video-recorded conversations from six varieties (American English, Asante Twi, Finland Swedish, Ghanaian English, German German, and Gulf Arabic). Consistent with previous work, we find that backchanneling increases with turn length: turns are shorter in videoconferencing (mean length=25.9 words) than previously reported for face-to-face interaction (mean length in Eiswirth 2020 = 100 words). With respect to the frequency of backchanneling, a conditional inference tree reveals that only the conversation group is significant, while variety is not. This suggests that individual communicative styles and rapport may be more relevant than interlocutors' linguistic backgrounds and that norms are still emergent.

**Keywords:** videoconferencing, interactional norms, backchanneling, variationist sociolinguistics

## 1. Introduction

While videoconferencing has been around for years, the COVID-19 pandemic has made it an indispensable part of life for many people around the world. Previous work has found systematic differences between face-to-face and video interactions. For example, speakers have been shown to emphasize their articulatory movements as well as their vowel contrasts to facilitate their interlocutors' understanding (Bleaman et al., 2022). Despite the ubiquity of video-mediated conversations, there is little quantitative linguistic work on how speakers' conversational norms vary between face-to-face and video-mediated interactions (but see Boland et al., 2021), especially in conversation and low resource languages. This pilot study addresses this desideratum by examining conversational norms on Zoom in six typologically distinct language varieties.

## 2. Data and method

Our data was collected as part of an experiential learning course at the University of Toronto Mississauga during the academic year 2021-22. So far, we have transcribed 10-30 segments from 13 Zoom conversations (see Table 1).

Setting	No. of conversations	No. of turns
American English	2	110
Asante Twi	1	75
Finland Swedish	2	136
German German	2	161
Ghanaian English	2	315
Gulf Arabic	4	453
TOTAL	13	1250

Table 1: Overview of the data sample.

The data was collected by the first four authors. It consists of informal conversations between three to four family members, neighbors, or friends.

We focus on a feature that we expect to differ between face-to-face and video-mediated interactions, both due to latency and alternative ways of showing engagement: backchanneling, i.e., the use of minimal responses such as *uh-huh* and *mhm* to signal engagement (Oreström, 1983; Eiswirth, 2020). We follow Eiswirth in quantifying the frequency of backchannels by counting the number of responses to a turn divided by the number of words in said turn and multiplying it by 100. This normalized frequency is the dependent variable in our statistical analysis.

We coded the data for the turn taker, their age, gender, level of education, and comfort level with Zoom, as well as the conversation group and the language variety used most during the conversation.

## 3. Results

Since determining turn length is crucial for calculating the frequency of backchanneling, we begin with a distributional analysis of turn length. Results indicate that the mean turn length in our data (25.9 words) is much shorter than in previous studies of face-to-face interaction (100 words in Duncan & Fiske 1977 and Eiswirth 2020). As seen in Table 2, there are substantial differences between varieties, but this is to be expected given the limited number of conversations in the data set.

In line with Eiswirth (2020), we find that the number of responses increases with the number of words uttered by the turn-taker (see Figure 1). In other words, the longer someone speaks, the more we indicate that we are still paying attention. This is true for all six language varieties we examined.

Setting	Mean number of words in turn
American English	22.2
Asante Twi	41.6
Finland Swedish	37.5
German German	20.5
Ghanaian English	15.9
Gulf Arabic	29.5

Table 2: Distribution of number of words in turn by setting.

Note that we are exclusively counting verbal responses here since we are still working on operationalizing non-verbal responses. The use of reaction buttons was rare, with only a single occurrence in the Arabic data.

Finally, we conducted a conditional inference tree analysis in R (R Core Team, 2021) to determine the relative influence of the predictors (Tagliamonte & Baayen, 2012). Since the sample is not balanced for social factors, we only included conversation groups and variety as predictors. Results show that conversation group is significant, while variety is not. We conclude that individual preference and rapport among group members trump whatever cross-linguistic differences may exist in terms of backchanneling frequency, reflecting that conversational norms are still emergent.

#### 4. Conclusion

This is the first study to investigate cross-linguistic differences in backchanneling during video-mediated conversations. While we find differences in turn length between different language varieties, we do not find differences in the frequency of backchanneling.

Future work will focus on creating a stratified sample that will allow us to explore the influence of demographic factors such as age, gender, and education, as well as participants' comfort level with Zoom. It will also explore the types of responses individuals offer.

#### 5. Acknowledgements

We gratefully acknowledge financial support from the UTM Office of the Vice Principal Research.

#### 6. References

- Bleaman, I. L., Cugno, K. & Helms, A. (2022). Medium-shifting and intraspeaker variation in conversational interviews. *Language Variation and Change*, 34(), pp. 305--29.
- Boland, J. E., Fonseca, P. Mermelstein, I & Williamson, M. (2021). Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*. First view.
- Duncan Jr., S. & Fiske, D. W. (1977). *Face-to-face interaction: Research, methods, and theory*. Hillsdale, NJ: L. Erlbaum Associates.
- Eiswirth, M. E. (2020). Increasing interactional accountability in the quantitative analysis of sociolinguistic variation. *Journal of Pragmatics*, 170, pp. 172--188.
- Oreström, B. (1983). *Turn-taking in English conversation*. Gleerup: Lund.
- R Core Team (2021). R: A language and environment for statistical computing. R foundation for statistical computing. <https://www.R-project.org/>.
- Tagliamonte, S. A. & Baayen, H. (2012). Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change*, 24(2), pp. 1--7

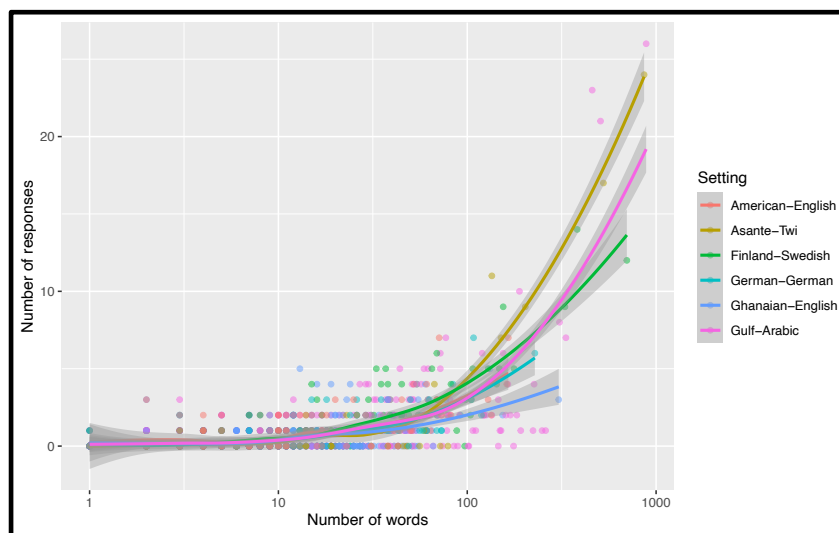


Figure 1: Number of interlocutor responses by number of words in turn

# ChrisTof: A Novel Corpus of Christian Online Forums

Sebastian Reimann, Lina Rodenhaut, Tatjana Scheffler, Frederik Elwert

Ruhr University Bochum  
firstname.lastname@ruhr-uni-bochum.de

## Abstract

Online forums represent a well-established and long-running form of computer mediated communication (CMC) that is however still relatively underexplored. Their significant online presence, thematic structure, and potential for community building renders them especially interesting for a wide range of research in the humanities and social sciences. For religious studies, online forums constitute a unique data source that provides access to lay people’s religious beliefs, reasoning and argumentation. Forum discussions also allow the near real-time observation of community construction and semantic change (Del Tredici et al., 2019).

We present a novel corpus of data from online forum posts in a custom-designed TEI XML format (Reimann et al., Submitted). In our XML structure, posts are grouped together according to the discussion threads in which they were produced, so that the original thread structure is preserved. The data includes complete archives from two Christian online forums in German (`jesus.de` and `mykath.de`) and two English subreddits related to Christianity (`r/TrueChristian` and `r/OpenChristian`). Table 1 provides detailed information on the size of the subcorpora.

Forum	Threads	Posts	Dates
Jesus.de	35,916	1,661,361	2007–2022
Mykath	13,577	1,157,653	2001–2022
r/OpenChristian	15,888	158,172	2010–2022
r/TrueChristian	55,986	1,084,214	2012–2022

Table 1: Overview of the forums in our corpus, the dataset size and the timespan covered.

We stored a comprehensive amount of post metadata. This includes post IDs, the timestamp of the post and reactions to posts such as Reddit upvotes and downvotes. Additionally, we preserved as much of the forum markup as possible, which means that formatting (e.g., boldface), as well as structural information such as quotations from scripture or other posts, are retained for future analyses. The user names have been automatically pseudonymized for privacy reasons.

Each post was automatically sentence segmented and tokenized using SoMaJo (Proisl and Uhrig, 2016) and all sentences and tokens were given unique IDs. The retained thread structure sets our corpus apart from previous corpora of forum-like CMC represented in TEI. We expect that this will facilitate comprehensive analyses of discourse relations between posts and the discourse structure of discussion threads, and thus provide new insights into the interactive nature of CMC.

The main purpose of our data is the annotation of metaphorically used words according to the Metaphor Identification Procedure VU Amsterdam (MIPVU) (Steen et al., 2010) and quantitative as well as qualitative analyses of religious metaphors in the forums. We will present the results of an initial annotation round, where we applied MIPVU to several complete threads from our corpus. Additionally, we use the corpus to train topic models for an analysis of the different Christian communities, of which we will also present early results.

Due to copyright and data protection issues, we will not upload our data publicly. However, we will make the data available to interested researchers – please contact the first author, stating the intended purposes.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958.

Del Tredici, M., Fernández, R., and Boleda, G. (2019). Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Proisl, T. and Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics (ACL).

Reimann, S., Rodenhaut, L., Elwert, F., and Scheffler, T. (Submitted). By a thread: Encoding online forum data in TEI. *Journal of the Text Encoding Initiative*.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., and Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.



# Building a Parallel Discourse-Annotated Multimedia Corpus

Hannah J. Seemann\*, Sara Shahmohammadi†, Tatjana Scheffler\*, Manfred Stede†

\*Ruhr-Universität Bochum

†Universität Potsdam

hannah.seemann@rub.de, sara.shahmohammadi@uni-potsdam.de, tatjana.scheffler@rub.de, stede@uni-potsdam.de

## Abstract

We present the construction process of a novel parallel discourse-annotated multimedia corpus in German, including data collection, pre-processing, paragraph-level alignment between documents, and annotation of discourse structure. The goal of the corpus is to enable the analysis of discourse-level variability across different media, in the context of the collaborative research cluster SFB 1287 “Limits of Variability in Language”.<sup>1</sup>

The presented corpus contains texts from two parallel computer-mediated media that present (roughly) the same information in two communicative situations: podcasts and blog posts. The data was collected from two domains, business and (popular) science, from sources where the original authors created their content both in a podcast format, as well as on a blog. After the podcasts have been automatically transcribed and manually checked, each podcast episode and its corresponding blog post was annotated manually for parallel segments. This paragraph-level alignment was carried out so that the discourse structure between texts covering the same information in two media can be compared regarding their linguistic features. The resulting corpus comprises 73 episodes in each medium (14,598 tokens in the blog posts, and 125,182 tokens in transcribed podcasts).

The blog posts and the corresponding parallel podcast segments have been annotated for discourse structure in two frameworks: Rhetorical Structure Theory (RST) and Questions Under Discussion (QUD). Both theories represent discourse structure as a tree. RST derives plausible global text structures by connecting discourse units using discourse relations (Mann and Thompson, 1988). It has been primarily designed for analyzing well-written text. On the other hand, the QUD model treats discourse as a series of implicit and explicit questions that are answered successively in dialogue (Ginzburg, 1996; Roberts, 2012).

We use this novel opportunity of a parallel RST and QUD annotated corpus to find similarities and differences between the discourse models (Shahmohammadi et al., 2023). Future tasks that are made possible with this corpus include the more detailed comparison of discourse structure in different media, as well as the analysis of linguistic features in different media, and different discourse structure frameworks.

The corpus will be released under a Creative Commons license at <https://osf.io/59acq/>.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287.

Ginzburg, J. (1996). Dynamics and the semantics of dialogue. *Logic, language and computation*, 1:221–237.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5, December.

Shahmohammadi, S., Seemann, H., Stede, M., and Scheffler, T. (2023). Encoding Discourse Structure: Comparison of RST and QUD. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 89–98.

---

<sup>1</sup><https://www.sfb1287.uni-potsdam.de/>

# Negotiating knowledge in cooperative learning scenarios: a multimodal approach to practices of computer-mediated and face-to-face communication in the university classroom

Sarah Steinsiek

University of Duisburg-Essen

**Keywords:** computer-mediated communication, pragmatics, cooperative learning, multimodality, data collection

The contemporary ‘digital condition’ (*Kultur der Digitalität*, Stalder, 2016) has given rise to innovative concepts of learning and teaching. Learning activities and interactions between teachers and learners as well as among learners do not exclusively take place in the physical classroom anymore but also – partly or completely – in the digital sphere using the potentials of computer-mediated communication (CMC). On my poster I will give an outline of my dissertation project, which I have been working on since January 2022 at the University of Duisburg-Essen, and discuss issues related to the collection of a multimodal corpus of interactions including (i) audio and video recordings of Zoom and face-to-face meetings (in peer-to-peer and plenary discussions with teachers), (ii) collaborative text annotation of digitized papers and discussion threads, (iii) cooperative text production with Etherpads on the learning platform *Moodle*, (iv) interviews with selected students, and (v) logfiles of private text messaging among students. These data are used for the investigation and modelling of digital and face-to-face practices of negotiating knowledge in a hybrid learning scenario in higher education (university seminar in linguistics). The learning scenario is designed to foster students’ competencies in comprehending researchers’ perspectives and approaches in linguistics papers and to improve their skills in discussing theoretical concepts derived from their readings. The scenario is characterized by the following features (examples with rough English translations illustrate the different stages of the setting):

- Student teams cooperatively elaborate on the theoretical frameworks and findings reported in papers and book chapters and discuss them on the basis of key questions provided by the teacher. They annotate and discuss texts using the *Moodle* activity type ‘Textlabor’ where they can comment on the text and also verbalize when they have difficulties understanding certain text passages.

**Example 1** (comment in a Textlabor discussion thread on a figure in a paper by Auer (2000), November 23, 2022):

Student 1: “Also ich kann das hier gar nicht gut nachvollziehen. Können wir da Freitag drüber reden?” *“I really can’t comprehend this [figure]. Can we talk about that on Friday?”*

- Based on their annotations and discussion threads, the teams talk about their findings and questions regarding text passages in Zoom meetings. They

use Etherpads to take notes and/or edit texts containing their results.

**Example 2** (Zoom meeting two days after student 1 posted his comment represented in Ex. 1, November 25, 2022):

Student 2 shares her screen that shows the Textlabor comment from Ex. 1: “so (---) das ist eh die frage (.) im zweiten text”. *this is uh the question (.) in the second text.*

Student 1 expresses his displeasure without repeating or rephrasing his comment: “ACH ja (.) hm ja (.) also DIE grafik [...] (das war) also WIRklich” [kollektives Lachen] *oh right (.) um yeah (.) well THAT figure [...] (that was) HONestly* [students laugh collectively]

[...]

Student 2: „ich kann das auch nicht verstehen (---) ehm (3.0) SCHWIERig (---) °h OH (.) d\_DAS ist ein [beispiel] von kombination also man s\_also der autor meinte in (dieser) äußerung [...]“ *“i don’t understand it either (---) um (3.0) DIFFicult (---) °h OH (.) th\_THAT is an [example] of combination so you s\_so in this statement the author meant [...]*

- In class they discuss their results and open questions face-to-face with the other teams and the teacher.
- The students use private messaging apps and/or other communication channels of their own choice to organize their team work.

**Example 3** (WhatsApp messages on November 25, 2022, 90 minutes prior to the Zoom meeting):

Student 3: „Meint ihr, wir sollten gemeinsam Absatz für Absatz im Meeting durchgehen [...]“ *Do you think we should discuss paragraph by paragraph in our meeting [...]*

Student 1: [...] „Vielleicht gucken wir uns erstmal an, was im Textlabor bearbeitet wurde und orientieren uns dann an den Aufgaben?“ *Maybe we should take a look at the comments in the Textlabor first and then focus on the assignments [provided by the teacher]?*

[...]

Student 1: „[...] finde besonders den Auer-Text echt schwierig...“ [...] *especially the text by Auer is really difficult...*

On my poster I will include sample data and give insights into my analyses. In my analyses I combine the perspectives and concepts of interactional linguistics (Imo/Lanwer, 2019), research on ‘digital practices’ (see Androutsopoulos, 2016; Beißwenger, 2016) and of negotiating knowledge in classroom discourse (see e.g. Morek/Heller/Quasthoff, 2017).

At the current state, the collected data set has the status of a “corpus in the wider sense” (sensu Beißwenger/Längen, 2022): a collection of audio and video files, stored logfiles and text documents as well as transcribed audio and video files for the purpose of linguistic and conversational analysis. On the poster I will present the procedure of data collection in three linguistics classes (April 2022 to July 2023) with a special focus on the handling of ethical and GDPR issues and the challenge to deal with the observer’s paradox, i.e. the challenge to design the observation process as unobtrusive as possible:

- Prior to the data collection, the students were informed about the project without expanding on the research questions in order to avoid priming effects.
- It was pointed out that participation in the data collection is voluntary and non-participation does not have any negative implications.
- The students gave informed consent by specifying which data types may or may not be collected (*Gestufte Einverständniserklärung*, Stukenbrock, 2022: 313).
- Data were collected in a “natural”, i.e. non-experimental setting by using unobtrusive recording devices and, whenever possible, in my absence (see Stukenbrock, 2022: 312).

A main motivation for my presentation at the conference is to get in touch with other researchers with experience in using state-of-the-art corpus technology and to discuss issues of representing and analyzing (multimodal) corpora with heterogeneous data types.

## References

- Androutsopoulos, J. (2016). Mediatisierte Praktiken: Zur Rekontextualisierung von Anschlusskommunikation in den Sozialen Medien. In A. Deppermann, H. Feilke & A. Linke (Eds.), *Sprachliche und kommunikative Praktiken* (pp. 337–367). Berlin, Boston: de Gruyter.
- Beißwenger, M. (2016). Praktiken in der internetbasierten Kommunikation. In A. Deppermann, H. Feilke & A. Linke (Eds.), *Sprachliche und kommunikative Praktiken* (pp. 279–310). Berlin, New York: de Gruyter.
- Beißwenger, M. and Längen, H. (2022). Korpora internetbasierter Kommunikation. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 431–448). Paderborn: Brill|Fink (= UTB 5711).
- Imo, W. and Lanwer, J. P. (2019). *Interaktionale Linguistik.*

*Eine Einführung.* Stuttgart: Metzler.

- Morek, M., Heller, V. and Quasthoff, U. (2017). Erklären und Argumentieren. Modellierungen und empirische Befunde zu Strukturen und Varianzen. In I. Meißner & E. L. Wyss (Eds.), *Begründen – Erklären – Argumentieren. Konzepte und Modellierungen in der Angewandten Linguistik* (pp. 11–45). Tübingen: Stauffenburg.
- Stalder, F. (2019), *Kultur der Digitalität.* Berlin: Suhrkamp.
- Stukenbrock, A. (2022). Audio- und Videographie. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 307–323). Paderborn: Brill|Fink (= UTB 5711).

# Linguistic features, device affordances, and contextual factors: A mixed-methods, two-corpora approach

Jenia Yudytska

University of Hamburg

E-mail: yevgeniya.yudytska@studium.uni-hamburg.de

## Abstract

In CMC research, the role of the particular technological device used to send messages is rarely taken into account; messages sent by computer and phone are implicitly treated as broadly similar (cf. Jucker & Dürscheid, 2013). In contrast, laypeople often believe their messages vary between device types, e.g., regarding message length, use of emoji, capitalisation, etc. The present study thus investigates the potential influence of device on such microlinguistic features. Rejecting any technological determinism, i.e., that computer-sent and phone-sent messages differ categorically, the study instead favours an affordance-based approach (cf. Hutchby, 2001). Device properties (e.g., keyboard type, autocorrect) afford the use of various linguistic features more easily or less, which may lead to linguistic variation in CMC messages. However, affordances are one influence among many, as contextual factors like synchronicity can also play a role in linguistic variation, as can individual user style.

Drawing inspiration from computational sociolinguistics (cf. Nguyen et al., 2015), the empirical study uses quantitative and qualitative methods to investigate both device affordances and their interactions with other factors. To explore both aspects, a two-strand approach was designed, each relying on its own type of corpus. Section (1) focuses solely on the influence of device affordances, and uses a large-scale corpus, so as to explore general trends found across device types. Section (2) focuses on interactions between affordances and contextual factors, and thus uses a smaller-scale corpus with richer information about both message context and the users. This two-strand, two-corpora approach allows for a richer understanding of both the possibilities and limits to the influence of device affordances.

Thus, for Section (1), a large-scale corpus of a million anonymous Twitter messages was collected, with the only metadata being device type. Five categories of microlinguistic features were examined: length, acronyms and abbreviations, emoji and emoticons, punctuation, and non-standard orthography. Quantitative analysis found weak but relatively consistent differences across them, for example, a higher frequency of emoji on the phone. For Section (2), a small-scale corpus of 50,000 messages from the platforms Twitter and Discord was collected from the same eleven participants. The two platforms differ in regard to contextual factors like synchronicity and audience size, and thus it is possible to compare how the influence of device affordances differs across them. Quantitative analysis found variation for both device type and platform, while a fine-grained qualitative analysis showed that users differed also in the extent to which they adhered to or circumvented device affordances. For example, mirroring the findings in the large-scale corpus, across both Twitter and Discord phone-based emoji frequency was overall found to be higher, while contrary results were also found for individual users due to their personal device and platform-related habits. The study thus illustrates both the interaction of device, contextual factors, and style, as well as the usefulness of complementary large-scale and smaller-scale corpus analysis.

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, 35(2), pp. 441-456.

Jucker, A.H. & Dürscheid, C. (2013). The linguistics of keyboard-to-screen communication. A new terminological framework. *Linguistik Online*, 56(6), pp. 39-64.

Nguyen, D., Doğruöz, A.S., Rosé C.P. & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3), pp. 537-593.

**Keywords:** device, affordance, corpus scale, microlinguistic features

# Deontic Authority in Computer-mediated Communication Between University Teachers and Students: A Comparative Study of German and Chinese

ZANG Yinglei

University of Duisburg-Essen

**Keywords:** computer-mediated communication, deontic authority, WeChat, email, interactional linguistics, conversation analysis

My poster presentation presents work in progress. I will shed light on the central question of my doctoral project which I have been conducting since October 2022 at the University of Duisburg-Essen. It is focused on deontic authority between teachers and students in German as well as Chinese universities respectively, i.e., how it is constructed, demonstrated and negotiated through different practices. The data are drawn on the one hand from WeChat-interactions between Chinese university teachers and students, and on the other hand from email correspondences between German university teachers and students. The study takes a comparative perspective and tries to provide insights into the computer-mediated communication in institutional contexts for two different languages and cultures.

Deontic authority which makes up the main theoretical framework of my study is about getting the world to match the words, i.e., determining what ought-to-be (Stevanovic 2013). It is no exaggeration to say that most of our actions have something to do with deontic authority. In institutional communication, for instance that between teachers and students, interlocutors have to deal with (i) the relation of *deontic status* which are determined, e.g., by their institutional roles, and (ii) *deontic stance*, i.e., locally and interactionally positioned expressions regarding deontic rights (Frick/Palola 2022). It is the dynamics of deontic authority in authentic interactions that makes this topic interesting.

I examine my data within the methodological framework provided by (i) interactional linguistics (Couper-Kuhlen/Selting 2018) and conversation analysis and (ii) research on the affordances and practices of computer-mediated communication (e.g., Beißwenger 2016). I focus on types of interactions which are typical for digital, institutional communication between teachers and students – making appointments, discussing term papers or bachelor and master theses, answering questions related to lectures and seminars etc. – and try to explain how deontic authority comes into play.

The following example from my WeChat-data shows some interesting practices of the teacher (Q) and the student (Nan). Q is the supervisor of Nan's master thesis. Firstly, this student is observed to perform transformed turns and adjusts her actions (e.g., making extreme case formulations and promises from post 14 to 19) successively in front of the perceived deontic authority of the teacher. With an institutional task in her mind, she infers the intention of the teacher and gives him deontic authority. By performing resistance, i.e., refusing Q's advice, she however doesn't totally relinquish her deontic right. Secondly, the teacher

responds to turns of the student only selectively (an eventually more obvious proof is abandoning the dialog abruptly), and carries the sequential development forward for his own intention, which, I argue, demonstrates more deontic stance by controlling the sequential organization. At the same time, the teacher refers to and relies on his epistemic primacy (post 4-8, 15) as a vehicle and veil in order to impose his deontic authority, i.e., he demonstrates his more know-that in terms of writing a thesis in order to persuade the student to accept his advice.

## Example:

Q 14:09

论文的整体框架出来了没有

Have you already finished the main structure of your paper?

WeChat #1, (03.02.2022, 14:09 PM)

Q 14:09

什么时候给我交一个整体的论文

When will you hand over a complete paper

WeChat #2, (03.02.2022, 14:09 PM)

Nan 14:16

老师，二月 10 号发给您整体的论文~

Teacher, I will send you the complete paper on 10th February~

WeChat #3, (03.02.2022, 14:16 PM)

Q 14:16

时间来得及吗

Do you have enough time

WeChat #4, (03.02.2022, 14:16 PM)

Q 14:16

我需要时间看

I need time to read it

WeChat #5, (03.02.2022, 14:16 PM)

Q 14:16

你还要修改

And you still have to revise it

WeChat #6, (03.02.2022, 14:16 PM)

Q 14:16

我还要再看

And I have to read it again



WeChat #7, (03.02.2022, 14:16 PM)

Q 14:16

够吗

Is it enough  
WeChat #8, (03.02.2022, 14:16 PM)



Q 14:16  
建议你延期毕业  
I advise you to postpone your graduation  
WeChat #9, (03.02.2022, 14:16 PM)

Nan 14:16  
 这周日给您  
 I will give you this weekend  
WeChat #10, (03.02.2022, 14:16 PM)

Q 14:16  
我不是第一次提醒你  
This is not the first time I remind you  
WeChat #11, (03.02.2022, 14:16 PM)

Q 14:16  
延期到今年年底毕业  
Postpone your graduation till the end of this year  
WeChat #12, (03.02.2022, 14:16 PM)

Q 14:16  
这样你还有大半年的时间好好写论文  
So you still have the most year to focus on your paper  
WeChat #13, (03.02.2022, 14:16 PM)


Nan 14:16  
老师, 我的论文大致已经写完了 目前在按您第一次提的意见修改  
Teacher, I have already finished the most part of my paper  
 Currently I am revising it according to your first advice  
WeChat #14, (03.02.2022, 14:16 PM)


Q 14:16  
你时间不够  
Your time is not enough  
WeChat #15, (03.02.2022, 14:16 PM)

Q 14:16  
利用假期跟家人充分沟通  
Make use your vacation to communicate with your family  
WeChat #16, (03.02.2022, 14:16 PM)

Q 14:16  
准备延期毕业  
Prepare to postpone your graduation  
WeChat #17, (03.02.2022, 14:16 PM)

Nan 14:16  
老师, 我这段时间一定尽全力修改  
Teacher, I promise to try my best to revise it recently  
WeChat #18, (03.02.2022, 14:16 PM)

Nan 14:16  
 所有的论文真的已经写完了  
I have really finished my paper

 WeChat #19, (03.02.2022, 14:16 PM)

This study is intended to be qualitative. I have already obtained approximately 1200 sequences of WeChat interactions between teachers and students based on voluntary donation and under consideration of privacy protection. As a next step I am planning to collect email correspondences between German teachers and their students. For this purpose, I have created a data-collection plan which also pays attention to privacy protection of the involved interlocutors. At the same time, I am working at analyzing the WeChat data and try to develop further my categories of analysis.

At the moment, the WeChat sequences are represented as more or less 'raw' data (1: screenshots of the original sequences as they have been displayed on the students' smart devices, 2: the written and graphic content of the sequences stored in text documents, 3: a prose representation of the metadata relevant for my analysis). It is thus not (yet) a "corpus in the narrower sense" (sensu Beißwenger/Lüngen, 2020). Through presenting my project at the CMCCORPORA conference, I am interested to discuss and learn how other researchers represent and handle their CMC data for the purposes of documentation and analysis, and learn about tools that may be useful for storing and annotating my data (two languages, two different types of CMC).

## References

- Beißwenger, M. (2016): Praktiken in der internetbasierten Kommunikation. In In A. Deppermann, H. Feilke & A. Linke (eds.), *Sprachliche und kommunikative Praktiken* (pp. 279-310). Berlin, New York: de Gruyter.
- Beißwenger, M., & Lüngen, H. (2020). CMC-core: a schema for the representation of CMC corpora in TEILE CMC-core : un schéma de représentation des corpus de la CMR en TEI. *Corpus 20*.
- Couper-Kuhlen, E., & Selting, M. (2017). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press.
- Frick, M., & Palola, E. (2022). Deontic Autonomy in Family Interaction: Directive Actions and the Multimodal Organization of Going to the Bathroom. *Social Interaction. Video-Based Studies of Human Sociality*, 5(1).
- Stevanovic, M. (2013). Deontic rights in interaction: A conversation analytic study on authority and cooperation.

# IDA - Incel Data Archive: a multimodal comparable corpus for exploring extremist dynamics in online interaction

Selenia Anastasi, Tim Fischer, Florian Schneider, Chris Biemann

University of Genoa, Language Technology Group (Hamburg University)

selenia.anastasi@edu.unige.it, tim.fischer@uni-hamburg.de,

florian.schneider-1@uni-hamburg.de, biemann@informatik.uni-hamburg.de

## Abstract

Extremist online communities are rapidly growing locally, posing potential threats to European and non-European countries. To gain insight into the dynamics of interaction within these web-based extremist groups, we present IDA, the Incel Data Archive. IDA is a multilingual and multimodal corpus compiled from Incel forums in both Italian and English languages. With its collection of forums, blogs, and websites, the Incelosphere serves as an ideal case study for examining interaction dynamics within extremist online communities from a cross-cultural perspective. Therefore, our work makes a twofold contribution: firstly, it provides an original cross-cultural perspective on the Incel phenomenon, and secondly, it extensively discusses the challenges and opportunities encountered when constructing a multimodal and multilingual corpus from discussion forums. To achieve this, we employ a mixed-method approach to Computer Mediated Communication. In order to shed light on important differences between the two communities, we conducted an exploratory analysis based on a novel topic modeling technique based on Transformer architectures. This approach allowed us to delve into the themes present in the two corpora. The results of our thematic exploration demonstrate not only variations in the discussion topic favoured by each community but also differences in the targets of their hateful content.

**Keywords:** CMC corpora, Incels, Online Extremism, Multimodality, Multilingual Corpora

## 1. Introduction

After spreading within Reddit, Incels communities gradually aggregated outside mainstream social networks, creating the formation of an independent insular cluster of local-based communities. Recently, several studies (Gillett and Suzor, 2022; Trujillo and Cresci, 2022) supported the hypothesis that moderation and quarantine practices adopted by mainstream platforms, may foster the growth of hateful insular peripheral communities akin to echo chambers. The creation of the dataset presented in this work was motivated by the need to draw upon spontaneous examples of Computer Mediated Discourse that exhibited similar content from various perspectives, framing this phenomenon at a local level. Moreover, even though the discourse of the Incelosphere is characterised by its hateful, misogynistic and anti-feminist contents, we argue that a corpus consisting of data from the Incelosphere may be useful in answering broader research questions that address the general understanding of the digital ecosystem in which extremist users interact. Our contribution is thus twofold: first, we intend to contribute to a deepen understanding of Incel communities from a cross-cultural perspective. Secondly, as datasets from these sources have not yet been made openly available for academic purposes, this study aims to fill this gap by addressing some of the challenges that accompany the construction of a multimodal and bilingual corpus in Italian and English. From a methodological perspective, we intend to offer our perspective and solutions to aid in the construction of a corpus intended to study of the forums-based communities by drawing on the Computer Mediated Discourse research field. We believe that this perspective is particularly relevant because it considers both the level of user interaction, the *affordances* of the discussion fora, as well as how the community and its sociocultural context influence each other.

## 2. The Incelosphere so far

Anglophone Incel communities have been studied from a wide variety of perspectives, ranging from psychology to discourse analysis. Many of these studies were focused on Reddit groups (r/ForeverAlone and r/Incels subreddits), which are archived in datasets and can be used as corpora. In Sociology, studies focused on the discursive practices, rhetoric and argumentation style, symbolism, and sexual imagery of Incel communities (Massanari, 2017; Waśniewska, 2020; Tranchese and Sugiura, 2021; Aiston, 2023; Prazmo, 2022), male and female identity construction (Ging, 2019; Chang, 2022; Thorburn et al., 2023), target of the hateful content (Pelzer et al., 2021), thematic and rhetorical connections to far-right oriented groups (Nagle, 2017), anti-feminism, values, normative orders, and group beliefs (Sugiura, 2021; O Malley et al., 2022; Heritage and Koller, 2020). Empirical analyses and terrorism studies have sought to trace, also through dynamic cross-platform approaches, the development of violent extremism in the main Anglophone communities (Ribeiro et al., 2021; Baele et al., 2023), as well as their misogynistic stances (Jaki et al., 2019; Farrell et al., 2019). For Baele and colleagues, “incel discourse demonstrates typical markers of extremist language” that is “an essentialist categorisation of society into sharply delineated *in-groups* and *out-groups* where the latter are linguistically dehumanised, and a conspiratorial narrative presenting the in-group as the victim of an all-powerful structure of oppression” (Baele et al., 2023). Moving from the latter consideration, our study aims to frame the studied communities as insular clusters that have spontaneously arisen among individuals who, while sharing the same ideology and similar ways of articulating it, may display different levels of extremism.

### 3. Online discussion forums and Computer Mediated Discourse

The Computer Mediated Discourse (CMD) approach traditionally concerns the study of discourse in interactions where communication occurs through computers or mobile devices (Herring and Androutsopoulos, 2015). While much of the research has focused on texts, recent attempts have been made to incorporate graphic, audio, and video elements, as well as stylistic and stylometric elements at the level below the utterance. Additionally, the CMD approach distinguishes itself from other discourse approaches by considering the importance of platform-specific affordances and how they shape interaction, an aspect we aimed to preserve in our corpus. Indeed, forums-mediated conversations are not simply digitised conversations, but rather a distinct type of interaction with their own conditions of production and interpretation. Nonsynchronous digital interaction promotes the presence of complex sequential organisations, with connections to previous shifts and the management of multiple lines of interaction in parallel. This necessitates participants to develop new methods for indexing sequential connections, self-introduction, greetings, and attention calls. Taking these aspects into account, the primary objective of this study is to illuminate the language and dynamics of interaction within Incel extremist communities, bridging the gap in resources that are openly available and can be used to examine this phenomenon from a cross-cultural perspective. The specificity of the corpus should not come as a surprise. With the spreading of new social networks sites such as TikTok, and the growing interest in particular phenomena related to digital communication, we have witnessed the development of several corpora tailored for specific purposes in recent years. As generic linguistic corpora such as the WaCky corpus (Baroni et al., 2009) do not enable researchers to delve into specific topics, more recent studies have focused on creating corpora from online content related to specific themes, such as anti-vaccine movements, fake news, and conspiracy theories (Miani et al., 2021). In the next paragraph, we offer a more detailed description of the corpus design, data collection criteria and annotation processes.

## 4. Corpus constructions

### 4.1. Collection criteria

To ensure that the samples between the two language-based macro communities are homogeneous, both in terms of characteristics of the medium and local situational factors, we took as a starting point the affordances offered by all the forums examined and on the similarities between the discussion topics, and the user’s identity claim as Incel. The selection of the forum was carried out with qualitative methods, including expert-domain close reading, for the purpose of analyzing the similarities between the two communities and identity claims their user-base. Thus, we selected only those forums that showed the greatest similarity in structure, affordances and purposes. According to relevant literature (Lilly, 2016), we considered the different communities present within the Manosphere (PUA, MG-TOW, MRA, etc.), and assessed the different user-reported

positioning and framing with respect to issues of masculinity and anti-feminism. Having to place each of these groups on an ideal continuum ranging from “not at all toxic” to “very toxic”, according to (Farrell et al., 2019) and (Ribeiro et al., 2021), Incels Anglophone communities shows a sharp rise in the mean toxicity score compared to PUAs and MRAs. For this reason, we believe that researching the Incel forums may be a worthy case study for a cross-cultural investigation on the rise of new online extremism. After the selection of the forums, we define relevant sections and threads according to our purposes. We chose to select and collect only specific freely accessible threads that did not require any formal subscription to the two forums. This was due to two main reasons: first, the ethical one - avoiding to violate the privacy policies of the platform; second, to reduce the risk for the researchers to be subjected to potential violence and other forms of retribution.

### 4.2. Dataset collection

We collected the data and processed the dataset using well-established methods (Holtz et al., 2012). Both forums are structured hierarchically in sections, threads, and posts. Every section can contain a varied number of threads of different lengths that relate to roughly one topic, and consisting of asynchronous conversation flows in which can involve various users.

For the composition and collection of the dataset, we implemented multiple crawlers, one for each forum, in order to systematically download threads and posts of the sections of our interest. Given the URLs to the sections of interest (e.g. Introduction, Inceldom Discussion, Off-Topic), the crawler performs the following steps:

1. Visit each section. Collect URLs to all threads of that section
2. Visit every thread. Extract metadata of the thread and collect URLs to all of its posts.
3. Visit every post. Extract metadata of the post and its content. If available, download linked materials such as image, video or audio data.

To be more specific, the crawlers extract `title`, `permalink`, `date` and `id` for threads, and `speaker`, `content`, `permalink`, `date`, `id`, `thread id`, `title`, `image urls` and `reply to` for posts. With this procedure, the created dataset captures the hierarchical structure of the forums of sections, threads and posts as well as the conversational flow of the threads and posts of referring, citing and replying to other users. The final dataset comes in various formats, including CSV, JSON, HTML and PDF. The two latter formats are a reconstruction of the original forum format and well suited for annotation tasks. The crawlers are implemented with Scrapy<sup>1</sup>, a Python framework for extracting data from websites. To navigate the forums and to extract metadata, content or linked materials, it is required to specify CSS and XPath selectors that point directly to the desired content. These identifiers are

---

<sup>1</sup><https://scrapy.org/>



specific to every website and forum, which makes the development of such crawlers a careful and time-consuming endeavour.

The English dataset is about 10 times the size of the Italian. Further statistics of the crawled datasets can be seen in Table 1.

A phase of post-processing has been devoted to managing external links and videos embedded in user posts. These information have been automatically replaced by appropriate labels. A second challenge involved the anonymization of user names in threads and posts to ensure data privacy.

	English	Italian
Number of threads	369.174	35.624
Number of posts	7.359.727	740.278
Average posts per thread	20	21
Average post lengths (in chars)	161,45	281,90
Number of images (total)	425.259	20.183
Number of images (unique)	72.22%	93.69%

Table 1: Main statistics of the English and Italian datasets.

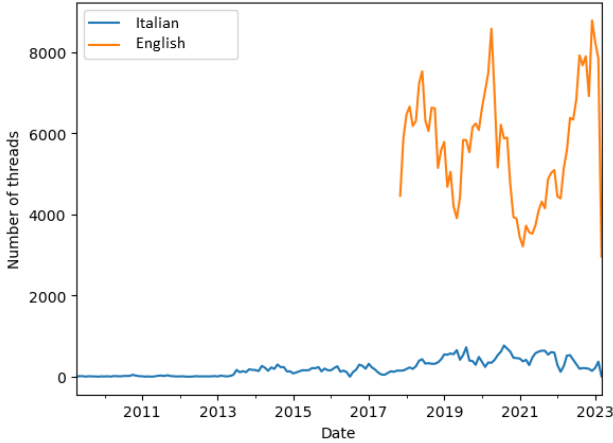


Figure 1: Number of new threads over time for English and Italian forum.

## 5. Corpus exploration

### 5.1. Methodology

The preliminary exploration and comparison of the dataset contents was executed in two phases. The first phase involved the use of topic modeling for extracting topics through an unsupervised approach. The second phase employed a Corpus Assisted Discourse Analysis, following the steps outlined in Baker’s proposed model (Baker, 2006): Description, Interpretation, Explanation, and Evaluation. The Explanation phase incorporated cross-referencing with the data from other sources such as newspapers and findings from previous research, particularly concerning the Anglophone community. Since the Italian community is studied to a lesser extent, the collection of all the relevant information has been carried out during the course of the last year by periodically accessing the forum and investigating the

practices of the community in close-reading. This qualitative analysis focused on a thorough reading and revolved around identifying key actors and topics being discussed within the forum, with a specific emphasis on aspects potentially influenced by sociodemographic variables such as entertainment, sexuality, and employment. The interpretation phase for the topics that were generated by the topic modeling was supported by exploring the meanings of the keywords contained within each topic list using the concordance tool in Sketchengine, on a subcorpus of both datasets consisting of approximately 3 million words each. Alongside the description of the dataset, our work, we showcases the potential for future research based on the Incelosphere corpus.

For the first phase, regarding the topic modeling, we randomly sampled 10% of the threads from the English forum (36.917), balancing the smallest Italian corpus (35.624). Although criticised (Brookes and McEnery, 2019), in Social Sciences and Digital Humanities, a widely used technique for exploring large unlabeled corpora is topic modeling. In our analysis, we replaced the classic approach based on bag-of-words representations and LDA (Blei et al., 2003) with a new approach based on transformer architectures (Vaswani et al., 2017), which allows for the extraction of words not only in relation to their distribution throughout the documents, but also in relation to their context of occurrence. Topic modeling based on BERT embeddings (Grootendorst, 2022) proved to be reliable for its high versatility and stability across domains, the possibility to perform analysis on multilingual data, and the ability to automatically extract the appropriate number of topics based on the sample size (Egger and Yu, 2022). This allowed us to obtain highly disambiguated word lists and minimised output manipulation. We used the Sentence Transformer (Reimers and Gurevych, 2019) model ALL-MPNET-BASE-v2<sup>2</sup> to compute vector representations of the threads, as it yielded the best clustering in our experiments. Topic modeling allowed us to obtain some preliminary insights on the topic trends, both synchronically and diachronically (see Fig.2).

### 5.2. Cross-media and cross-cultural analysis

The first step in analysing the contents of both datasets has been to visualise the flow of new messages over time (see Fig. 1). We found that the flow of new threads differs significantly between the two forums. The Italian forum displays a relatively stable pattern of new messages per day, whereas the English forum exhibits distinct peaks in 2018, 2020, and 2023, as well as a notable decrease in 2019 and 2021. The reason behind this trend remains to be accurately determined; however, from a cross-media perspective, we notice a correspondence between the decrease of messages and how the media gives attention to this community cyclically, mostly when there are crime events associated to it. Notably, there have been 50 documented cases of incel violence since 2014, including the murder of five people by Jake Davison in Plymouth in August 2021 and Gabrielle Friel’s weapons stockpiling in 2019 for a terror-

<sup>2</sup><https://www.sbert.net/docs/pretrainedmodels.html>

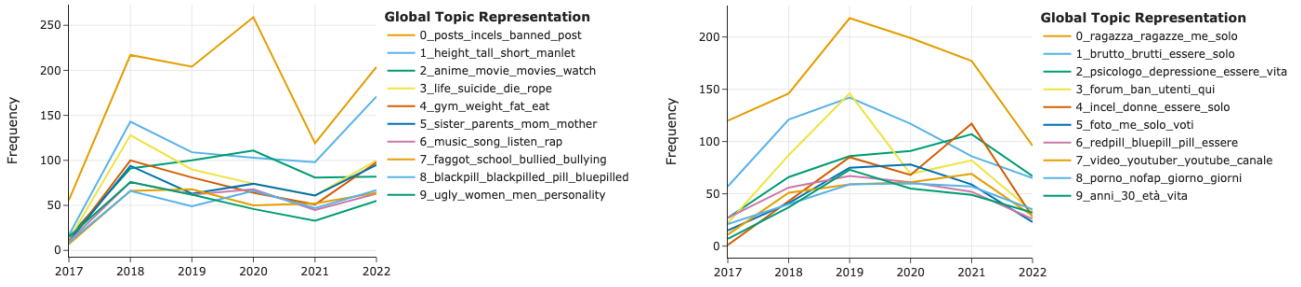


Figure 2: Dynamic topic modelling comparison of the English (Top) and Italian (Bottom) forum. Best viewed digitally with colour and zoom.

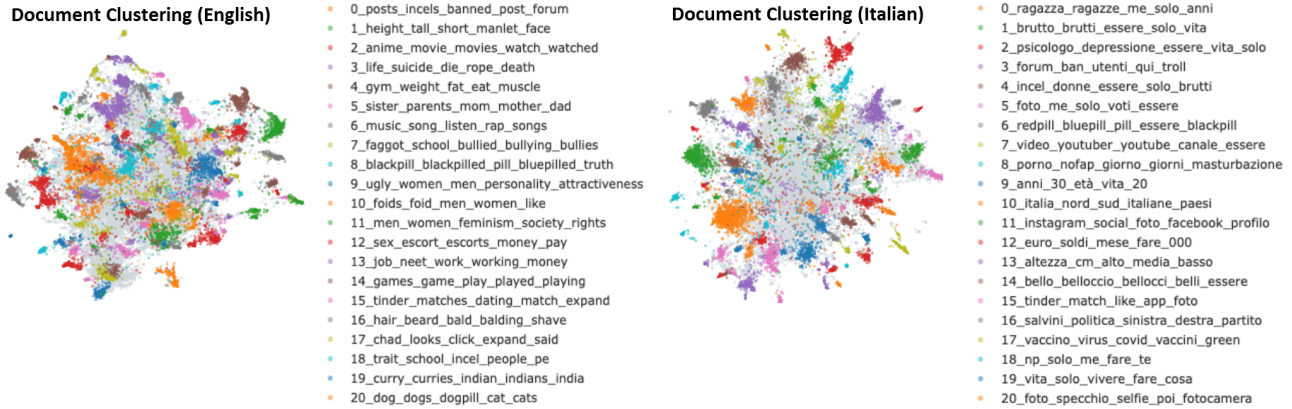


Figure 3: Static topic modelling comparison of the English (Left) and Italian (Right) forum. Best viewed digitally with colour and zoom.

ist attack in Scotland<sup>3</sup>. Furthermore, according to previous literature (Baele et al., 2023), we found that both, in Anglophone and Italian groups, there was an increase in the flow of messages in correspondence with the pandemic and post-pandemic years. The relationship between media attention and the growth of online extremist communities has already been observed elsewhere (Sugiura, 2021), but the correlation can be better supported by further and more in-depth analyses.

Results from the dynamic topic modeling (see Fig. 2) contain some clues about their differences: in the Anglophone community, main topics are mostly related to the user’s activity (both internal and external); mainstream entertainment such as movies and music; aesthetics issues, particularly height and weight; reference to bullying, suicide, death and rapes, as well as reference to feminine family components (sisters and mothers) and women in general. Interestingly, in the Italian forum the 2017 marks a turning point in user interest. Prior to 2017, the most frequent theme appears to have been the identity traits that characterize the user base and gives the group its name (*being ugly*), while after 2017, discussions are directed towards girls and women (or in slang, “non-persons”), anti-feminism, loneliness, and mental health. Topics concerning the internal life of the forum are still present, such as “ban” and “users”, indicating possible concerns on boundary maintenance. The static clustering of the two datasets (see Fig. 3) shows

the prevalence of social and affective concerns in the Anglophone group, such as unemployment, family care, sexuality and prostitution. This also emerges from the the Italian forum, where mainstream platforms such as YouTube and Instagram appear to play a prominent role. Moreover, from the Italian data in particular, aesthetic evaluation seems to be the prevailing community practice. This is not surprising, confirming the cornerstones of Incel’s theories in the so-called “LMD theory”, acronym for Look, Money, and Status, according to which both men and women are considered, and consider themselves, “as sexual objects to be evaluated and inserted in a hierarchical order characterised mainly by aesthetics” and economical status (Dordoni and Magaraggia, 2021). The widespread reference to ethnic categorisations such as “white”, “black”, “Indians” (or the incel slang variant, “curry/curries”), “Jews”, and “Asians”, along with keywords such as “race” and “ethnicity” in the Anglophone group is worthy of further investigation. In contrast, this pattern does not seem to emerge in the Italian forum, where stereotypes address the difference between men and women of southern and northern Italy. This aspect marks a point of continuity between the two communities, and can provide important clues for future analyses aimed at revealing mutual influence between the cultural ground of the user base and their radical instances.

## 6. Future Works

For further analyses, we plan to apply computational techniques such as network analyses and in-depth hate speech detection. These analyses can provide additional insights

<sup>3</sup><https://www.theguardian.com/lifeandstyle/2021/mar/03/incel-movement-terror-threat-canada>

both on the linguistic level (is there any difference in the way hate is expressed between the two linguistic communities? Who are the target groups?) and on the level of social structures and internal hierarchies. Finally, we plan to annotate the textual data in order to reveal interactional and rhetorical patterns, while the images will be annotated to provide a new benchmark for misogyny recognition.

## 7. References

- Aiston, J. (2023). *Argumentation strategies in an online male separatist community*. Ph.D. thesis, Lancaster University.
- Baele, S., Brace, L., and Ging, D. (2023). A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and Political Violence*, pages 1–24.
- Baker, P. (2006). *Using corpora in discourse analysis*. Bloomsbury Publishing.
- Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Brookes, G. and McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.
- Chang, W. (2022). The monstrous-feminine in the incel imagination: investigating the representation of women as âfemoidsâ on r/braincels. *Feminist Media Studies*, 22(2):254–270.
- Dordoni, A. and Magaraggia, S. (2021). Modelli di mascolinità nei gruppi online incel e red pill: Narrazione vittimistica di sé, deumanizzazione e violenza contro le donne. *AG About Gender-International Journal of Gender Studies*, 10(19).
- Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7.
- Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96.
- Gillett, R. and Suzor, N. (2022). Incels on reddit: A study in social norms and decentralised moderation. *First Monday*.
- Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and masculinities*, 22(4):638–657.
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Heritage, F. and Koller, V. (2020). Incels, in-groups, and ideologies: The representation of gendered social actors in a sexuality-based online community. *Journal of Language and Sexuality*, 9(2):152–178.
- Herring, S. C. and Androutsopoulos, J. (2015). Computer-mediated discourse 2.0. *The handbook of discourse analysis*, pages 127–151.
- Holtz, P., Kronberger, N., and Wagner, W. (2012). Analyzing internet forums. *Journal of Media Psychology*.
- Jaki, S., De Smedt, T., Gwózdź, M., Panchal, R., Rossa, A., and De Pauw, G. (2019). Online hatred of women in the incels.me forum. *Journal of Language Aggression and Conflict*, 7(2):240–268.
- Lilly, M. (2016). *‘The World is Not a Safe Place for Men’: The Representational Politics Of The Manosphere*. Ph.D. thesis, Université d’Ottawa/University of Ottawa.
- Massanari, A. (2017). # gamergate and the fapping: How reddit’s algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3):329–346.
- Miani, A., Hills, T., and Bangerter, A. (2021). Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24.
- Nagle, A. (2017). *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.
- O Malley, R. L., Holt, K., and Holt, T. J. (2022). An exploration of the involuntary celibate (incel) subculture online. *Journal of interpersonal violence*, 37(7-8):NP4981–NP5008.
- Pelzer, B., Kaati, L., Cohen, K., and Fernquist, J. (2021). Toxic language in online incel communities. *SN Social Sciences*, 1:1–22.
- Pražmo, E. (2022). In dialogue with non-humans or how women are silenced in incelsâ discourse. *Language and Dialogue*, 12(3):383–406.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., and Zannettou, S. (2021). The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207.
- Sugiura, L. (2021). *The incel rebellion: The rise of the manosphere and the virtual war against women*. Emerald Group Publishing.
- Thorburn, J., Powell, A., and Chambers, P. (2023). A world alone: Masculinities, humiliation and aggrieved entitlement on an incel forum. *The British Journal of Criminology*, 63(1):238–254.
- Tranchese, A. and Sugiura, L. (2021). âi donât hate all women, just those stuck-up bitchesâ: How incels and mainstream pornography speak the same extreme language of misogyny. *Violence against women*, 27(14):2709–2734.
- Trujillo, A. and Cresci, S. (2022). Make reddit great again: assessing community effects of moderation interventions on r/the\_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Waśniewska, M. (2020). The red pill, unicorns and white

knights: Cultural symbolism and conceptual metaphor in the slang of online incel communities. *Cultural conceptualizations in language and communication*, pages 65–82.

# The Reply Function in WhatsApp Chat Communication

Tianyi Bai

University of Mannheim, Germany

E-mail: tianyi.bai@students.uni-mannheim.de

## Abstract

This paper is based on empirical observations and focuses on the reply function in WhatsApp-Chats communication. Although this function has been available for several years, it has received little attention in academic research. Through the corpus analysis of both the collected data in the Mobile Communication Database 2 (MoCoDa2) and self-collected chat data from WhatsApp, this study identified various functions of the reply function in different chat contexts. In one-on-one chats, the reply function can serve diverse functions, including thematic reference, forming pair sequences, and improving comprehensibility. In contrast, in group chats, it can be used to address one or more participants, continue a previous conversation, or clarify misunderstandings. Structurally, while the quotation's placement can be influenced by various factors, messages that have been sent to the chat are quoted sequentially. This paper summarizes the functions and structures of using the reply function in individual and group chats, respectively, and contributes to research on internet-based and internet-supported chat communication.

**Keywords:** reply function; chat communication; MoCoDa2; WhatsApp

## 1. Introduction

The reply function is used to respond to certain messages in one-on-one or group chats (source: <https://faq.whatsapp.com>). In WhatsApp, users can swipe the message to the right and type their reply. Then, the quoted message and the reply can be sent to the recipient or group. Alternatively, users can access the reply function by pressing and holding the quoted message and clicking on the "Reply" button. This feature is also available in other messaging apps such as Telegram and Discord.

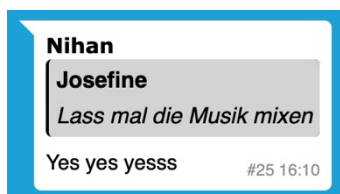


Figure 1: Screenshot of the reply function (MoCoDa2: #cKzoG)

As group chats involve multiple participants, it can be difficult for the specific participants to determine who is being addressed. Additionally, individuals often encounter the issue of information overload, where excessive number of messages are displayed on their screens, many of which may not be pertinent or necessitate a response. However, even in one-on-one chats with only one addressee, responders still utilize the reply function. This suggests that the reply function is not only used for addressing individuals correctly but also for providing feedback on specific content. These observations lead to the following research questions:

1. When do responders use the reply function? What is the relationship between the post quoted and the reply?
2. What are the structures of the chats using the reply function in one-on-one chats and group chats?

## 2. Literature Review and Methodology

The nomenclature used by researchers into this subject has

been extremely varied. It includes "computer-mediated communication" (Herring, 1996), "computer-based communication" (Beck, 2006, and Crystal, 2011), "internet-mediated communication" (Yus, 2011) and "interactional linguistics" (Imo 2014 and Hausendorf, 2015). However, the mix and overlaps of these theories have been clarified by the interpretation and argumentation of Jucker & Dürscheid (2012), who coined a term "keyboard-to-screen communication". Apart from physical keyboards, the sender also uses virtual keyboard embedded in the app on the smartphone to write the message and sends it to the recipient's screen. As Beißwenger (cf. 2007: 1) notes, the communication represented by chats, SMS, WhatsApp, and instant messages, is reliant on computer networks and infrastructure in order to function effectively. Specifically, in chat communication, responders can scroll the screen to enable an overview of the whole chat (Lee & Barton, 2013: 40). The challenges of scrolling arise from the fact that multiple topics can run parallel to each other in "simultaneous" discussions (cf. AlQbailat, 2020), and that available digital infrastructures can influence the individual actions of online interlocutors (cf. Androutsopoulos, 2016: 8). Therefore, it is beneficial to observe the use and structure of the response function to find out when and how people usually use this new technical function.

Chat, a type of keyboard-to-screen communication, is an internet-based service that enables synchronous written text communication in groups as well as in dynamic communication (Beck, 2006: 118), with the characteristics of synchronicity (or quasi-synchronicity) (see Dürscheid & Flick, 2014), written orality (cf. Günthner & Wyss, 1996: 70; Koch & Oesterreicher, 2007; Beck, 2006: 73; Dürscheid, 2016; Androutsopoulos, 2007: 80) and sequential organization (cf. Schegloff, 2007). Writing on the platform and with the end device is an "interactive action" and "the contributions follow each other quasi-synchronously" (cf. Dürscheid & Brommer, 2009: 16). The sequential and serial ordering (cf. Meier et al., 2020) reconstructs the flexible reordering of the conversation (cf.



ibid.) and follows a written language formulation pattern (cf. Beißwenger & Storrer, 2012: 92).

There are two sources of data for this paper: the Mobile Communication Database 2 (MoCoDa2), and self-collected data from WhatsApp. The MoCoDa2, developed in a cooperation among the University of Duisburg-Essen, the University of Hamburg, and WWU Münster, collected and processed 927 chats on everyday communications with 38,590 messages and 3,361 chat contributors, for the purpose of linguistic scientific research. Thirty chats were manually selected and classified by the author in MoCoDa2. These chats comprised of 11 one-on-one chats and 19 group chats. Since the reply function was not annotated in the database, all chats in the database had to be manually read through. Among the 927 chats, only 30 of them exhibited the quote function. Additionally, the author glanced over all personal daily chats in WhatsApp from September 2022 to December 2022 and selected five chats that typically reflected the reply function and its structure. The data in MoCoDa2 is publicly available, and the self-collected data is under the verbal consent of the participants. For the analysis, one-on-one chats and group chats were considered separately. The chats were donated by the conversation participants and are, therefore, authentic, and represent original dialogic communication situations in chats. The empirical authenticity of the data collection is advantageous for precise analysis (cf. Deppermann, 2008: 105; cf. Becker-Mrotzek & Brünner, 2006: 5).

### 3. Findings and discussion

#### 3.1 The Reply Function in One-on-One Chats

Since each message in one-on-one chats is directed at a specific interlocutor, the reply function not only addresses the intended recipient but also considers the message content, thus serving multiple content-related functions. Firstly, the interlocutor addresses the thematic content and provides further expansion on the previous topic. By quoting the previous message, the response to the information given above can be completed through the action of citing. Secondly, the interlocutor forms pairs of sequences such as “event-reaction sequence” or “question-answer sequence” to achieve typical communication goals. For instance, if an event, experience, or interesting video occurs, the interlocutor expects a corresponding comment or reaction to it. If a question is asked, the other interlocutor expects an answer. Thus, the reply function allows speakers to answer multiple events in an organized manner without causing chaos. Thirdly, it is also noticeable that one can simulate monologic and continuous speech by quoting one’s own words. The quoted message, which is also written by oneself, and the quoting message, which is currently being sent in a response sentence, form, in combination, a continuous story. What the interlocutor states afterward follows what they have already stated. This gives the possibility that, despite the limitations of chat communication, the interlocutor can still speak/write continuously without being disturbed by interruptions or

contributions from the other side.

In addition, the use of the reply function is mainly associated with the timing of the messages. Normally, interlocutors will read the messages and respond to them, following the predetermined order of the posts. Thus, if the reply function is used, the quoted messages will appear in sequential order, that is, the interlocutor typically quotes are the first message that appears on the chat or screen. Certain types of messages including voice messages that establish a particular sequence employ the reply function in a manner that necessitates the reader’s response to the initial message prior to processing any subsequent messages. However, the order of quoting interesting content is different from the arrangement followed by the timing. When videos or eye-catching images are shared in the chat, individuals are inclined to respond to these messages prior to any others. This change of sequential order of reply function results in “interesting” content being quoted and responded to first, with other messages being attended to subsequently.

```
#f6F8W
#150 01:59 Leyla Meme (die Simpsons im Auto, alle
gucken Homer böse an. Text: Wenn dein
Witz die Unterhaltung zerstört, aber
du ganz genau weißt, dass er der
Hammer war)
#151 09:32 Leyla uRL Kategorie 1 das müssen wir
nächstes Mal unbedingt probieren 😊
#152 10:15 Jonathan Majestic - 10/10
#153 10:17 Jonathan Ein Glück keine Bachata Songs mit TS,
wär halt locker ein Genickbruch 😊
#154 10:18 Jonathan >> #150 Leyla Meme...
was?! 😊
#155 10:19 Jonathan >> #151 Leyla uRL Kategorie 1...
Ich mein was ist die Coreo? Tanz mir
den Apache?
#156 10:22 Leyla >> #152 Jonathan Majestic - 10/10
😊
#157 10:22 Leyla >> #153 Jonathan Ein Glück keine
Bachata Songs mit TS, ...
TS?
#158 10:22 Leyla >> #155 Ich mein was ist die
Coreo?...
😊😊😊
#159 11:08 Jonathan Taylor Swift
#160 11:09 Leyla Achso 😊
```

Figure 2: Example of a one-on-one chat (MoCoDa2: #f6F8W)

```
#f6F8W
#150 01:59 Leyla Meme (The Simpsons in the car, everyone
looking at Homer angrily. Text: If your joke
ruins the conversation, but
you know for sure it was hilarious)
#151 09:32 Leyla uRL Category 1, we definitely have to try
it next time 😊
#152 10:15 Jonathan Majestic - 10/10
#153 10:17 Jonathan Luckily no Bachata songs with TS,
would be a neck-breaker 😊
#154 10:18 Jonathan >> #150 Leyla Meme...
what?! 😊
#155 10:19 Jonathan >> #151 Leyla uRL Category 1...
I mean, what's the choreo? Dance me
the Apache?
#156 10:22 Leyla >> #152 Jonathan Majestic - 10/10
😊
#157 10:22 Leyla >> #153 Jonathan Luckily no Bachata...
TS?
#158 10:22 Leyla >> #155 Jonathan I mean, what's the...
😊😊😊
#159 11:08 Jonathan Taylor Swift
#160 11:09 Leyla Oh, okay 😊
```

Figure 3: Translation of Figure 2

This example, #f6F8W, demonstrates how the reply function forms a pair sequence in a one-on-one chat. The conversation involves Leyla and Jonathan, who are dance partners and are developing a romantic relationship despite Leyla being Jonathan's subtenant. Prior to the start of this chat, they had already bid goodnight, but nevertheless, Leyla sent a meme (#150). The following morning, Leyla provided more about the dance organization and sends a private link. Jonathan reacted positively to the recommendation but conveyed his dislike for TS, another singer. In #154, Jonathan responded to the previous night's meme, exhibiting surprise and delight with emojis. He then engaged in a joke in #155 by pointing out the Coreo (a type of dance) and Apache (a pop singer). Leyla replied to Jonathan's messages by giving two comments/reactions (#156 and #158), seeking clarification regarding the meaning of TS. This section is concluded with Leyla acknowledging the clarification.

### 3.2 Reply Function in Group Chats

In group chats, it is not necessary to respond to every message, as some messages may not be targeted at a specific individual or may be irrelevant or unattractive. However, all messages are sent to the same virtual chat room, which may lead to problems when addressing and organizing parallel topics. Confusion can occur when several topics are discussed in the group chat, with different individuals being involved in one or more topics. The reply function can help organize the conversations and assign topics by classifying messages into different subtopics. This approach clarifies which individuals are discussing which topics at any given moment.

The reply function can also be used when addressing a third person or a group. In this situation, it functions more as a quoting tool than a direct response to the original sender. The interlocutor can quote a specific word or information from a previous message and use it in their own message, potentially addressing it to the whole group or a third party. In this case, the function of responding to original sender is less significant than that of quoting the specific information directed to the potential addressee. It's important to note that the original sender and the addressee can be independent of each other since the reply function allows for multiple uses. Furthermore, the reply function also has additional functions in group chats, such as disambiguation and forwarding of previous topics.

```
#jrP4p
#17 Jennifer 21:06 Muss eigentlich noch wer von euch am 1.12
arbeiten? 🤔
#18 Susanne 21:10 Ich...ich kann da abend auch nicht hase...
#19 Rolf 21:11 Würde dir das Geld geben, fänd ich
einfacher 😊
Getränke klappt ja ganz gut immer :) Und
moi 🍷🍷 werde daher nicht so lange
chillen
#20 Rolf 21:11 >> #18 Susanne Ich...ich kann da abend...
Menno :(
#21 Jennifer 21:12 >> #20 Menno :(
Ja das meinte ich doch mit aufteilen 😊
Wann denn? Ich muss erst um 15 Uhr 🕒
#22 Jennifer 21:12 >> #18 Susanne Ich...ich kann da abend...
Schade!
```

Figure 4: Example of a group chat (MoCoDa2: #jrP4p)

```
#jrP4p Date
#17 Jennifer 21:06 Does anyone else have to work on December
1st? 🤔
#18 Susanne 21:10 I... I can't in the evening either...
#19 Rolf 21:11 I'd give you the money, that would be
easier 😊
Drinks always work quite well :) And
moi 🍷🍷 so I won't hang out for too long
#20 Rolf 21:11 >> #18 Susanne I... I can't ...
Aw :(
#21 Jennifer 21:12 >> #20 Aw :(
Yeah, that's what I meant by splitting it
😊 When then? I don't finish until 3pm 🕒
#22 Jennifer 21:12 >> #18 Susanne I... I can't ...
Too bad!
```

Figure 5: Translation of Figure 4

Example #jrP4p illustrates how the reply function works in group chats. Before #17, they discussed when they could make an appointment and who wanted to sign up for it. As some members may not be able to attend for various reasons, in #17, Jennifer asked who was unable to participate due to work commitments. Then, Susanne mentioned her absence because of a work obligation, which becomes the first subtopic. After a minute, Rolf introduced the second subtopic, proposing, for the sake of convenience, that they should buy the drinks together. Both Susanne and Rolf did not use the reply function because Susanne's response was not closely related to the question, and Rolf's message was a new subtopic and not a direct response to any previous message. When #17, #18, and #19 appeared on the screen of this group chat, chat participants must decide which subtopic to continue discussing. Rolf, leaving the topic of drink, first used the reply function to respond to Susanne, sympathizing with her. Jennifer then commented first on the drinking topic and subsequently wrote "too bad" regarding Susanne's absence.

## 4. Conclusion

To conclude, this paper provides a detailed analysis of the functions and structures of the reply function in chat communication, specifically in one-on-one and group chats. The study utilizes the MoCoDa2 corpus and self-collected data to identify patterns in the use of the reply function. The findings suggest that the reply function serves different purposes in different contexts. In one-on-one chats, it is used to refer thematic contents, form pair sequences, and simulate a continuous monologue. In contrast, in group chats, it is used to address one or more individuals or clarify misunderstandings. Additionally, the paper highlights the importance of considering various factors that influence the sequence of quotations in chat communication. Overall, this study contributes to a better understanding of the role of the reply function in internet-based and internet-supported chat communication. By using the reply function, which allows one to respond to a specific message, communication becomes more effective and precise. The reply function is also helpful for organizing conversations and prioritizing certain contributions. When multiple people are active in a group chat and discussing various topics, the reply function can help focus on a specific conversation. By replying to a specific message, one shows

interest in a message or topic and can express a response accordingly. This contributes to a pleasant and respectful conversation atmosphere and increases the possibility of reconstructing conversations. The reply function, as a meaningful addition to chat platforms, has the potential to improve communication and interaction between people.

## 5. Acknowledgement

The author expresses gratitude to Prof. Dr. Florence Oloff for her support and advice.

## 6. References

- AlQbailat, N. M. I. (2020). *Internet linguistics: a conversational analysis of online synchronous chat and face-to-face conversations of efl undergraduate students in jordan*. <http://hdl.handle.net/10016/29806>
- Androutsopoulos, J. (2007). Bilingualism in the mass media and on the internet. *Bilingualism: A social approach*, 207-230. [https://doi.org/10.1057/9780230596047\\_10](https://doi.org/10.1057/9780230596047_10)
- Androutsopoulos, J. (2016). 13 Theorizing media, mediation and mediatization. *Sociolinguistics: theoretical debates*, 282.
- Beck, K. (2006). *Computervermittelte Kommunikation im Internet*. Oldenbourg Wissenschaftsverlag. <https://doi.org/10.1524/9783486839203>
- Becker-Mrotzek, M., & Brünner, G. (2006). *Gesprächsanalyse und Gesprächsführung*, 2. Auflage Stuttgart.
- Beißwenger, M. (2007). Corpora zur computervermittelten (internetbasierten) Kommunikation. *Zeitschrift für germanistische Linguistik*, 35(3), 496-503. <https://doi.org/10.1515/zgl.2007.035>
- Beißwenger, M., & Storrer, A. (2012). Interaktionsorientiertes Schreiben und interaktive Lesespiele in der Chat-Kommunikation. *Zeitschrift für Literaturwissenschaft und Linguistik*, 168, 92-124. <https://nl.ijs.si/janes/wp-content/uploads/2014/09/beisswengerstorrer12.pdf>
- Crystal, David (2011): *Internet Linguistics: A Student Guide*. London: Routledge.
- Deppermann, A. (2008). *Gespräche analysieren. Eine Einführung*. 4. Aufl. Wiesbaden. <https://link.springer.com/content/pdf/10.1007/978-3-531-91973-7.pdf>
- Dürscheid, C. (2016). Neue Dialoge–alte Konzepte?. *Zeitschrift für germanistische Linguistik*, 44(3), 437-468. <https://doi.org/10.1515/zgl-2016-0023>
- Dürscheid, C., & Brommer, S. (2009). Getippte Dialoge in neuen Medien. Sprachkritische Aspekte und linguistische Analysen. *Linguistik online*, 37(1). <https://doi.org/10.13092/lo.37.511>
- Dürscheid, C., & Frick, K. (2014). Keyboard-to-Screen-Kommunikation gestern und heute: SMS und WhatsApp im Vergleich. *Sprachen*, 149-181. [https://www.researchgate.net/profile/Christa-Duerscheid/publication/281188294\\_Keyboard-to-Screen-Kommunikation\\_gestern\\_und\\_heute\\_SMS\\_und\\_WhatsApp\\_im\\_Vergleich/links/57206f4708aefa64889a9519/Keyboard-to-Screen-Kommunikation-gestern-und-heute-SMS-und-WhatsApp-im-Vergleich.pdf](https://www.researchgate.net/profile/Christa-Duerscheid/publication/281188294_Keyboard-to-Screen-Kommunikation_gestern_und_heute_SMS_und_WhatsApp_im_Vergleich/links/57206f4708aefa64889a9519/Keyboard-to-Screen-Kommunikation-gestern-und-heute-SMS-und-WhatsApp-im-Vergleich.pdf)
- Günther, U., & Wyss, E. L. (1996). *E-mail-Briefe-eine neue Textsorte zwischen Mündlichkeit und Schriftlichkeit* (pp. 61-86). Peter Lang. <https://digitalcollection.zhaw.ch/handle/11475/2592>
- Hausendorf, H. (2015). Interaktionslinguistik. In Sprachwissenschaft im Fokus. *Positionsbestimmungen und Perspektiven* (pp. 43-69). de Gruyter. <https://doi.org/10.1515/9783110401592.43>
- Herring, S. (1996). Linguistic and critical analysis of computer-mediated communication: Some ethical and scholarly considerations. *The information society*, 12(2), 153-168. <https://doi.org/10.1080/911232343>
- Imo, W. (2014): Interaktionale Linguistik. In: Staffeldt, Sven und Jörg Hagemann (Hrsg.): *Pragmatiktheorien*. Tübingen: Stauffenburg, 49–82. <https://link.springer.com/content/pdf/10.1007/978-3-476-05549-1.pdf>
- Imo, W. (2016). Dialogizität – eine Einführung. *Zeitschrift für germanistische Linguistik*, 44(3), 337-356. <https://doi.org/10.1515/zgl-2016-0019>
- Jucker, A. H., & Dürscheid, C. (2012). The linguistics of keyboard-to-screen communication. A new terminological framework. *Linguistik online*, 56(6/12), 1-26. [http://www.linguistik-online.org/56\\_12/juckerDuerscheid.html](http://www.linguistik-online.org/56_12/juckerDuerscheid.html)
- Koch, P., & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik*, 35(3), 346-375. <https://doi.org/10.1515/zgl.2007.024>
- Lee, C., & Barton, D. (2013). *Language online: Investigating digital texts and practices*. Routledge. [https://www.academia.edu/download/53705587/\\_David\\_Barton\\_Carmen\\_Lee\\_Language\\_Online.pdf](https://www.academia.edu/download/53705587/_David_Barton_Carmen_Lee_Language_Online.pdf)
- Meier, S., Viehhauser, G., & Sahle, P. (Eds.). (2020). *Rekontextualisierung als Forschungsparadigma des Digitalen* (Vol. 14). BoD–Books on Demand. [https://kups.ub.uni-koeln.de/29390/1/SIDE14\\_Rekontextualisierung\\_als\\_Forschungsparadigma\\_des\\_Digitalen.pdf](https://kups.ub.uni-koeln.de/29390/1/SIDE14_Rekontextualisierung_als_Forschungsparadigma_des_Digitalen.pdf)
- Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I* (Vol. 1). Cambridge university press.
- Yus, F. (2011). *Cyberpragmatics: Internet-mediated communication in context*. John Benjamins Publishing Company. <http://library.oapen.org/handle/20.500.12657/30723>
- Mobile Communication Database 2 (MoCoDa2): <https://db.mocoda2.de> (last accessed on 24.12.2022)
- WhatsApp: [https://faq.whatsapp.com/3240283752856917/?helpref=uf\\_share](https://faq.whatsapp.com/3240283752856917/?helpref=uf_share) (last accessed on 21.04.2023)



# Ellipsis Points in Messaging Interactions and on Wikipedia Talk Pages

Michael Beißwenger, Eva Gredel, Lena Rebhan, Sarah Steinsiek

University of Duisburg-Essen

## Abstract

In this paper, we examine the usage of ellipsis points (EP) in two genres of computer-mediated communication (CMC) using corpora. In two studies, we describe and compare the formal and functional characteristics of EP usage in WhatsApp chats and on Wikipedia talk pages. (1) We present a typology of pragmatic functions of EP in WhatsApp interactions that has been derived from the analysis of a corpus sample and discuss how the practices of EP usage in these data originate from traditions of writing. (2) We investigate typographic and allographic variation of EP on Wikipedia talk pages and examine whether the categories resulting from study 1 also fit for this type of CMC discourse. Our analyzes show that EP are frequently used for the sequential organization of written interactions and relationship management between interlocutors in different CMC genres.

**Keywords:** ellipsis points, pragmatics, computer-mediated communication, corpora, text messaging, Wikipedia

## 1. Introduction

In the past decade there has been a growing interest of linguistics in the pragmatics of written interactional discourse in computer-mediated communication (CMC). This paper adds to the pragmatic knowledge on how interlocutors face the challenges of written interpersonal communication in the digital sphere by an examination of the usage of ellipsis points (henceforth: *EP*) in messaging interactions and on Wikipedia talk pages. The reported work builds on previous research on EP as elements of the writing system and in CMC. We analyze a random sample from a WhatsApp corpus and query results from the Wikipedia corpora provided via the corpus infrastructure of the Leibniz Institute for the German Language (IDS). In a first study, we present a typology of pragmatic functions of EP that has been derived from the analysis of a randomized sample extracted from the Mobile Communication Database (MoCoDa2) and discuss to what extent the ‘novel’ practices of EP usage found in text messaging interactions originate from traditions of writing. In a second study, we investigate typographic and allographic variation of EP and examine whether the typology from study 1 is also suitable for the analysis of a random sample of posts from Wikipedia talk pages. The results of both studies illustrate the flexibility of the written tradition to be adapted to new domains of communication and social interaction.

## 2. Ellipsis points as elements of the writing system and in CMC

According to the official rules of standard German orthography, ellipsis points are to be used to indicate that elements of words, sentences or texts have been omitted (see AR, 2018: 100, § 99). They are used, for example, in academic papers to denote omissions within quotations or, in private correspondence, to allude to rather than write out taboo words (“Du bist echt ein A...!”, ‘You’re a real a...hole!’). However, ellipses are also often used rather stylistically. In direct speech in literary texts for instance, ellipses can serve as a linguistic device to instruct the reader to imagine the respective written text parts as utterances spoken by a literary figure:

“[N]ovelists developed special conventions involving choice of vocabulary and syntactical features, but they also imposed new conventions of layout and punctuation up on the printer to make it as clear to the reader as possible that the representation of spoken language was intended” (Parkes, 1992: 93).

Following Parkes (1992), EP are punctuation devices that originate from practices of the mimetic representation of spoken language in the written medium (see Bredel, 2011: 13). Examples 1 and 2 illustrate the use of these practices in contemporary literature (the functional categories used in the captions will be introduced in Sect. 4). As we will show in Sect. 4 and 5, it is important to bear in mind these practices from literature when it comes to the analysis of EP in CMC interactions. Another important background for a pragmatic reconstruction of practices of EP usage is provided by Bredel (2011: 47) who considers the involvement of the reader an essential feature of how EP support the cooperation of writers and readers in text communication: They instruct readers to activate their own knowledge (of the co-text and/or context) and fill in missing information on a lexical, syntactic or even pragmatic level.

Ex. 1: Transmodal segmentation with EP in a comic book (*Too Much Coffee Man saves the universe*, 1997, p. 1):



Ex. 2: Implying and other-party selection with EP in a novel (Stan Jones: *Village of the Ghost Bears*, 2009, p. 38):

“We’ve got to get that guy out of One-Way Lake,”  
Active told the pilot. “If you could just. . . .”

“Sorry, man, it’ll have to wait till tomorrow,” Cowboy  
said. “He’s not going anywhere, right?”

Active frowned. “I still don’t like leaving him up there.  
This time of year, everything’s on the move and hungry.  
Bears, wolves, foxes, ravens. Wolverines too.”

Cowboy gave him a what-can-I-do? shrug. “One more  
day won’t hurt.”

The two studies we present in our paper focus on the use of ellipses in written CMC interactions that are organized in sequences of posts, namely text messaging (WhatsApp, Sect. 4) and on Wikipedia talk pages (Sect. 5), where interlocutors tailor their messages (CMC posts) to fit the interactional context and sequential structure (= strategy of *interaction-oriented writing*, Storrer, 2012; 2018). We will show that, from a pragmatic perspective, the uses of EP in this type of discourse have more in common with those in literary language than with uses that conform to the official standard rules which hold for the written standard language of “traditional” types of text.

### 3. Related work

In the past years there has been increasing research interest in the pragmatics of CMC (e.g., Herring/Stein/Virtanen, 2013, Meier-Vieracker et al., 2023) with a special focus on practices of adapting the resources of the writing system to the requirements of sequential interaction (e.g. Beißwenger, 2016, Androutsopoulos/Busch, 2020). In this research context, EP – as an element of the contemporary orthographic standard with a history that traces back to practices of adapting the writing system for the mimetic representation of spoken language – can be considered a resource that is downright predestined for the requirements of written interactional discourse, while the official orthographic standard restricts their use to the indication of omissions on the word, sentence or text level.

Androutsopoulos (2020) gives a detailed overview and critical appraisal of the international state of research on the use of EP in CMC. In our own work, we build on the examination of EP presented in Androutsopoulos’ paper. The author expands on the functional typology of EP suggested by Meibauer (2007). While Meibauer’s typology is neither empirically based nor takes into account written practices in CMC (but only the use of ellipses in ‘traditional’ text genres), Androutsopoulos analyzes 353 Facebook posts by Greek high schoolers and shows that the function of ellipses to indicate omissions (see Meibauer, 2007) is of no significance in this type of CMC at all (see Androutsopoulos, 2020: 154). Instead, EP in message-final position are used to convey a certain overtone or imply something and those used within posts are a means of text segmentation (ibid.: 150; Meibauer refers to this function as *connection*). In this sense, they have a syntactic function similar to other punctuation marks. However, ellipses are more salient, which is why Androutsopoulos (2020: 155) terms them as “eine Art Allzweck-Segmentierer” – an all-

purpose remedy for segmentation.

In his study on register variation of German middle and high school students, Busch (2021) analyzes WhatsApp chats and shows that ellipses are also used to mitigate face threats, as a means of cohesion, and as a technique for sequential organization/other-selection, i.e. to directly address and elicit input from other interlocutors (see ibid.: 391). Busch points out that EP can take on several functions at once (see ibid.: 405).

In summary, both Androutsopoulos (2020) and Busch (2021) show that EP serve many different purposes – except for the one purpose that is codified in the official rules of German orthography: to signal the omission of words or text components.

### 4. Investigating ellipsis points in messaging interactions

The case study on WhatsApp data reported in this section is based on two random samples of WhatsApp messages drawn in 2021 and 2022 from the *Mobile Communication Database (MoCoDa2)*, a crowdsourced corpus of German WhatsApp chats that is freely available online for research purposes and teaching under the following link: <https://db.mocoda2.de/> (Beißwenger et al., 2019). The goals and subtasks of the study were the following:

- (1) In the light of the findings of Androutsopoulos (2020) and Busch (2021), a critical revision of the functional typology of EP by Meibauer (2007) based on empirical data can be considered a desideratum – not only from the perspective of CMC research but also from the perspective of pragmatic research in written practices (including, but not limited to CMC) in general. One goal of the study was to explore relevant categories of such a typology based on a small but randomized sample (N=100 posts with 108 true positive occurrences of EP). The focus of this exploratory study was that of CMC research, however, categories for ‘traditional’ texts suggested in previous research (especially Meibauer, 2007) were integrated with some necessary modifications, even though they prove not to be relevant for the analysis of our CMC sample.
- (2) The occurrence of instances of the functional categories from our typology was quantified in order to receive a first impression of their relevance in CMC (Androutsopoulos, 2020 gives detailed descriptions of the practices found in his data but does not quantify them). To achieve this goal, a second randomized sample of similar size (N=100 posts comprising another 108 true positive instances) from the same corpus was coded by the two researchers in a hermeneutic procedure.

By providing evidence for these two goals, the study adds to knowledge on the pragmatics of EP in CMC. In the following we present the key results for (1) and (2). A detailed description of the analyzes and findings is given in Beißwenger/Steinsiek (2023).

Similar to the findings in Androutsopoulos (2020) and Busch (2021), our quantitative analysis shows that the

standard-conformant use of EP obviously does not play a role in WhatsApp interactions. Instead, ‘jobs’ related to sequential organization, the construction of interactional coherence (sensu Herring, 1999) and relationship management are dominant in the data.

A simplified version of the typology resulting from subtask (1) and the frequencies found as a result of subtask (2) are given in Tab. 1.

Type	Occ.	%
<b>Omission</b>	1	0,93
<b>Implying</b>	28	25,93
<b>Sequential Organization</b>		
– other-selection	13	12,04
– self-selection	7	6,48
<b>Segmentation</b>		
– visual	41	37,96
– transmodal	16	14,81
<i>More than one possible interpretation</i>	2	1,85
<b>Total</b>	<b>108</b>	<b>100,00</b>

Table 1: Pragmatic functions of EP in the WhatsApp sample and frequency of instances per category.

According to our findings, the most frequent types of ellipsis usages are the following (for a detailed description of all types see Beißwenger/Steinsiek, 2023):

#### Segmentation:

- *Visual*: The ellipsis serves as a marker of boundaries between sentences, syntactic components or communicative units and supports the reading (scanning) process of the recipient. Example (#Aqkwk):

Norbert: Gruess dich! Jetzt hast du auch meine nummer ...  
lg, norbert *'Hi there! Now you have my number too ... br, norbert'* [The German acronym ‘lg’ stands for ‘Liebe Grüße’, which corresponds to ‘br’ (‘best regards’) in the translation.]

- *Transmodal*: The ellipsis is used to simulate nonverbal signs in spoken language like gaps or changes of intonation, for instance to lay stress on something (see Androutsopoulos, 2020: 135), e.g. a punch line (‘typographic silence’, see Busch, 2021: 387). Example with English translation (#n3716):

Emma: Melde dich wenn ich dir wieder gut genug bin, so lange nerve ich dich nicht. Bin echt etwas enttäuscht muss ich sagen.. Trotzdem wünsche ich dir später eine gute Nacht und schöne Träume, viel Spaß noch auf dem Geburtstag und pass auf dich auf ja 🥰 (shortened)  
*'Let me know when I'm good enough for you again, I won't bug you anymore until then. I'm honestly a little disappointed.. Good night though and sweet dreams, have fun at that birthday party and take care 🥰'*

**Implying**: The recipient is supposed to infer what the author is implying based on common knowledge or by making assumptions about them and their opinions. Example: Christina replies to Johannes’s question whether

she could pick up him and a friend (#rgsLe):

Christina: Hab selber Alkohol getrunken  
*'I've already had a couple of drinks myself'*  
Johannes: ....pff 🙄

**Sequential organization**: Based on common knowledge of sequential organization and conditional relevance in spoken conversations, the ellipsis is intended to be interpreted as an imitation of ‘next speaker selection’.

- *Other-selection* (more or less explicit): The recipient is supposed to (1) take on the role of the author and reply to the current post or to (2) infer that the current author has nothing (more) to contribute at this point of the ongoing conversation. Example of a more explicit other-selection (#y91fl):

Muriel: Schick mal deine emailadresse.hab ich irgendwie nich mehr.. *'Give me your email address.i somehow don't have it anymore..'*

The example illustrates that other-selection may pose a potential face-threat: Without the EP, Muriel’s request for the email address might be interpreted as an order. Thus, in this case, the EP additionally serve as a softener to mitigate the face-threatening act that pragmatically results from the request (see Beißwenger/Steinsiek, 2023).

- *Self-selection* (imitation of floor keeping strategies in spoken conversation): Based on previous posts, (1) a planned expansion can be projected by the usage of EP in final position or (2) an ellipsis in initial position can be interpreted as a cohesive device. Example (#OGOME):

Luisa: Ach Quatsch stört mich nie :)  
*'Oh that [if your place is a mess] doesn't bother me at all :)'*  
Luisa: ... bei anderen :D in meiner wg treibt mich das zur Weißglut aber das ist ein anderes Thema 🙄 *'... as long as it's not my place :D the mess in my dorm drives me crazy but that's a different issue 🙄'*

Both other- and self-selection help to establish interactional coherence (Herring, 1999) under the conditions of CMC. The results support the assumption that the usage of EP in WhatsApp interactions resembles the functions found in literary prose (and also in comics and graphic novels) where they are used to represent features of spoken language in written utterances assigned to fictional characters (illustrated in examples 1 and 2). Thus, even though the usage of EP in WhatsApp appears to be ‘non-conformant’ with standard orthography, we consider these practices as building on traditions that have already been established in our literate culture, and that are adopted as devices to face the challenges of natural interpersonal conversation in the digital sphere.

## 5. Investigating ellipsis points on Wikipedia talk pages

In this study, we are interested in how EP are used in a further CMC genre – namely Wikipedia talk pages. For this objective, we access data from the most recent releases of the Wikipedia corpora compiled at the IDS (see Margaretha & Lungen, 2014) via COSMAS II (2023). These corpora contain data from three different types of German Wikipedia talk pages: article talk pages (wdd19), user talk pages (wud17), and the redundancy talk pages (wrdd17). On the article talk pages, authors negotiate the online encyclopedic content of the respective associated Wikipedia entries, on the user talk pages the edits of individual authors are discussed, and on the redundancy talk pages Wikipedians decide whether an article should be deleted (see Gredel, 2020). Although these different types of talk pages each serve different purposes in Wikipedia, they have similarities at the linguistic level: They share features of CMC genres such as a dialogic, sequential structure and an informal writing style with non-standard language (see Storrer, 2017).

Nevertheless, the postings on the talk pages differ from those in other CMC genres like Facebook and WhatsApp: Wikipedia shows less everyday communication, but rather interaction with the overall goal to collaboratively create an online encyclopedia. Against this background, special practices for negotiating collaborative text production have developed on the Wikipedia talk pages (see Beißwenger, 2016; Gredel, 2017). In order to further explore the specific practices of using EP in CMC genres, this study represents an important addition to previous work on EP in CMC genres (see Androutsopoulos, 2018; 2020; Beißwenger/Steinsiek, 2023).

The case study on Wikipedia talk pages accesses EP in the Wikipedia corpora via two different types of search queries. The first type of queries retrieves the occurrences of EP from the corpora where the EP were entered as an html entity (Unicode U+2026) via the Wikipedia editor. This query only returns occurrences with the norm-compliant form variant (three dots) and shows for all three subcorpora that authors frequently use the html entity “...” on Wikipedia talk pages (Tab. 2).

Corpus Sigle	Corpus size (in token)	Occurrences of EP	pMW
wdd19	414,929,118	60,184	144.7
wud17	326,214,993	33,364	102.3
wrdd17	1,951,044	140	71.76

Table 2: Results of the search query for the html entity “...” (Unicode U+2026) in COSMAS II.

The second type of queries reveals occurrences of EP from the corpus where Wikipedia authors entered three (or more) “single” points without inserting the beforementioned html entity. On the one hand, this type of corpus query returns a large number of false positives that must be sorted out by manual annotation. On the other hand, these results are particularly interesting for linguistics, since it is through

this type of search queries that allographic variants are discovered that do not conform to the German norm codified in the official rules of standard German orthography in the “Amtliche Regelwerk” (AR 2018). With regard to the Wikipedia data, it can be noted that the spectrum of different allographic variants of EP is much larger than in other CMC genres: While allographic variants with more than five points do not or hardly occur in the WhatsApp data (Beißwenger/Steinsiek 2023) as well as in the Facebook data (Androutsopoulos 2020), they are present in the Wikipedia Data. Instances of EP with up to 15 “single” points can be reconstructed via the corpora:

- (1) Anbei zwei Fotos von mir zur allfälligen Verwendung im Artikel (Ich habe sie leider mit “Greif” anstatt mit “Stral” gekennzeichnet, was ich nicht mehr wegbringe) dringend [= user name as part of the signature] 22:50, 04.02.2012  
‘Attached are two photos of mine for use in the article (I unfortunately marked them “Greif” instead of “Stral”, which I can’t get rid of)’
- (2) ..... Habe es weggebracht und korrigiert.-- dringend 12:19, 25.02.2012  
‘..... Took it away and corrected it.’ (WDD19/A0067.40545)

User “dringend” has made what they consider to be an incorrect edit to a Wikipedia article on 04.12.2012 which they would now like to revert. For technical reasons, they are initially unable to do so (posting 1). According to the timestamp of the second posting, they report back on 25.02.2012 that they have now succeeded in making the revert themselves. At the beginning of their second posting, they set 15 single points in initial position, which connect the two postings like cohesion devices. With the fivefold reduplication of the norm-compliant variant the user shows symbolically that a long period of time (after all 19 days) has passed between their postings in which they have tried to fix the error.

### Sequential Organization

The pragmatic function of the allographic variant of EP in posting (1) is “sequential organization” – subtype “self-selection” (Beißwenger/Steinsiek, 2023) with an iconic dimension. This example suggests that in Wikipedia, too, the pragmatic functions of EP go far beyond those formulated in the Amtliche Regelwerk (2018). Initial corpus evidence makes clear that other (sub)types are also relevant on Wikipedia talk pages:

- (3) Und meine beiden Fragen habt ihr immer noch nicht beantwortet ... -- Phi Φ 21:02, 20.04.2013 And you still haven't answered my two questions ... (WDD19/B0070.75479)

In posting (3), the request to answer formulated questions is generated by posting-final EP. This occurrence of EP fulfils the function “sequential organization” – subtype “other-selection”.

### Omission and Segmentation

The other pragmatic functions of EP according to Beißwenger/Steinsiek (2023) are also relevant on



Wikipedia talk pages. The author of a Wikipedia edit in the article on a brewery is asked whether other types of beer from the respective brewery should not be listed in the entry. He answers the following:

- (4) Kann sein. Es steht ja dort: „viele Biersorten“ und dann „darunter sind“, erhebt also keinen Anspruch auf Vollständigkeit. Wenn du sie für wichtig hältst, darfst du sie austrin...äh... einbauen. Gruß. -- Peng nfu-peng 12:28, 12.07.2007 *Maybe. It says there are many types of beer and then among them, so it doesn't claim to be exhaustive. If you think they are important, you may drin...er... include them. Greetings.* [the German *äh* is a hesitation marker which is translated here with the English *er*] (WDD19/E0027.51475)

In this example, the author jokingly stages that he realizes a Freudian slip in his contribution to the discussion, which he then corrects himself – in the sense of a self-repair: The first of the two EP instances in example (4) is of the type “omission”, since the German word *trinken* (English: *drink*) is not written out completely here. The second instance is a transmodal segmentation in order to verbally realize the time delay in reformulating the Freudian slip.

### Implying

Example (5) is preceded by a conflict of two authors over the definition of the terms *Marxism* and *Socialism*:

- (5) Danke für deine Aufklärung bzgl. Trotskismus und Rosdolsky ... :-( Wenn du das Manifest kennst, kannst du doch nicht behaupten, der (sic!) Sozialismus sei „integrativer Bestandteil des Marxismus“. -- redtux 17:44, 03.05.2008 *Thank you for your enlightenment regarding Trotskyism and Rosdolsky... :-( If you know the manifesto, you can't claim that (sic!) socialism is “an integral part of Marxism.”* (WDD19/L0043.38230)

From the EP in combination with the emoticon it can be deduced that the thanks verbalized by author “redtux” is not actually meant that way. In this context, the EP thus imply that the statement is meant differently than it was formulated. The statement is to be read as calculated inconsistency.

As these examples made clear, the occurrences of EP must be subjected to detailed analysis in order to investigate them appropriately in their respective contexts. Results of this analysis will be presented at the conference. For this analysis, we compiled a random sample of 100 posts from Wikipedia talk pages (corpus wdd19) containing EP. The random sample will include hits on both search strings (html entity and “single” points). In the qualitative part of our corpus study, we then address the following questions: What allographic variants of EP can be found on Wikipedia talk pages? In which position of a post do the EP appear? What is their function? This qualitative study builds on the function typology according to Beißwenger/Steinsiek (2023) to empirically investigate EP in the CMC genre of Wikipedia talk pages.

## 6. Conclusion and outlook

In this paper we have shown for two genres of CMC that even though the use of ellipsis points in messaging interactions and on Wikipedia talk pages seems to be deviant from the codified orthographic standard at first glance, it is not to be considered a radically new family of practices that emerge from the affordances and interactional nature of the digital sphere. Instead, these practices relate to traditions of writing where the representation of spoken language in the written medium is treated as a challenge of literary expression. What in fact is novel about the use of EP in the CMC genres analyzed here is that interlocutors on the internet obviously rely on their cultural experience with the written tradition and take up the practices found there to face the challenges of written interpersonal interaction in natural communication. This hypothesis promises to be productive for further corpus-based investigations of the interdependence of practices of interaction-oriented writing in the digital sphere and the written tradition.

## 7. References

- Androutsopoulos, J. (2018). Digitale Interpunktion: Stilistische Ressourcen und soziolinguistischer Wandel in der informellen digitalen Schriftlichkeit von Jugendlichen. In A. Ziegler (Eds.), *Jugendsprachen. Aktuelle Perspektiven internationaler Forschung*. Berlin/Boston: De Gruyter, pp. 721–748.
- Androutsopoulos, J. (2020). Auslassungspunkte in der schriftbasierten Interaktion. Sequenziell-topologische Analysen an Daten von griechischen Jugendlichen. In J. Androutsopoulos & F. Busch (Eds.), *Register des Graphischen. Variation, Interaktion und Reflexion in der digitalen Schriftlichkeit*. Berlin/Boston: De Gruyter, pp. 133–158.
- Androutsopoulos, J. & Busch, F. (eds., 2020): *Register des Graphischen: Variation, Interaktion und Reflexion in der digitalen Schriftlichkeit*. Berlin/New York: de Gruyter (Linguistik – Impulse & Tendenzen 87).[AR 2018] Deutsche Rechtschreibung. Regeln und Wörterverzeichnis. Aktualisierte Fassung des amtlichen Regelwerks entsprechend den Empfehlungen des Rats für deutsche Rechtschreibung 2016. Mannheim.
- Beißwenger, M. (2016). Praktiken in der internetbasierten Kommunikation. In A. Deppermann, H. Feilke, A. Linke (Eds.): *Sprachliche und kommunikative Praktiken*. Berlin/Boston: De Gruyter, pp. 279–310.
- Beißwenger, M., Imo, W., Fladrich, M. & Ziegler, E. (2019): <https://www.mocoda2.de>: a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions. In *European Journal of Applied Linguistics* 7(2), 333–344.
- Beißwenger, M. & Steinsiek, S. (2023). Interpunktion als interaktionale Ressource. Eine korpusgestützte Untersuchung zur Funktion von Auslassungspunkten in der internetbasierten Kommunikation. In M. Beißwenger, E. Gredel, L. Lemnitzer & R. Schneider (Eds.), *Korpusgestützte Sprachanalyse. Grundlagen, Anwendungen und Analysen*. Tübingen: narr francke

- attempto (Studien zur deutschen Sprache 88), pp. 287--310.
- Bredel, U. (2011). *Interpunktion*. Heidelberg: Winter.
- Busch, F. (2021). *Digitale Schreibregister. Kontexte, Formen und metapragmatische Reflexionen*. Berlin/Boston: De Gruyter.
- [COSMAS II 2023] Corpus Search, Management and Analysis System. Leibniz-Institut für Deutsche Sprache. Mannheim. <http://cosmas2.ids-mannheim.de>.
- Gredel, E. (2017). Digital discourse analysis and Wikipedia: Bridging the gap between Foucauldian discourse analysis and digital conversation analysis. In *Journal of Pragmatics* (115), pp. 99--114.
- Gredel, E. (2020). Digitale Diskursanalysen: Das Beispiel Wikipedia. In H. Lobin, K. Marx, A. Schmidt (Eds.): *Deutsch in Sozialen Medien*. Berlin/Boston: De Gruyter, pp. 247--264.
- Herring, S. C. (1999). Interactional Coherence in CMC. In: *Journal of Computer-Mediated Communication* 4(4).
- Herring, S., Stein, D. & Virtanen, Tuija (eds., 2013): *Pragmatics of Computer-Mediated Communication*. Boston: De Gruyter Mouton (Handbooks of Pragmatics 9).
- Margaretha, E. & Lungen, H. (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In M. Beißwenger, N. Oostdijk, A. Storrer & H. van den Heuvel (Eds.): *Building and Annotating Corpora of Computer-mediated Communication: Issues and Challenges at the Interface between Computational and Corpus Linguistics*. Journal for Language Technology and Computational Linguistics (JLCL), 29(2), pp. 59--82.
- Meibauer, J. (2007): Syngropheme als pragmatische Indikatoren: Anführung und Auslassung. In: S. Döring & J. Geilfuß-Wolfgang (Eds.): *Von der Pragmatik zur Grammatik*. Leipzig: Universitätsverlag, pp. 21--37.
- Meier-Vieracker, S., Bülow, L., Marx, K. & Mroczynski, Robert (eds., 2023): *Digitale Pragmatik*. Heidelberg: Metzler (Digitale Linguistik 1).
- Parkes, M. B. (1992). *Pause and Effect. An Introduction to the History of Punctuation in the West*. Aldershot: Scolar Press.
- Storrer, A. (2018): Interaktionsorientiertes Schreiben im Internet. In: A. Deppermann & S. Reineke, Silke (Eds.): *Sprache im kommunikativen, interaktiven und kulturellen Kontext*. Berlin/Boston: de Gruyter, pp. 219--244.
- Storrer, A. (2017). Grammatische Variation in Gespräch, Text und internetbasierter Kommunikation. In M. Konopka & A. Wöllstein (Eds.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, Berlin/New York: De Gruyter, pp. 105--125.
- Storrer, A. (2012). Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia. In: H. Feilke, J. Köster & M. Steinmetz (Eds.): *Textkompetenzen in der Sekundarstufe II*. Stuttgart: Fillibach bei Klett, pp. 277--304.

# The representation of the ‘Jew’ as enemy in French public Telegram channels within the identitarian-conspiratorial milieu

Laura Bothe

Universität Heidelberg

E-mail: laurabothe@stud.uni-heidelberg.de

## Abstract

The ‘Jew’ as enemy is not new, neither in France nor in Europe. However, according to the CNCDH Report, discourses reminiscent of conspiracy theories have resurfaced during the Covid-19 pandemic (CNCDH, 2022). The hatred is partly driven by a milieu that situates itself between conspiracism and identitarianism and that prefers to spread its ideas on the internet and social networks (Froio 2017). Super-conspiratorial narratives (Soteras 2019) circulate in these platforms and stigmatise the ‘Jew’ (Schwarz-Friesel 2013). What are the denominative patterns that the milieu uses in France to designate them? To answer this question, a corpus of 90,000 messages from ten Telegram messenger channels emitted between January 2018 and May 2022 was analysed. The given social media channels are particularly characterised by the homogeneity of its users. Approaches from DA and CxG were applied to the corpus in order to find recurring patterns.

**Keywords:** Telegram, discourse analysis, CxG, conspiracy theories

## 1. Introduction

During 2022 French electoral campaign, conspiracies like the great reset, great replacement or Q-Anons super-conspiracies (Soteras, 2019) were discussed in the media, and even by center politicians like Valérie Pécresse<sup>1</sup> (*Les Républicains*). These theories mirror a reactionist discourse that made its way into the political sphere in the last decades in France (Durand and Sindaco, 2015) in which minorities often serve as scapegoats (Giry, 2016). The ‘Jew’ as an enemy, responsible for the decline of French society is still omnipresent (Commission nationale consultative des droits de l’homme, 2022). Politicians and press serve as a vehicle for a discourse that has its origins in an identitarian-conspiratorial milieu. To go to the roots of this discourse, a corpus of 90.000 Telegram messages was exploited to analyze representations of the ‘Jew’ as enemy. Telegram and its channels provide an ideological “huis-clos” in which discourses are articulated by mindlike users fueled by official accounts, so-called press reviews and anonymous channel owners.

## 2. The Corpus

After regularly monitoring 25 ideologically homogeneous Telegram channels linked to the identitarian, the ultra-Catholic and the conspiratorial sphere, ten of them were chosen for the construction of a corpus. To represent the diversity of the Telegram messenger, three press reviews, two individuals with clear names (one male and one female) and five channels of anonymous administrators were selected.

### 2.1 Telegram as source for linguistic analysis

Since the Covid-19 pandemic more and more linguists have been interested in analyzing hate speech (Solopova, Scheffler and Wyatt, 2021; Vergani et. al., 2022), disinformation networks (Willaert et al., 2022) or antisemitism and conspiracy narratives (Steffen et al., 2023) generated on Telegram. The messenger that allows asynchronous and anonymous conversations is seen as a “harbinger for freedom” by e.g., extremist groups (Wijermars & Lokot, 2022). Moreover, ideologically homogenous channels tend to lack counter-discourse and regulation by the channel’s often anonymous administrators. Steffen et. al assume that a negative attitude towards Jews becomes more visible in these public but mostly hidden spaces than in heterogeneous environments like Twitter (Steffen et. al 2023, 1090).

### 2.2 Key figures of the Telegram corpus

The data collection on 17 May 2022 allowed the downloading of a total of 90,023 posts, 77,035 of which contained linguistic signs. The download was made directly from the Telegram desktop app. The dataset was cleaned semi-automatically and transferred into the corpus tool TXM, developed by ENS Lyon (Heiden, Magué, & Pincemin, 2010). The corpus has a total of 4 417 995 words. Three sub-corpora represent the genres of the channels. The three press reviews had a reach of 18,036 subscribers as of 12 July 2022, with the majority subscribed to *Egalité & Réconciliation* (E&R) and *fdesouche*, both of whom come from the national-identitarian milieu<sup>2</sup>. The third press

<sup>1</sup><https://www.radiofrance.fr/franceinter/le-grand-remplacement-de-valerie-pecresse-ne-passe-pas-chez-les-republicains-9466575> (last accessed 04-27-2023).

<sup>2</sup> E&R was created by Alain Soral whose ideology was already subject to an article of Bernard Bruneteau in n° 62/2 of the *revue d’histoire moderne & contemporaine* in 2015 (DOI: 10.3917/rhmc.622.0225.). His discourse was analyzed by Lucy Raymond in n°104 de *Quadern* (DOI: 10.4000/quaderni.2140i.). As for *fdesouche*, his founder declared himself neo-Nazi

according to an article of *Le Monde* from 2017 ([https://www.lemonde.fr/politique/article/2017/04/14/pierre-sautarel-l-apprenti-droitier\\_5111064\\_823448.html](https://www.lemonde.fr/politique/article/2017/04/14/pierre-sautarel-l-apprenti-droitier_5111064_823448.html), last accessed on 04-27-2023).

Channel	Subcorpus	First posted	Subscribers (as of July 12th 2022)	Comments allowed	Number of posts	with signs	Wordcount	Dimensions (%)
Canal Natio	anonymous	26.03.20	7.797	No	5.436	4.201	202.379	4,6
Chroniques	anonymous	09.06.20	1.935	Yes	2.343	1.560	161.945	3,7
kadosh	anonymous	15.08.21	8.032	Yes	6.493	2.279	188.464	4,3
LVC	anonymous	02.12.20	8.483	Yes	28.972	23.125	1.274.568	28,8
Trad. catholique	anonymous	13.09.19	1.905	No	1.833	1.630	268.135	6,1
Female Individ.	individual	18.08.21	1.431	No	512	327	18.322	0,4
Male Individ.	individual	13.03.20	11.419	Yes	1.769	1.290	152.571	3,5
fdesouche	press review	13.11.19	9.599	No	23.999	23.987	1.315.264	29,8
MI	press review	15.01.21	1.749	No	3.060	3.036	185.201	4,2
E & R	press review	28.02.18	6.688	No	15.615	15.600	651.146	14,7
				<b>Total</b>	90.032	77.035	4.417.995	100,0

Table 1: Key Figures of the corpus

review (*media-presse-info*, *MI*) can be linked to identitarian Catholics, the channel having close ties to the far right, ultra-catholic *Civitas* organization. The individuals prove to be the least productive of the channel-administrators. The male emitter, who is supposed to have radicalized himself alongside Alain Soral and Dieudonné emits nearly 80% of these messages. The female, a former member of the FN and convicted to six months of prison for public provocation to racial hatred, has a small channel where she shares links and videos from other channels rather than emitting her own. A third sub-corpus groups together the anonymous channels. The most productive of these is a channel accounts for around 30% of the total of messages in the corpus.

### 3. Approaching the ‘Jew’ as enemy through linguistic recurrences

In the work by Schwarz-Friesel and Reinharz on language and hostility towards Jews in 21st century Germany, the authors argue that anti-Semitic linguistic structures constitute and transmit mental models into the collective communicative memory, in which Jews are conceptually represented as the “other” (2013, 6). According to their book the naming of Jews which has been subject to negative amalgams for decades creates an image of Jews as enemies or outsiders. The authors provide examples of underspecified paraphrases such as *die “Religionsgemeinschaft, die uns am Wickel hat”* (the religious community that has us wrapped around its finger) and *“die Banker an der Ostküste”* (the bankers on the East Coast) which has become a fixed formula for referring to American Jews (Schwarz-Friesel and Reinharz, 2013, 37). Through the recurrence and transmitted conceptual patterns of these structures, speakers are able to easily identify the very often negative context of such statements.

#### 3.1 Discourse Formulae in French Discourse Analysis

Recurrent patterns in discourse analysis (DA) can be analyzed through the lens of discursive formulaicity (Faye,

1972; Krieg-Planque, 2003). This concept generally refers to any statement with a fixed structure that fits within a discursive dimension, functions as a social reference, and has a polemical aspect (Krieg-Planque, 2009, 63). The formula is the result of the discursive shaping of a lexical-syntactic association “that speaker fashion and take up by investing it with positioning issues and values” (Krieg-Planque, 2010, 104). Its relative fixity allows the formula to be identified through its frequency in public discourse.

#### 3.2 Conventionalized constructions as an approach to linguistic recurrency

Since Filatkina (2018), a construction grammar approach to formulaicity has made its way into historical discourse analysis. Idiomatic or formulaic language is described here as words which develop their meaning only in combination with others. For those patterns to make sense “and allow speakers to achieve their communicative goals, they must necessarily be conventionalized” (Filatkina, 2018, 4). The conventionalization of a construction can for example be studied through the lens of *frame semantics* (Ziem, 2008). According to Ziem, frames make “relatively stable, discursively solidified background knowledge cognitively available” (2013, 232). Ziem’s statement implies that conventionalized cognitive knowledge is manifested through structured turns. He argues that meaning, or a “predication”, is conventionalized if it is frequently used by a community of speakers (Ziem, 2013, 234). Among all possible predications, the most frequently used become *default values* that the speaker memorizes as implicit knowledge (Ziem, 2008, 242).

#### 4. The results: The ‘Jew’ as enemy in identitarian-conspiratorial Telegram channels

Both approaches mentioned above, the discourse formula and the notion of construction and their *default value* were used to explore the corpus at hand. To extract a first set of frequent expressions a list of 63 terms<sup>3</sup> found in the works of Schwarz-Friesel and Reinharz (2013) and Sarfati (1999)<sup>4</sup> were searched in TXM and cooccurrences displayed. Three

<sup>3</sup> juif; youpin; hébreu; hébraïque; sémite; israélite; judéo-; Israël ; Lapid; Netanyahu; Sharon; le peuple ; Torah; Talmud; Rabbin; Pharisien; œil pour œil ; juif éternel; judaïsme; ; judaïsation; judaïser; sabbat; hérésie; hérétique; Rothschild; Rockefeller; Goldman; Soros; lobbyiste/lobbyisme; lobby; sion.\*; complot;

complot.\*; protocole; négation.\*; Kabbale; franc-maçonnerie; hérésie; hérétique; intrigant; intrigue; primitif; brutal; pervers; avide; infidèle; usurier; sale; crochu; Jérusalem; solution finale; Shoah ; Dreyfus; sang ; race; origine; goy/goyim; ashkénaze

<sup>4</sup> The terms were then classified into five groups (heresy,



### 4.1 Recurrent constructions

#### 4.1.1. *Le/la juif/ve + Name*

Predictations of *the Jew XY* in their messaging context within the Telegram corpus

Category	Percentage
Lobby	41%
Heresy	18%
Cruel individual	17%
Complotism	12%
Israel	12%

conspiracies, ‘Jew’ equals state of Israel, cruelty, lobbyism) which were also used as annotation scheme for the analysis of default values. The classification was done to simplify the analysis and to help synthesize the century-old traditions of anti-Jewish stereotypes that are transmitted through predications. Nevertheless, it is not intended to be universally valid.

However, classifying those stereotypes has proved to be difficult as religious and secular stereotypes are intertwined into conspiracy theories<sup>5</sup> (Soteras, 2019).

[French original] Le juif Dennis Prager explique pourquoi les Juifs quittant la religion reste malgré tout religieux en se convertissant à la religion du gauchisme (rendre le monde meilleur mais sans Dieu). (Telegram post from 01-05-2022)

[illegible]

#### 4.1.2. The qualifier *judéo*

Within the framework of this work only one annotation was made. Despite the lack of an inter-annotator agreement, the results underline the complexity of the endeavor of classifying the predications.

<sup>7</sup> CQL-Query in both corpora: [lemma = "judéo.\*"]

*judéosceptique* appears within their channels. Fig. 3 also suggests that press reviews show small interest in the lexeme.

Only E&R uses it, creating the amalgam *judéo-nazi*, a term

The jew as enemy, moreover, is characterised particularly by his influence on the (Western) world, the destruction of a christianized world to be protected and an associated lobbyism for mondialism.

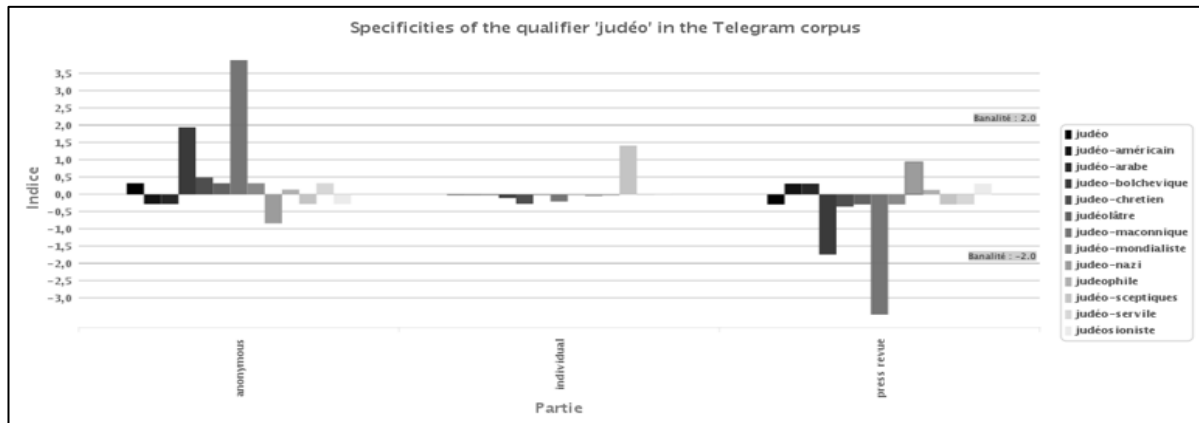


Figure 3: Repartition of the qualifier *judéo-* in the Telegram corpus

that appears for the first time in Mai 2022 to designate Ukrainian individuals after the Russian invasion.

[2] *Chutzpah: Judeo-Nazi Zelensky denounces Hitler's Nazism*

[French original] Chutzpah : le judéo-nazi Zelensky dénonce le nazisme hitlérien

(Telegram post from 05-08-2022)

[3] *Judeo-Nazi Mikhail Kavun (Pravy Sector financier) arrested in Russia*

[French original] Arrestation du judéo-nazi Mikhail Kavun (financier de Pravy Sector) en Russie

(Telegram post from 05-15-2022)

The flexibility of the qualifier *judéo-* could be seen as a sign for a high type frequency (Ziem, 2008, 360). The object itself may be cognitively present to the recipient however there is not one single default value but many and as such the lexeme must be specified by explicit predications.

In the Telegram corpus, only *judeo-christian* demonstrates a high rate of explicit predications (85%). The conventionalization of this construction relies on the frequent use of the syntagm in the media and a broader discourse where *judeo-christian* normally refers to the common roots of the European society (Greene, 2021; Jolibert, 2014; Teixeira, 2008). The explicit context of the messages in the Telegram corpus can be explained by its divergent use. In all explicit messages of the corpus the qualifier is linked to the opposite namely to denounce the supposed predominance of the Jewish over the Christian. For all other occurrences of the qualifier *judéo-* the predications tend to be implicit.

Examining both pattern's conventionalization through the lense of their possible explicit and implicit predications brings to light various negative representations that form the mosaic of its meaning in the given milieu. In the corpus at hands, it is above all the entanglement of religious, conspirational and secular prejudices that can be observed.

<sup>8</sup> The French term community presents itself problematic because it evokes a homogenized group of people, without taking into

#### 4.2 *La communauté que vous connaissez bien, a discursive formula ?*

Nevertheless, another expression found in the corpus could have formulaic potential in the sense of DA: *La communauté que vous connaissez bien*. Taken from an interview broadcasted on television with a former French general, in the summer of 2021, this syntagma made its way through various media platforms and into the streets of France during the protests against the vaccine pass. When asked who controls today's media, the General answers, "Well, it's the community you know well", referring to a generalized Jewish community<sup>8</sup>. Used to avoid the denomination 'Jew', the formula circumvents censorship while stigmatizing Jews as string pullers. In June and July, it is frequently used to stigmatize individuals on Telegram (Examples 4 and 3).

[4] *Cross-breeding is great! Axel Kahn, a member of the community you know well, explains why cross-breeding is a good thing.*

[French original] Le métissage, c'est super ! Axel Kahn, membre de la communauté que vous connaissez bien, explique pourquoi le métissage est une bonne chose.

(Telegram post from 07-06-21)

[5] *Is there any way of finding out just how much this Daniel Křetínský is part of the community you know well?*

[French original] Y aurait-il moyen de vérifier à quel point ce Daniel Křetínský fait partie de la communauté que vous connaissez bien ?

(Telegram post from 04-19-22)

It occurs in variations such as "the community we know well/all" or "a certain community we know well". With memes and GIFs, speakers show creativity in their desire to adopt ideological stances. A meme referring to a scene of the movie *Star Wars II* circulates already in August 2021. Anakin, who is transformed into Darth Vader later in the

consideration the fluidity and plurality of religiosity and secularity in Jewish environments (Endelstein 2016).

film, discusses the system of governance of the planets with Princess Padme. This conversation is repeated in the original meme. It alludes to the violent seizure of power that Anakin is planning in the film.



Figure 4: Example of a meme *La communauté que vous connaissez bien* on Telegram

Despite the formula's rare usage in the homogeneous Telegram channels after its peak in June and July 2021, *la communauté que vous connaissez bien* still circulates in computer mediated communication. After a phase where it encountered a lot of counter-discourse in August 2021, the formula shows itself transcendent and able to accommodate many different scenarios on Twitter. In 2022, the expression resurfaces around Kanye West's anti-Semitic comments, the ban of the Russian soccer team at the World Champion ships and xenophobic statements directed toward the LGBTQ community (Fig.5).



Figure 5: Examples of Tweets in 2022 on *la communauté que vous connaissez bien*

[6] *What kind of shitty job is this the more time goes by the more our society regresses what's the point of dressing up as a woman when you're a man another strike of the community you know well* (Fig. 5, left-hand side)

[7] *Why isn't the country of the community you know well excluded?* (Fig. 5, middle)

[8] *The community you know so well really has a long arm.* (Fig. 5, right hand side)

Being taken from an interview that clearly refers to the Jews as “the community who controls the world”, the formula reflects the above-mentioned prejudice and is used to pick up these prejudices in the given contexts within and outside of Telegram. As such it takes part in the representation of the *Jew* as influential and potentially dangerous *other* to a “Christian” society.

## 5. Conclusion

In tracing the identitarian discourse about Jews, we have come across some forms of denomination that help to grasp the overall picture of the *Jew* as ‘enemy’. For this purpose, CxG methods as well as traditional discourse analysis could help to gain an impression of the functioning of semantic and conceptual discours pattern and how they spread in the Telegram corpus and beyond. However, especially in social media such as Telegram, but also on Twitter chains of messages can be polygonal (Longhi, 2020). Thus, attention must be paid to hyperlinks and technographics (Paveau, 2017) like memes and gifs or stickers, because these play a major role in giving sens to a certain structure and more precisely in the conventionalization of linguistic expressions. Telegram and the homogenous character of the chosen channels seem to display underlying structures used in the milieu discussed in this paper. The mostly hidden but still accessible channels give the emitters the impression of being unobserved, among like-minded people and protected from the censorship of the so called mainstream. Even though channel administrators alert members about algorithms that detect harmful speech, people seem to tend to express more unconventional opinions that seem less accepted in the wider society. The corpus at hands could also be used to examine this further by looking into representations of other enemies like muslims or comunists. A comparison with more heterogenous social media environments such as Twitter or other public online discourse such as comments to newspaper articles related to Jewishness and Israel would give more insight about the representation of the jew as enemy. Furthermore, the intersection of methods from two linguistic traditions proves productive and complementary to approach the complex, and ambivalent concept that is enmity.

## 6. Bibliographical References

- Commission nationale consultative des droits de l'homme (CNCDDH). (2022). *La lutte contre le racisme, l'antisémitisme et la xénophobie: année 2021 : [rapport présenté à Monsieur le premier ministre]*. La documentation Française.
- Durand, P., and Sindaco, S. (2015). *Le discours « néo-réactionnaire »: transgressions conservatrices*. Paris: CNRS éditions.
- Endelstein, L. (2016). Religion et communautés juives : Pour une approche spatiale de la diversité communautaire. *L'Information géographique*, 80(1), pp.14-21.
- Faye, J.-P. (1972). *Théorie du récit: introduction aux « Langages totalitaires »; la raison critique de narrative l'économie*. Collection Savoir. Paris: Hermann.
- Filatkina, N. (2018). *Historische formelhafte Sprache*. De Gruyter.

- Giry, J. (2016). Le conspirationnisme. Archéologie et morphologie d'un mythe politique. *Diogenes*, 249-250 (1), pp.40--50.
- Greene, T. (2021). Judeo-Christian Civilizationism: Challenging Common European Foreign Policy in the Israeli-Palestinian Arena. *Mediterranean Politics* 26 (4): 430--50.
- Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM: Une plateforme logicielle open-source pour la textométrie - conception et développement. In *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data*. Rome, pp. 1021-32.
- Jolibert, B. (2014). Que Peut-on Entendre Par « morale Judéo-Chrétienne » ?. *L'enseignement Philosophique* 64e Année (1): 54--73.
- Krieg-Planque, A. (2003). « Purification ethnique »: Une formule et son histoire. CNRS Éditions.
- . (2009). *La notion de formule en analyse du discours: cadre théorique et méthodologique*. Besançon: Presses universitaires de Franche-Comté.
- Longhi, J. (2020). Les usages stratégiques du commentaire sur Twitter comme contributions aux processus d'idéologisation. *Repères-Dorif*, 22, online.
- Paveau, M.-A. (2017). *L'analyse du discours numérique : Dictionnaire des formes et des pratiques*. Paris: Hermann.
- Sarfati, G.-E. (1999). *Discours ordinaires et identités juives : La représentation des juifs et du judaïsme dans les dictionnaires et les encyclopédies de langue française, du Moyen Age au XXe siècle*. Paris: Berg.
- Schwarz-Friesel, M., Reinhartz, J. (2013). *Die Sprache der Judenfeindschaft im 21. Jahrhundert*. Berlin, Boston: De Gruyter.
- Solopova, V., Scheffler, T. and Popa-Wyatt, M. (2021). A Telegram Corpus for Hate Speech, Offensive Language, and Online Harm. In *Journal of Open Humanities Data* 7, 9.
- Soteras, E. (2019). Les enjeux politico-religieux du conspirationnisme à l'ère postmoderne. *Sociétés*, 142 (4), pp.7--18.
- Steffen, E., Mihaljevic, H., Pustet M., Bischoff N., Do Mar Castro Varela, M., Bayramoglu, Y. and Oghalai, B. (2023). Codes, Patterns and Shapes of Contemporary Online Antisemitism and Conspiracy Narratives – an Annotation Guide and Labeled German-Language Dataset in the Context of COVID-19. In *Proceedings of the International AAAI Conference on Web and Social Media* 17, pp.1082--92.
- Teixidor, J. 2008. Judaïsme et Christianisme et Non Pas « judéo-Christianisme ». *Cités*, 34: 43--52.
- Wijermars, Mariëlle, and Tetyana Lokot. (2022). Is Telegram a “Harbinger of Freedom”? The Performance, Practices, and Perception of Platforms as Political Actors in Authoritarian States. *Post-Soviet Affairs* 38 (1–2), pp. 125--45.
- Vergani, M., Martinez Arranz, A., Scrivens, R., and Orellana, L. (2022). Hate Speech in a Telegram Conspiracy Channel During the First Year of the COVID-19 Pandemic. *Social Media + Society* 8 (4): 20563051221138758.
- Ziem, A. (2008). *Frames und sprachliches Wissen: kognitive Aspekte der semantischen Kompetenz*. Berlin: De Gruyter.
- . (2013). Wozu Kognitive Semantik? In D. Busse & W. Teubert (Eds.), *Linguistische Diskursanalyse: neue Perspektiven*. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 217--240.

# Studying Socially Unacceptable Discourse Classification (SUD) through different eyes: "Are we on the same page ? "

**Bruno Machado Carneiro, Michele Linardi, Julien Longhi**

ENSEA Engineering School, ETIS UMR-8051 CY Cergy Paris Université, AGORA CY Cergy Paris Université  
bruno.machadocarneiro@ensea.fr, {michele.linardi, julien.longhi}@cyu.fr

## Abstract

We study Socially Unacceptable Discourse (SUD) characterization and detection in online text. We first build and present a novel corpus that contains a large variety of manually annotated texts from different online sources used so far in state-of-the-art Machine learning (ML) SUD detection solutions. This global context allows us to test the generalization ability of SUD classifiers that acquire knowledge around the same SUD categories, but from different contexts. From this perspective, we can analyze how (possibly) different annotation modalities influence SUD learning by discussing open challenges and open research directions. We also provide several data insights which can support domain experts in the annotation task.

**Keywords:** SUD Classification, Machine Learning, Deep Learning, Transfer Learning, Annotation Guidelines

## Acknowledgment

The work presented in this paper is part of the ARENAS project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No:101094731.

## 1. Introduction

During these last two decades, the massive popularisation of social media has been changing the way people communicate, interact and collect worldwide news. The dissemination speed rate and the possibility to quickly reach a large audience are some clear advantages of modern social network platforms. By contrast, the potential anonymity and sense of impunity can bring out the worst in people and made them sharing ideas that would not be socially acceptable otherwise. Socially Unacceptable Discourse (Sulc and de Maiti, 2020) (SUD) typically occur in various form; The use of offensive and abusive language represent a common form of SUD, but it is also important to note that controversial narratives are not necessarily bad or immoral, but they closely relate to radicalization and ideologies. Clear contexts in the recent history are the Covid-19 crisis and the the Russian invasion of Ukraine. During these periods, we have witnessed several cases of public debate radicalization, especially favored by the circulation of distorted information (De Giorgio et al., 2022) that jeopardizes the knowledge acquisition of complex systems and environments.

Another particular trait of SUD is the presence of distinctive grammatical characteristics. To model these features, we require identifying several grammatical substructures such as residual representations, use of pronouns, and future tense (Ascone and Longhi, 2018; de Maiti et al., 2020). We note that, in publicly annotated corpus used so far by the Machine Learning community, no standard or common guidelines for SUD annotation exist (Fišer et al., 2017) despite the adoption of the same terminology and/or tags. It derives that different SUD definitions may potentially share overlapping characteristics, or on the other hand a single category may cover text instances with divergent features depending on the context. Furthermore, annotators bias can also play a decisive role as reported by previous

works (Badjatiya et al., 2019; Yuan et al., 2022a; Davidson et al., 2019).

In this scenario, it is reasonable to expect a poor generalization capability of ML SUD classifiers trained in a specific context (Yuan and Rizoiu, 2022). To that extent, we study and evaluate the capability of current state-of-the-art Deep Learning models to characterize SUD on different grounds. Other works have recently considered the zero-shot learning problem in hate speech detection, where transfer learning is tested and measured on binary (hate/no hate) (Torman et al., 2022) and on multi-class (Yuan and Rizoiu, 2022) classification. In this context, we sketch and propose a different approach that first aims to test transfer learning at a class level rather than a dataset level. This approach permits us to provide more interpretable insights on the SUD semantic and to test the transfer over different annotation guidelines on the same speech categories.

## 2. Socially Unacceptable Discourse Corpora

We report the corpora we consider in our study in table 1. We use data from various sources recently adopted to assess the performance of state-of-the-art ML solutions for automatic SUD detection (e.g., hate speech detection, sentiment, toxicity, radicalization, and ideology analysis).

We selected **13** publicly available datasets containing **470,768** samples distributed over 12 classes.

We generate a unique English text corpus by concatenating all the 13 datasets, denoting it with the label  $G^{SUD}$ . Note that the datasets we concatenate in  $G^{SUD}$  share multiple overlapping SUD labels, which identify the same SUD category. We consider the presence of bias and ambiguities as physiological, and identifying and analyse the concerned instances is under the lens of our research.

In figure 1(a), we report the instances distribution over SUD classes. Note that the *neither* class subsumes all texts that do not fall in any SUD categorizations proposed by the annotators. As expected, SUD classes have a sensitive lower support compared to the *neither* class denoting the typical class imbalance setting of the SUD detection problem.

Figure 1(b) illustrates the ratio of each dataset with respect to the global corpus. We observe that Jigsaw and Founta contain together more than 60% of the data.

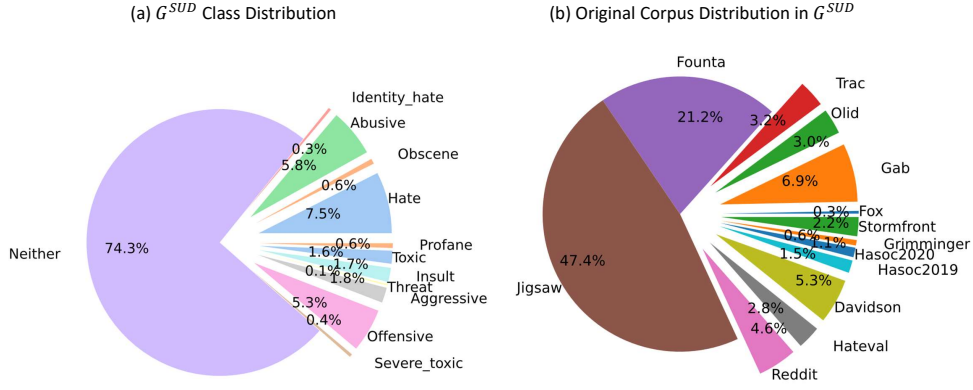


Figure 1: (a)  $G^{SUD}$  Class distribution, (b) Corpus distribution in  $G^{SUD}$

Dataset	Sample type	# Samples	Topic	Best performing SUD classifier	F1 Macro (%)
Davidson (Grimminger and Klinger, 2021)	Tweets	25,000	Generic	BERT	93
Founta (Swamy et al., 2019)	Tweets	100,000	Generic	BERT	69.6
Fox (Yuan and Rizoju, 2022)	Threads	1,528	Fox News Posts	BERT	65
Gab (Qian et al., 2019)	Posts	34,000	Generic	CNN	89.6
Grimminger (Grimminger and Klinger, 2021)	Tweets	3,000	US Presidential Election	BERT	74
HASOC2019 (Wang et al., 2019)	Facebook, Twitter posts	12,000	Generic	LSTM + Attention	78.8
HASOC2020 (Roy et al., 2021)	Facebook posts	12,000	Generic	XLNet-RoBERTa	90.3
Hateval (MacAvaney et al., 2019)	Tweets	13,000	Misogynist and Racist content	mSVM/BERT	75.4
Jigsaw (van Aken et al., 2018)	Wikipedia talk pages	220,000	Generic	Bi-GRU + Attention	78.3
Olid (Zampieri et al., 2019)	Tweets	14,000	Generic	CNN	80
Reddit (Yuan and Rizoju, 2022)	Posts	22,000	Toxic subjects	BERT	85
Stormfront (MacAvaney et al., 2019)	Threads	10,500	White Supremacy Forum	BERT	80.3
Trac (Aroyehun and Gelbukh, 2018)	Facebook posts	15,000	Generic	LSTM	64

Table 1: Best performing SUD classification model on each dataset.

## 2.1. Datasets

Here, we provide the details of each dataset we join in  $G^{SUD}$ .

**Davidson** (Davidson et al., 2017) contains around 25,000 tweets labelled as being hateful, offensive or neither of those randomly sampled from a set of 85.4 million tweets produced by 33,458 different users. Each sample was labelled by at least three different annotators.

**Founta** (Founta et al., 2018) contains about 100,000 tweets, labeled with four categories: abusive, hateful, normal, and spam. In this dataset, a variable number of users (between five and ten) have annotated each sample.

**Fox** (Gao and Huang, 2018) contains 1528 comments posted on ten different popular threads on the Fox News website. In these data, two native English speakers have produced labels to differentiate hateful from normal content following the same annotation guidelines.

**Gab** (Qian et al., 2019) contains 34,000 samples extracted from Gab, a social media, where users commonly share far-right ideologies (Jasser et al., 2021), annotated in the Amazon Mechanical Turk<sup>1</sup> platform, where at least 3 annotators provided a label for each sample.

**Grimminger** (Grimminger and Klinger, 2021) contains 3,000 tweets on 2020 presidential election topic in the United States. The samples were labelled between hate speech or not by three undergraduate students, who discussed the annotation guidelines during the labelling process.

**HASOC2019** (Modha et al., 2019) and **HASOC2020** (Mandl et al., 2020) are datasets proposed

in the Indo-European Languages (HASOC) challenge, which contain 12,000 English text samples extracted from Twitter and Facebook labeled between hateful, offensive, profane or neither of those.

**Hateval** (Basile et al., 2019) gathers around 13,000 tweets containing hateful and normal speech. The hateful content originates from accounts of potential victims of misogyny and racism.

**Jigsaw**<sup>2</sup> (van Aken et al., 2018) is a dataset provided in the Toxic Comment Classification Challenge. It contains about 220,000 samples extracted from Wikipedia talk pages differentiated into seven classes: toxic, severe toxic, obscene, threat, insult, identity hate, and neither of the previous.

**Olid** (Zampieri et al., 2019) contains 14,000 tweets annotated using the Figure Eight Data Labelling platform<sup>3</sup>. In this context, tweet selection is executed by keyword filtering and human annotation.

**Reddit** (Qian et al., 2019) has 22,000 samples extracted from Reddit, labeled for hate speech detection by Amazon Mechanical Turk users. Before the labeling task, the text got selected according to a list of toxic subjects on the Reddit platform.

**Stormfront** (de Gibert et al., 2018) contains 10,500 samples taken from a white supremacy forum called Stormfront and divided into four classes: hate, no hate, related, and skip. The related class contains statements that can not be considered hateful without considering their context. Text belonging to the skip class does not contain enough information to determine if it can be classified as hateful.

<sup>2</sup>[https://www.kaggle.com/c/](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge)

[jigsaw-toxic-comment-classification-challenge](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge)

<sup>3</sup><https://f8federal.com/>

<sup>1</sup><https://www.mturk.com/>

**Trac** (Kumar et al., 2018) dataset gathers 15,000 Facebook posts and comments classified into aggressive and non-aggressive language.

### 3. SUD Deep Learning Models

In this section, we introduce and describe the state-of-the-art Deep Learning models adopted for the SUD detection task in previous works. In Table 1, we show the best performer in each corpus. Here, we report the Macro F1 score, which is the recommended averaging method for F1 score when dealing with class imbalance. It is calculated by averaging the sum of the F1 score of each class.

Recall that the F1 score reports the harmonic mean of precision and recall of a classification model. For a particular input class, we compute the precision (P) and recall (R) of a SUD classifier as follows:  $P = \frac{TP}{TP+FP}$ , and  $R = \frac{TP}{TP+FN}$ , where TP denotes the number of correctly classified instances of the input class (true positive), FP denotes the number of occurrences that are wrongly assigned with the input class label (false positive), and FN represents the number the input class samples that are erroneously classified (false negative). Hence we have that  $F1 = 2 \times \frac{P \times R}{P+R}$ .

From Table 1, we observe that **BERT** (Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)) is the best performer model in the majority of the datasets. BERT adopts a Deep Learning (DL) architecture released by the Google AI Language team in early 2019, which is pre-trained by masked language model (MLM) and next sentence prediction (NSP) tasks over a large corpus of English data containing more than 3B words (Devlin et al., 2019). MLM consists of training the model to predict masked tokens in the corpus sentences, whereas the NSP training aims to predict if two sentences form a sequence in the original text. XLM-RoBERTa (Conneau et al., 2020) is a multilingual variant of the original BERT model.

BERT has clearly shown its superiority over other types of DL models previously adopted in SUD classification, such as Convolutional Neural Networks (CNN) (Qian et al., 2019) and Long-short term memory networks LSTM (Wang et al., 2019). The attention mechanism used by BERT represents a robust solution that can better learn long-range token dependencies, avoiding the limitation of LSTM networks, which assumes that each token depends only on previous ones. By contrast, BERT learns relationships considering all the tokens in a sentence simultaneously.

In this work, we evaluate the SUD classification performance of BERT in the heterogenous corpus we construct. In the next section, we present all the research questions we address, discussing the results we obtain.

## 4. Experiments

### 4.1. Multiclass SUD Classification

To conduct our experimental evaluation, we use the BERT<sub>BASE</sub> (Devlin et al., 2019; Yuan and Rizoio, 2022) model pre-trained by WordPiece tokenizer algorithm. For the sake of reproducibility, we provide the code and the data used in the experiments along with the relative instructions in an online repository (Machado Carneiro et al., 2023).

Training set	F1 Score (%)		
	Macro	Weighted	Micro
$G^{SUD}$	53.9	86.8	87.1
$G^{SUD}$ Balanced	51.3	85	84.5
$G^{SUD}$ with Neither Undersampled	58.5	73.7	73.9
$G^{SUD}$ balanced with Neither Undersampled	56.8	72.5	72.1
$G^{SUD}$ (Binary classification)	88.5	91.3	91.2
$G^{SUD}$ balanced (Binary classification)	89.7	89.7	89.7

Table 2: Comparison between all experiments

To perform SUD classification, we connect BERT pooled output layers to a Multi-Layer Perceptron (MLP) architecture that contains 12 output neurons (one per class). We have fine-tuned the MLP layer of proposed model on the  $G^{SUD}$  corpus using a 80%/10%/10% splitting ratio for training, validation, and testing respectively. We have adopted a stratified sampling technique to keep the same class distribution throughout the three splits. Hyperparameters have been tuned by performing several complete training rounds, picking the setting with the best validation performance.

The research questions we want to address are the following: *Which are the state-of-the-art model generalization capability in a global context? What are the main challenges that hamper the SUD modelling effectiveness?*

Table 2 contains the results, where we report Macro, Weighted and Micro F1 score of the SUD classification. Note that the Weighted F1 weighs the global F1 average according each class support, whereas the Micro F1 score computes a global F1 making no distinction across classes. Considering that  $G^{SUD}$  contains highly unbalanced SUD classes, we repeat classification tasks after training our model on a balanced dataset. To that extent, we have performed random oversampling of minority classes as suggested by several works (Yuan and Rizoio, 2022; Swamy et al., 2019; MacAvaney et al., 2019).

Furthermore, given the dominance of the *neither* class, we also consider a setting with under-sampled non-SUD text (*neither* class). Here, we have selected 10% of the non-SUD samples in a stratified way, maintaining the same proportion of the *neither* class samples in every dataset.

We note that undersampling the *neither* class has a sensitive effect on the model prediction capability as the Macro F1 score increases. On the other end, reducing the neutral class causes an increment of model errors for the *neither* class (majority class) as we observe a significant reduction of the Weighted and Micro F1 scores. It follows that coping with such an imbalance between non-SUD and SUD samples represents a concrete challenge (typically occurring in a real-world scenario), which is amplified in the extended corpus under consideration.

We also notice that producing a balanced class scenario by performing random oversampling does not provide any significant benefit. This suggests that class imbalance is only a joint cause of the model discrimination capability.

To better understand how the adopted model discriminates SUD classes, we visualize the generated text representation (output of BERT output pooled layer). To reduce the dimensionality of the latent space, we apply t-distributed Stochastic Neighbor Embedding (t-SNE). Figure 2 shows the plot computed over the testing set, with a model trained



	Macro F1 Score (%)											
	Abusive	Aggressive	Hate	Identity Hate	Insult	Neither	Obscene	Offensive	Profane	Severe Toxic	Threat	Toxic
$G^{SUD}$	79.4	64.1	65.8	35.9	50	94.3	25.6	74.9	30.5	39.5	42.6	17.7
Davidson	-	-	41.4	-	-	88.5	-	89.2	-	-	-	-
Founta	81.7	-	33.2	-	-	95.5	-	-	-	-	-	-
Fox	-	-	13	-	-	82.6	-	-	-	-	-	-
Gab	-	-	86.4	-	-	88.6	-	-	-	-	-	-
Grimminger	-	-	10.8	-	-	93	-	-	-	-	-	-
HASOC2019	-	-	7.94	-	-	78.1	-	25	20.4	-	-	-
HASOC2020	-	-	6.67	-	-	91.1	-	29.7	39.1	-	-	-
Hateval	-	-	53.2	-	-	73.9	-	-	-	-	-	-
Jigsaw	-	-	-	37.9	53.1	97.5	26.9	-	-	40.4	46	18.1
Olid	-	-	-	-	-	85.8	-	45.3	-	-	-	-
Reddit	-	-	74	-	-	89.5	-	-	-	-	-	-
Stormfront	-	-	39.7	-	-	94.1	-	-	-	-	-	-
Trac	-	68.1	-	-	-	66.1	-	-	-	-	-	-

Table 3: Macro F1 Score of SUD classification per class and dataset.

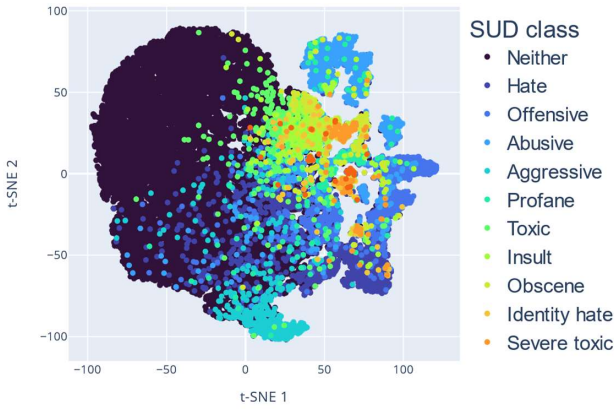


Figure 2: Two components t-SNE visualization of samples embedding produced by BERT output pooled layer.

on the complete corpus  $G^{SUD}$ . In Table 3 we report the Macro F1 score of SUD classification in  $G^{SUD}$  for each dataset and each class. Note that each line in this table corresponds to a different model, trained only on the specified dataset, while the first line is the result obtained using the model trained on  $G^{SUD}$ .

Here, we observe that some class features, i.e., *Abusive* (top-right), *Aggressive* (bottom-center) form fairly clear clusters. We can expect this behavior as each one of these class labels solely occurs in a single dataset, as depicted in Table 3.

Some other classes, i.e., *Hate*, *Offensive*, and *Toxic*, have more sparse values, which is one reason behind the absolutely low F1 score. Once again, these results get confirmed by the absolute low Macro F1 score both in the global corpus and in each single dataset.

Overall, the results explain the poor generalization capabilities of the studied classification model as this latter attains a low Macro F1 (58%) score on  $G^{SUD}$ . In detail, we note that problematic classes are not only those with the lowest number of training samples as one might expect. In fact, a performance drop occurs in  $G^{SUD}$  classes that share samples from multiple corpus, suggesting the presence of intraclass heterogeneous samples as depicted in Table 3.

In this sense, a clear example concerns the *hate* class that

contains samples from ten different datasets (out of thirteen). We note that shaky classification performance in each dataset of  $G^{SUD}$  (see Table 3) depends on divergent annotation criteria on a sensibly general concept, which can relate to different textual elements.

In Table 4, we depict the classification results obtained for each dataset in the global corpus  $G^{SUD}$ , and when the model was trained only using a single dataset (Individual). We note that only in two cases the global model performs better than the individual counterpart (for the Fox and Grimminger datasets). We believe that the relatively small support of these two corpora is the reason behind this improvement. Nevertheless, leveraging more knowledge from multiple domains does not constitute an advantage in practice.

Dataset	Macro F1 Score (%)		
	(a) Multiclass SUD Classification		(b) Binary Classification
	Classified in $G^{SUD}$	Individual	Classified in $G^{SUD}$
$G^{SUD}$	53.9	-	88.5
Davidson	73	75.1	93.9
Founta	70.1	74.7	92.9
Fox	<b>47.8</b>	41.6	59.2
Gab	87.5	89.9	86.2
Grimminger	<b>51.9</b>	46.9	64
HASOC2019	32.9	40.8	64.5
HASOC2020	41.7	48.4	88.2
Hateval	63.6	75.7	70.2
Jigsaw	45.7	52.6	87.7
Olid	65.6	75.2	72.3
Reddit	81.7	82.9	79.9
Stormfront	66.9	76.1	71.1
Trac	67.1	73.1	69.3

Table 4: (a) **Multiclass** SUD classification results (F1 score) with the model trained in  $G^{SUD}$  VS on each single dataset. (a) **Binary** SUD classification with the model trained in  $G^{SUD}$ .

## 4.2. Binary SUD Classification

For each of the experiments reported in this section, we have also tested the capability of the model to discriminate SUD and non-SUD text in  $G^{SUD}$  irrespective of the specific class. To that extent, we use the same configuration for the classification head, changing the output layer to perform binary classification and re-training the model. For this case, we obtain a relatively high Macro F1 score ( $\sim 90\%$ ). Such results suggest how the model discriminates well the *neither* class from the generic SUD in the



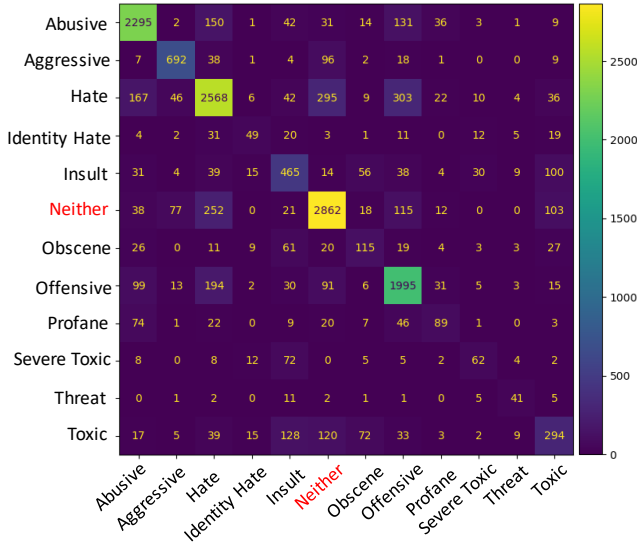


Figure 3: Confusion matrix of multi-class SUD classification.

global context we built, confirming the current trend observed in the ML literature so far. At the same time, effectively modeling multi-class SUD remains an open challenge.

## 5. Further Discussion and Perspective

To closely analyze the state-of-the-art limitation on SUD modeling, in figure 3, we plot the confusion matrix computed on the test set. In this case, we consider a test corpus with undersampled instances of neither class since, for this case, the classification model performs (slightly) in the best manner. Here, we can observe multiple critical cases that concern the labels *Identity Hate*, *Toxic*, *Obscene* and *Profane*. The classification model assigns a random label to these four classes that have overlapping features with all the others. Concerning classification performance, we note that the F1 score is not significantly dropping for these classes when the model applies to  $G^{SUD}$ . It derives that learned features are fairly conserved in the new global context.

This observation confirms the results proposed by prior studies (Yuan et al., 2022b; Fortuna et al., 2020), which already analyzed the relation among several classes in significantly smaller corpora.

We believe the large-scale scenario we propose motivates the need for a more consistent effort in the ML community to equip language models with more discriminant power. This concerns the capability to distinguish the source and the target of the SUD discourse (individual rather than group), as well as the elements that characterize the kind of narrative of each SUD class.

## 6. Conclusion and Future Work

In this work, we present an empirical evaluation of automatic SUD detection using the BERT model, a state-of-the-art Deep Learning architecture for SUD classification. To test generalization capability, we consider a large and heterogeneous context in which we obtain results that are not in line with the expected performance of the model trained

at the local level, i.e., in every single corpus. In this sense, we argue that to build more general and reliable models, the ML community should consider formal guidelines provided by language experts (mostly neglected so far), which can sensibly reduce local bias (e.g., annotation policy, context, etc.). In future work, we plan to closely analyze the inter-domain mismatches we observe at the class sample level. Such effort would be beneficial to understand how to improve textual feature learning and to communicate requirements and expectations from the annotation task.

We furthermore note that the results and the insights we obtained also have the potential for the research linguists, discourse analysis, or semantics, as they show, from a knowledge base constituted by the main works on SUD corpora, the semantic links, and conceptual relationships, between several labels or tags.

In fact, over and above terminology, it is crucial to clearly state and understand the specific features of hate speech, offensive speech, or extremist speech. These initial results are necessary to foster several research discussions in the Horizon Europe ARENAS project into which this work integrates.

Specifically, the semantic issues in discourse categorization have an impact not only in terminological and computational terms (for annotating and classifying) but also in legal, political, and sociological terms. The impact of different characterizations is not neutral, there are potential issues of moderation or condemnation (Longhi, 2021), and it is necessary to proceed cautiously and rigorously in the delimitation of the chosen descriptors and in the way they are defined and characterized.

Finally, the explicability of these categories and the classification provided by Artificial Intelligence is central to future research. Making transparent outcomes will enable us to propose valuable results for all those involved in the hate speech and extremism analysis. In the context of a multi-disciplinary project like ARENAS, which brings together scientists with different backgrounds (i.e., linguists, political scientists, etc.) and targets a heterogeneous audience, such as lawyers and journalists, the clarity of descriptors, and their ability to be understood by different stakeholders, is an essential element.

## 7. References

- Aroyehun, S. T. and Gelbukh, A. (2018). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *TRAC-2018*.
- Ascone, L. and Longhi, J. (2018). The expression of threat in jihadist propaganda. *Fragmentum*.
- Badjatiya, P., Gupta, M., and Varma, V. (2019). Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*.
- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, June.

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL 2020, Online, July 5-10, 2020*.
- Davidson, T., Warmusley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language.
- Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*.
- de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. In *ALW2*, October.
- De Giorgio, A., Kuvašić, G., Maleš, D., Vecchio, I., Tornali, C., Ishac, W., Ramaci, T., Barattucci, M., and Milavić, B. (2022). Willingness to receive covid-19 booster vaccine: Associations between green-pass, social media information, anti-vax beliefs, and emotional balance. *Vaccines*, 10.
- de Maiti, K. P., Fišer, D., and Erjavec, T. (2020). Grammatical footprint of socially unacceptable facebook comments. In *Language Technologies & Digital Humanities*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. <https://doi.org/10.48550/arXiv.1810.04805>.
- Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*.
- Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *International Conference on Language Resources and Evaluation*.
- Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. <https://doi.org/10.48550/arXiv.1802.00393>.
- Gao, L. and Huang, R. (2018). Detecting online hate speech using context aware models. <https://doi.org/10.48550/arXiv.1710.07395>.
- Grimminger, L. and Klinger, R. (2021). Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, April.
- Jasser, G., McSwiney, J., Pertwee, E., and Zannettou, S. (2021). Welcome to #gabfam: Far-right virtual community on gab. *New Media & Society*.
- Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *(LREC 2018)*.
- Longhi, J. (2021). Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International*, 318:110564.
- MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*.
- Machado Carneiro, B., Linardi, M., and Longhi, J. (2023). [https://github.com/mlinardiCYU/SUD\\_study\\_different\\_eyes.git](https://github.com/mlinardiCYU/SUD_study_different_eyes.git).
- Mandl, T., Modhab, S., Shahic, G. K., Jaiswald, A. K., Nandinie, D., Patelf, D., Majumder, P., and Schäfera, J. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages. <https://ceur-ws.org/Vol-2826/T2-1.pdf>.
- Modha, S., Mandl, T., Majumder, P., and Pate, D. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages.
- Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech.
- Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021). Leveraging multilingual transformers for hate speech detection.
- Sulc, A. and de Maiti, K. P. (2020). No room for hate: What research about hate speech taught us about collaboration? In *TwinTalks@DH/DHN*.
- Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.
- Toraman, C., Şahinuç, F., and Yilmaz, E. H. (2022). Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Language Resources and Evaluation Conference*.
- van Aken, B., Risch, J., Krestel, R., and Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis.
- Wang, B., Ding, Y., Liu, S., and Zhou, X. (2019). Ynu\_wb at hasoc 2019: Ordered neurons lstm with attention for identifying hate speech and offensive language. In *Fire*.
- Yuan, L. and Rizoio, M.-A. (2022). Detect hate speech in unseen domains using multi-task learning: A case study of political public figures.
- Yuan, L., Wang, T., Ferraro, G., Suominen, H., and Rizoio, M.-A. (2022a). Transfer learning for hate speech detection in social media.
- Yuan, S., Maronikolakis, A., and Schütze, H. (2022b). Separating hate speech and offensive language classes via adversarial debiasing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*.
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, June.

# A Pipeline for the Large-Scale Acoustic Analysis of Streamed Content

Steven Coats

English, Faculty of Humanities, University of Oulu, Finland

E-mail: [steven.coats@oulu.fi](mailto:steven.coats@oulu.fi)

## Abstract

Vast quantities of audio and video data are available from video sharing sites, streaming services, and social media platforms, but relatively little of this content has been utilized for acoustic, phonetic, or multimodal analysis of linguistic variation. This article describes a Python-based scripting pipeline for the extraction and analysis of audio from YouTube and other websites that use common streaming protocols. The pipeline comprises elements from the Python libraries yt-dlp and Parselmouth and uses the Montreal Forced Aligner for aligning audio with text. The scripts are customizable and suitable for the automatic extraction of video as well as audio and transcript data. An exploratory proof-of-concept analysis considers the first target of the /eɪ/ diphthong in American English: Starting from videos indexed in the Corpus of North American Spoken English, almost 9 million tokens of the segment were retrieved using the pipeline and their values in F1/F2 formant space mapped. As expected, the diphthong's first target has a more closed and back starting point for speakers in the American Southeast.

**Keywords:** Corpus linguistics, Phonetics, Formants, YouTube, DASH, CoNASE

## 1. Introduction

The study of linguistic and interactive properties of Computer-mediated communication (CMC) has historically been focused primarily on text content such as chat, instant messenger (IM) messages, or text-based posts on social media web platforms. In the past 15 years, however, continual increases in bandwidth availability and refinement of technical protocols have led to the widespread use of images, audio, and streamed video content in CMC, for example on video sharing and streaming sites or in online video meetings. Multimodality, or the concurrent use of text, speech, and video, has become central to CMC on the most widely-used video sharing and social media communication sites such as YouTube, Twitch, or TikTok.

As of 2023, most websites utilize the DASH protocol (Dynamic Adaptive Streaming over HTTP; Sodagar 2011) or the related HLS protocol (HTTP Live Streaming) to serve video, audio, and other content on the web. DASH allows the transmission of video and audio data in various formats and compression levels as well as automatic speech recognition (ASR) or manually-uploaded captioned transcripts of speech, user comments and interactions, and other types of data and metadata to the end user in a web browser.

For the researcher interested not only in text, but also in acoustic, phonetic, or gestural/kinesic properties of communication, multimodal content delivered via web streaming represents a valuable source of empirical data. This paper presents a pipeline for accessing streamed audio content on YouTube for phonetic analysis.<sup>1</sup> The pipeline, which is Python-based, makes use of several open-source tools, code libraries, or repositories: yt-dlp<sup>2</sup> for content download of audio, video, and transcript data; the Montreal Forced Aligner<sup>3</sup> for forced alignment of audio and transcript data; and Parselmouth-Praat<sup>4</sup> for identification

and extraction of acoustic features of interest. The pipeline, in a Python Jupyter format, consists of modular script blocks that can be modified and adapted for specific tasks on existing datasets without needing to apply all of the steps in the pipeline. While the example provided in this paper focuses on acoustic properties of audio segments, the pipeline is also suitable for the automated download of corpora of video content.

The rest of the paper is organized as follows: Section 2 provides a brief overview of a few tools used for forced alignment and acoustic analysis of online content, including web-based services. Section 3 details the components used in the pipeline, and Section 4 demonstrates the functionality of the pipeline by providing an exploratory analysis of geographical variation for the first target of the /eɪ/ diphthong in F1/F2 formant space in North American English, starting from videos indexed in the *Corpus of North American Spoken English* (Coats 2023). The exploratory analysis reveals a pattern that corresponds to results of previous research, confirming the potential usefulness of the pipeline. Section 5 summarizes the paper and provides a brief outlook for future developments.

## 2. Previous Work

Phonetic analysis of speech audio requires a transcribed text and a forced alignment of the transcript with the speech signal, permitting the acoustic analysis of words, phonemes, and other segments. Several tools for forced alignment have been built on the Hidden Markov Model Toolkit (HTK, Young 1993)<sup>5</sup> and Kaldi (Povey et al. 2011)<sup>6</sup>: The Penn Forced Aligner (Yuan & Liebermann 2008) and the MAUS aligner (Schiel 1999), for example, are built on HTK, while the Montreal Forced Aligner (McAuliffe et al. 2017) builds upon Kaldi. Other forced alignment tools include Julius (Lee et al. 2009),<sup>7</sup> and SPPAS,<sup>8</sup> developed for French on the basis of Julius but capable of aligning

<sup>1</sup> [https://github.com/stcoats/phonetics\\_pipeline](https://github.com/stcoats/phonetics_pipeline)

<sup>2</sup> <https://github.com/yt-dlp/yt-dlp>

<sup>3</sup> <https://montreal-forced-aligner.readthedocs.io>

<sup>4</sup> <https://github.com/YannickJadoul/Parselmouth>

<sup>5</sup> <https://htk.eng.cam.ac.uk>

<sup>6</sup> <http://kaldi-asr.org>

<sup>7</sup> [http://julius.osdn.jp/en\\_index.php](http://julius.osdn.jp/en_index.php)

<sup>8</sup> <https://sppas.org>

additional languages (Bigi 2015).

Composite tool suites and web-based speech processing platforms have incorporated these aligners into their functionality, making it easier to process audio recordings without having to install and configure the software locally. FAVE-Extract (Forced Alignment and Vowel Extract, Rosenfelder et al. 2011), for example, uses the Penn Forced Aligner, while WebMAUS (Kisler et al. 2017) and DARLA (Dartmouth Linguistic Annotation, Reddy & Stanford 2015), which use MAUS and MFA, respectively, are websites that allow users to upload audio files and transcripts for forced alignment. A recent option in DARLA allows users to generate ASR transcripts from audio files by sending them to Deepgram, a paid service that hosts large neural network speech-to-text models.

Studies have shown that the Penn Forced Aligner and the Montreal Forced Aligner can produce results comparable to those of human annotators. MacKenzie and Turton (2020), for example, used FAVE and DARLA to align samples of speech from six regional British English varieties. Comparing them with alignments produced by human annotators, they found that DARLA performed slightly better than FAVE, but that both tools perform well and produce alignments comparable to those created by human annotators. They remark that “the fact that they have been provided with phonological systems that differ – sometimes rather radically – from the systems they have been trained on has not hindered their performance” (2020: 9), and conclude “our analysis has shown impressive performances from both DARLA and FAVE, and we have full confidence in recommending that researchers who work on non-American and non-standard varieties of English use these tools for forced alignment” (2020: 11). Similarly, Gonzalez et al. (2020) found that the Montreal Forced Aligner generated accurate alignments for recordings of Australian English, even when using an American English model.

Once the audio and transcript have been aligned, acoustic analysis can be undertaken with Praat (Boersma & Weenink 2023) or other software, for example to investigate vowel quality and quantity, pitch, timing and prosody, or other features.

For YouTube, the PEASYV tool (Phonetic Extraction and Alignment of Subtitled YouTube Videos, Méli 2023) provides for individual videos functionality similar to that of the pipeline described in this paper. PEASYV makes use of yt-dlp and aligns transcripts with the Penn Forced Aligner and SPPAS. Source code for the tool, however, is not available, as of mid-2023. Notable is also youglish.com, a service through which users can search YouTube ASR transcripts for specific utterances; links to the utterance in

YouTube videos are returned.<sup>9</sup>

## 4. Pipeline components

The pipeline has been provided as a Jupyter Notebook hosted on GitHub which can be run on the Google Colab service. Due to restrictions on user accounts imposed by the database underlying the Montreal Forced Aligner, using the pipeline on a local or cloud machine may be more efficient than Colab for extensive data collection.

### 4.1 Yt-dlp

Yt-dlp is a fork of YouTube-DL, an open-source library for accessing YouTube or other streamed content. The fork provides some additional functionality, compared to the original library, and can be used to retrieve content not only from YouTube, but from many websites that stream using DASH or HLS protocols, including broadcasters, social media, and content sharing websites.

The yt-dlp component of the pipeline extracts ASR transcripts for video(s) of interest; these are tokenized and then converted to either a format in which the transcript is rendered as a standard text or a format in which each word token has timing information appended in the form `word_1.00`, where the numerical value indicates the time offset in seconds from the start of the corresponding video. SpaCy can be used in the pipeline for part-of-speech tagging.<sup>10</sup> The script works “out of the box” for any of the languages for which YouTube provides ASR captions.<sup>11</sup>

Texts prepared with word timing information in this manner can then be used to extract audio or video content from the corresponding videos, again using yt-dlp. With regular expressions, specific lexical items, word sequences, speech acts, or exchanges can be targeted for audio or video extraction. The pipeline script uses the timing information to retrieve the corresponding audio segment and transcript fragment for a variable-length “window” around the targeted word sequence: for example, if the regular expression targets the sequence “need to”, the window can be set to capture (three words) + “need to” + (three words), resulting in hits such as “then if we need to ask about the”.<sup>12</sup>

### 4.2 Montreal Forced Aligner

The extracted text fragment and its corresponding audio segment are aligned with the Montreal Forced Aligner, using an acoustic model trained on the librispeech dataset (Panayotov et al. 2015). The output is Praat TextGrid files which contain the exact start and end times for the words and phones within the corresponding audio; phones are represented with the ARPA dictionary (Gorman & Howell 2011).

### 4.3 Parselmouth (Praat)

Parselmouth (Jadoul et al. 2018) is a Python port of functions from Praat. In the exploratory analysis in Section

<sup>9</sup> Youglish uses YouTube’s API and automatically-generated metadata to associate individual videos with English varieties (American, British, Australian, etc.). The service provides access to the videos at YouTube’s website, but audio and video content are not available for download and further processing such as forced alignment without using additional tools.

<sup>10</sup> With the `en_core_web_sm` model

(<https://spacy.io/usage/models>).

<sup>11</sup> As of mid-2023, English, Dutch, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish.

<sup>12</sup> Aligning shorter segments prevents ASR or other errors from causing cascading alignment errors in the entire video. A window length from approximately 7–20 words was found to be effective.



5, Parselmouth is used to measure formant frequencies, but the software can be used to investigate other acoustic phenomena pertaining to the speech signal as well, such as pitch, intensity, timing phenomena, stress, or intonation. An advantage of using Parselmouth, compared to standalone Praat, is Python integration: while shell scripts can be used to pass data from Python to Praat, the process can be cumbersome, and integration of Praat functions, via Parselmouth, into common Python development environments such as Jupyter can facilitate analysis and visualization workflows.

## 5. /eɪ/ Nuclei in North America

This section describes an exploratory analysis of regional phonetic variation undertaken using the pipeline.

The Corpus of North American Spoken English (Coats 2023), a 1.3-billion-word corpus of geolocated YouTube ASR transcripts, was used as a starting point for extraction of /eɪ/ diphthongs. A regex script targeted monosyllabic words in CoNASE containing /eɪ/ and extracted a seven-word span of transcript and audio from the corresponding videos. These alignments were used to extract F1 and F2 formant values at nine measurement points during vowel duration for the monophthongs and diphthongs of American English.

Figure 1 demonstrates the results for videos from a single YouTube channel, that of the municipality of Hendersonville, Tennessee. The figure shows the trajectories in formant space for /eɪ/, as well as the diphthongs /aʊ/ and /oʊ/, for 10,745 vowel tokens extracted from 133 videos. Each circle represents a single measurement in F1/F2 space. The size of circles shows the number of measurements at the corresponding duration quantile. The mean trajectories of the diphthongs correspond to line segments joining the centers of the individual measurement points for that diphthong.

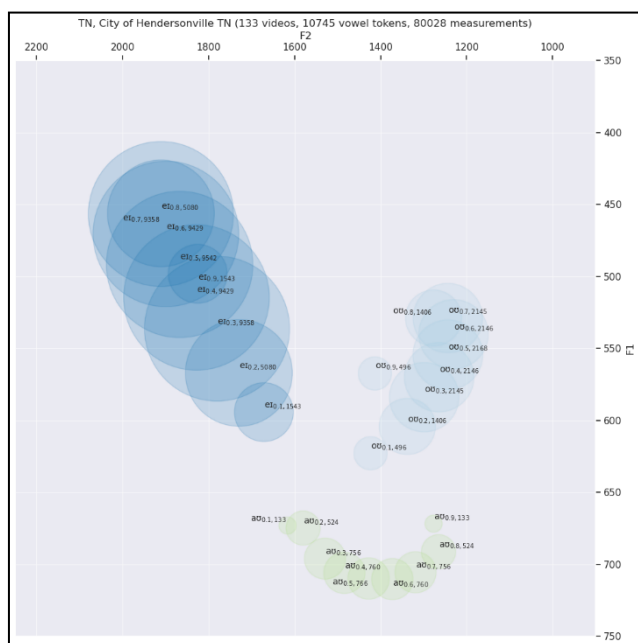


Figure 1: /eɪ/, /aʊ/, and /oʊ/ diphthongs for

Hendersonville, TN

This level of detail allows the analyst to consider characteristic qualities of vowels in different regions or locations.

Integration of the pipeline into Python development environments makes it possible to create interactive visualizations as well. Figure 2 is a screenshot of an interactive visualization of a sample of /eɪ/ diphthongs from another Tennessee locality, the town of Gallatin.<sup>13</sup> Diphthong trajectories for individual tokens are represented as lines; the circles on each line mark the measurements at the corresponding quantile. Users interacting with the plot can click on a line to hear the diphthong; the plot can be used to demonstrate relative closedness and backness of /eɪ/ for many speakers from this locality (and elsewhere in the American Upper South).

From a broader geographical perspective, the formant extraction procedure can provide an overview of variation in the phonemic inventory of American English. Figure 3 shows the Getis-Ord Gi\* value for the F2 value of the first target of the /eɪ/ diphthong, based on almost 9 million vowel tokens. As can be seen, the diphthong nucleus is somewhat more back in the American Southeast, but more front in the upper Midwest, Canada, and Southern California. This pattern largely corresponds to our knowledge of the distribution of formant values for this diphthong (e.g., Labov et al. 2006: 94; Grieve et al. 2013: 49), providing a preliminary confirmation of the validity of the phonetic extraction pipeline.

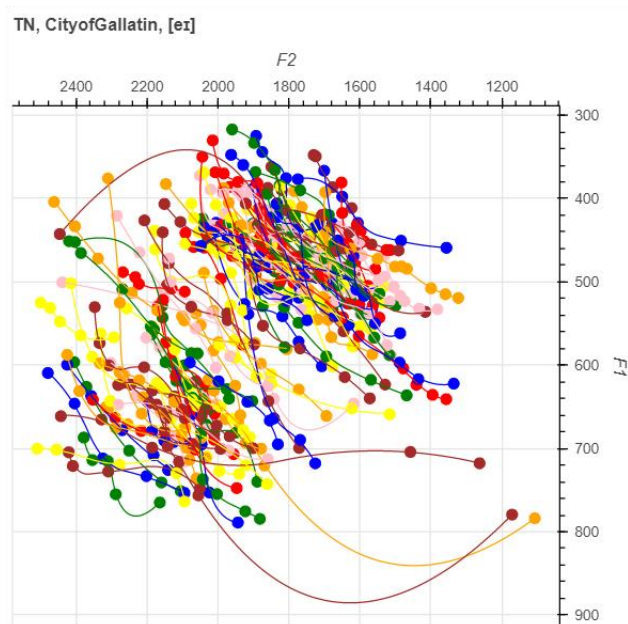


Figure 2: Screenshot of interactive /eɪ/ formant tracks for Gallatin, TN

<sup>13</sup> [https://cc.oulu.fi/~scoats/example\\_Gallatin\\_all.html](https://cc.oulu.fi/~scoats/example_Gallatin_all.html)

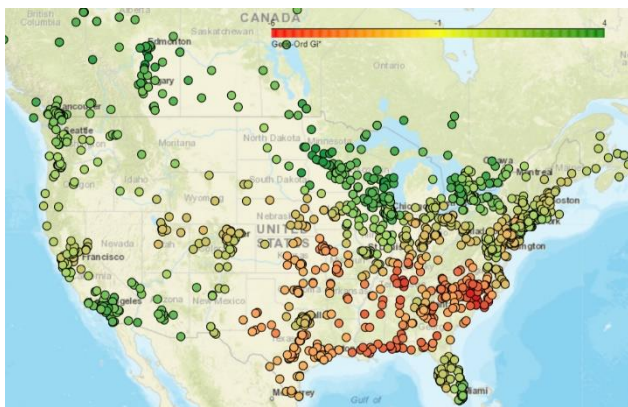


Figure 3: Getis-Ord  $G_i^*$  values for F2 nucleus of /eɪ/ diphthong (8,788,999 tokens)

## 6. Summary and Outlook

The acoustic analysis pipeline utilizes components from ytdlp, the Montreal Forced Aligner, and Parselmouth-Praat, and can be used to harvest transcript and acoustic data from YouTube. Content from other websites that utilize the common streaming protocols can also be harvested, including video data. The pipeline can be used to create custom corpora for acoustic and multimodal analysis, or can serve as the starting point for acoustic analyses of large existing corpora of YouTube transcripts, such as CoNASE or CoBISE (Coats 2023, 2022). The pipeline represents a potentially useful framework for the creation of corpora and the acoustic analysis of naturalistic speech from a range of geographical contexts, content types, and pragmatic situations.

## 7. References

- Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician - International Society of Phonetic Sciences* 111, 54–69.
- Boersma, P. & Weenink, D. (2023). *Praat: doing phonetics by computer* [Computer program]. Version 6.3.09. <http://www.praat.org>
- Coats, S. (2023). Dialect corpora from YouTube. In Beatrix Busse, Nina Dumrukic, and Ingo Kleiber (eds.), *Language and linguistics in a complex world*, 79–102. Berlin: de Gruyter. <https://doi.org/10.1515/9783111017433-005>.
- Coats, S. (2022). The Corpus of British Isles Spoken English (CoBISE): A new resource of contemporary British and Irish speech. In Karl Berglund, Matti La Mela, and Inge Zwart (eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference, Uppsala, Sweden, March 15–18, 2022*, 187–194. Aachen, Germany: CEUR. <http://ceur-ws.org/Vol-3232/paper15.pdf>
- Gonzalez, S., Grama, J. & Travis, C. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard* 5. <https://doi.org/10.1515/lingvan-2019-0058>
- Gorman, K. & Howell, J. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3), 192–193.
- Grieve, J., Speelman, D. & Geeraerts, D. (2013). A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography* 1, 31–51. <https://doi.org/10.1017/jlg.2013.3>
- Jadoul, Y., Thompson, B. & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, 1–15. <https://doi.org/10.1016/j.wocn.2018.07.001>
- Kisler, T., Reichel, U. D. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.
- Labov, W., Ash, S. & Boberg, C. (2006). *The Atlas of North American English*. Berlin: Mouton de Gruyter.
- Lee, A. & Kawahara, T. (2009). Recent development of open-source speech recognition engine Julius. In *Proceedings of APSIPA ASC 2009*, pp. 131–137.
- MacKenzie, L. & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6, no. sl. <https://doi.org/10.1515/lingvan-2018-0061>
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.
- Méli, A. (2023). *PEASYV: Phonetic Extraction and Alignment of Subtitled YouTube Videos*. <https://adrienmeli.xyz/peasyv.html>
- Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. <https://doi.org/10.1109/ICASSP.2015.7178964>
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US*. IEEE Signal Processing Society.
- Reddy, S. & Stanford, J. (2015). A Web Application for Automated Dialect Analysis. In *Proceedings of NAACL-HLT 2015*.
- Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H. & Yuan, J. (2014). *FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2* <https://doi.org/10.5281/zenodo.22281>
- Schiel, Florian. (1999). Automatic phonetic transcription of non-prompted speech. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)*, 607–610.
- Sodagar, I. (2011). The mpeg-dash standard for multimedia streaming over the internet. *IEEE multimedia*, 18(4), 62–67.
- Young, S. J. (1994). *The HTK hidden Markov model toolkit: Design and philosophy*.
- Yuan, J. & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*.

# *megageil, mega geil, and voll mega: Intensification in YouTube comments*

Louis Cotgrove

Leibniz-Institute for the German Language

E-mail: [cotgrove@ids-mannheim.de](mailto:cotgrove@ids-mannheim.de)

## Abstract

This paper analyses intensification in German digitally-mediated communication (DMC) using a corpus of YouTube comments written by young people (the *NottDeuYTSch* corpus). Research on intensification in written language has traditionally focused on two grammatical aspects: syntactic intensification, i.e. the use of particles and other lexical items and morphological intensification, i.e. the use of compounding. Using a wide variety of examples from the corpus, the paper identifies novel ways that have been used for intensification in DMC, and suggests a new taxonomy of classification for future analysis of intensification.

**Keywords:** intensification, semiotics, computer-mediated communication, youth language, corpus linguistics, pragmatics, interaction

## 1. Introduction

Digitally-mediated communication (DMC) has often been described as containing a higher concentration and broader variety of “expressive markers” than other written language (Hilte, Vandekerckhove, & Daelemans 2019), and similar has been said of youth language (Tagliamonte 2016), in particular, the use of intensifiers.<sup>1</sup> Intensification has traditionally been defined as a grammatical process referring to the modification (both amplifying or reducing) of the quality of an element in a sentence by another element (an intensifier) (Bolinger 1972; Quirk et al. 1985). For example, the adverb *sehr* (‘very’) modifies an adjective to increase its quality, e.g. *sehr geil* (‘very awesome’), or the prefix *semi* reduces the quality of the adjective to which it is attached, e.g. *semigeil* (‘semi-awesome’). This paper adopts the above definition of intensification with one proviso: the element used to intensify does not need to be a word or part of a word, rather it can be any digital sign or process.

The paper first examines existing approaches to intensification before analysing how young people intensify in DMC, using a corpus of YouTube comments written by young people between 2008 and 2018, the

*NottDeuYTSch* corpus (Cotgrove 2018), as the basis for the investigation. The examples used in the analysis all come from the corpus. The article also uses the variety of intensification in youth DMC to suggest a potential framework of analysis for the classification of intensifiers according to their grammatical and visual characteristics.

## 2. Approaches to intensification

Research on intensification has focused on several different thematic areas, such as sociolinguistic aspects, i.e. which intensifiers are used by a particular social group (e.g. Macaulay 2006; Tagliamonte 2008; Reichelt & Durham 2017), or how intensification is used with a particular word class, most popularly adjectives (e.g. Kirschbaum 2002; Claudi 2006; Reichelt & Durham 2017). However, the focus of this paper is on the forms that intensifiers can take, and earlier research on written German identified three means of intensification: the two grammatical forms of intensification: morphological and syntactic intensification (see Kirschbaum 2002), and a stylistic-based means, referred to as expressive intensification (see Aitchison 1994: 19-20).

- (1) [...] das Video ist einfach so urgeil!!!!  
[...] the video is simply so utterly awesome!!!!

Morphological intensification is the use of compounding, where the base lexeme is intensified, most frequently with

<sup>1</sup> A list of alternative terms for intensifier can be found in Stratton (2020: 188), such as ‘degree word’, ‘Gradierer’, and ‘Intensitätspartikel’, however, in this article, the term ‘intensifier’ is used.

the pattern **prefix + adjective**, e.g. *urgeil* in Example 1, although many other combinations are valid, e.g. *affengeil* (noun + adjective).

- (2) so so geil wie sau XD  
so so awesome as hell XD
- (3) [...] tanzt bei Party Rock aber geil o.o  
Wow [...] dances to Party Rock awesomely o.o
- (4) Haha, ja, die Augenbrauen waren etwas strange.  
Haha, yes, the eyebrows were somewhat strange.

Syntactic intensification is the use of “various grammatical categories” to modify the quality of a word or phrase. These include adverbs (e.g. *so* in Examples 1 and 2), phrases (e.g. *wie Sau* in Example 2), particles (e.g. *aber* in Example 3), and indefinite pronouns (e.g. *etwas* in Example 4).

- (5) geil geil geil einfach geil  
awesome awesome awesome simply awesome

Expressive intensification is the use of self-repetition of lexical items, e.g. *geil* in Example 5 (which also contains syntactic intensification). It is commonly seen as a rhetorical device in poetry and has multiple functions, such as anaphora and epistrophe, such as to create rhythm and movement in the text, or to link ideas, but it can also be used to intensify emotions or feelings (Attridge 1994).

Gutzmann (2011) and Schmidt (2022) argued that lexical choice could also provide an intensifying effect, citing the indexical differences between the use of *dog* and *cur* in the sentence “This dog/cur howled the whole night” - the two are near synonymous, but *cur* has a more negative connotation, which they argue demonstrates intensification. While the indexical aspects of lexical choice can affect the strength or meaning of a message (Silverstein 2003), this does not fit within the definition of intensification used in this paper and as such is not investigated further in youth DMC, as *dog* is replaced entirely, rather than having a quality scaled.

Although mainstream research in the German language focused on morphological and syntactic intensification (Stratton 2020: 186), research in DMC additionally

identified a number of grapheme-based ways of intensification, such as the repetition of individual letters, e.g. *geeeiiiilll*, the use of capital letters (shouting capitals), e.g. *GEIL*, or indeed, a combination of the two, e.g. *EEEEEEEEEEEEIL* (Runkehl, Schlobinski, & Siever 1998; Androutsopoulos 2000). Despite the long-standing DMC-focused literature on graphemic intensification, it has only recently begun to be legitimised and analysed alongside other forms intensification. Philipp et al. (forthcoming: 2), for example, suggested that graphemic, syntactic and morphological intensification should be incorporated into a more general model, which also would include the repetition of intensifiers, e.g. *sehr sehr cool*. However, this paper shows that intensification in youth DMC in fact goes beyond the model suggested by Philipp et al. (forthcoming), and Section 4 analyses examples from the *NottDeuYTSch* corpus, demonstrating the extent to which features of DMC can be used to convey intensification.

### 3. Intensification in youth DMC

An analysis of YouTube comments in the *NottDeuYTSch* corpus reveals that intensification in youth DMC contain methods to intensify that have not been previously covered in existing research in the field. These include new ways of intensifying that would be classified within existing categories, as well as ways of intensifying that require an additional category.

- (6) Dass du so oft geklickt wurdest ist doch gar kein Wunder. Du bist einfach geilomatico!!!!  
It is no wonder at all that you get so many views.  
You are simply awesomesauce!!!!
- (7) einfach nur Geilheit  
just simply awesomeness

Example 6 demonstrates morphological intensification through the use of suffixation (*omatico*), and Example 7 also demonstrates intensification through suffixation (*-heit*) as well as derivation, changing the word class from an adjective to a noun. These processes have not been considered as within the existing definition of morphological intensification. However, in youth DMC, such constructions are relatively common and productive, for example we find *geilo*, *geili*, and *geilonachstman* in



(8) Ich finds mega geil xDDDD  
I find it mega awesome xDDDD

(9) ich finde es megageil! :D  
I find it mega-awesome! :D

(10) hater is the BeSt!!!

(11) Das Video is M Ü L L  
The video is R U B B I S H

(12) DIGGA.....DU HAST DIE PUNCHLINES  
GEFLOWT!!!!!!!!!!!! DAS WAR —>FRESH<—  
BRO.....YOU FLOWED THE  
PUNCHLINES!!!!!!!!!!!! THAT WAS —  
>FRESH<—

as ubiquitous as they are within DMC. While graphicons can undoubtedly influence the reception and tone of a message, I would argue that they have a different function, i.e. they do not directly intensify a word or phrase but provide illocutionary force (or other metacommunicative function) to the message (Cotgrove 2022: 242-244).

#### 4. Conclusion and Future Research

Through the examination of YouTube comments in the *NottDeuYTSch* corpus, this paper has demonstrated the wide variety of ways in which young people intensify in DMC. The innovation and creativity in the examples, identified through a corpus-based approach, have shown the need to expand the current understanding of what digital features can be used to intensify and how they can be categorised. The paper has shown that the definitions of existing categories of intensification need to be expanded, i.e. morphological, syntactic, and graphemic, and that it is necessary to introduce a new category of intensification, typographical, that will help researchers more fully understand the variety of ways in which it is possible to intensify in DMC.

The paper also serves as the basis for the potential development of a new general framework or taxonomy for intensification, hopefully serving as a base for future research in the field. This could include the incorporation of phonological intensification, i.e. the use of intonation or emphasis, to help analyse multimodal DMC or phonological differences between syntactic and morphological intensification, or examine the differences in intensity between different types and combinations of intensification. Such a framework could also help analyse whether typographical and graphemic means of intensifying are gradually replacing morphological and syntactical ways of intensifying.

#### 5. References

- Aitchison, Jean. 1994. “‘Say, Say It Again, Sam’: The Treatment of Repetition in Linguistics”, *SPELL: Swiss Papers in English Language and Literature*: 15–34 <<https://doi.org/10.5169/seals-99896>>
- Androutsopoulos, Jannis. 2000. ‘Non-Standard Spellings in Media Texts: The Case of German Fanzines’, *Journal of Sociolinguistics*, 4: 514–33 <<https://doi.org/10.1111/1467-9481.00128>>
- Attridge, Derek. 1994. ‘The Movement of Meaning : Phrasing and Repetition in English Poetry’, *SPELL: Swiss Papers in English Language and Literature*: 61–83 <<https://doi.org/10.5169/SEALS-99899>>
- Bolinger, Dwight. 1972. *Degree Words* (De Gruyter) <<https://doi.org/10.1515/9783110877786>>
- Claudi, Ulrike. 2006. ‘Intensifiers of Adjectives in German’, *Language Typology and Universals*, 59.4 (De Gruyter (A)): 350–69 <<https://doi.org/10.1524/stuf.2006.59.4.350>>
- Cosentino, Gianluca. 2017. ‘Stress and Tones as Intensifying Operators in German’, in *Exploring Intensification: Synchronic, Diachronic and Cross-Linguistic Perspectives*, Studies in Language Companion Series, ed. by Maria Napoli & Miriam Ravetto (John Benjamins Publishing Company), pp. 193–206 <<https://doi.org/10.1075/slcs.189.10cos>>
- Cotgrove, Louis Alexander. 2018. ‘The Importance of Linguistic Markers of Identity and Authenticity in German Gangsta Rap’, *Journal of Languages, Texts, and Society*, 2: 67–98
- . 2022. ‘#GlockeAktiv: A Corpus Linguistic Investigation of German Online Youth Language’ (unpublished PhD Thesis, Nottingham: University of Nottingham) <<https://eprints.nottingham.ac.uk/id/eprint/69043>>
- Gutzmann, Daniel. 2011. ‘Expressive Modifiers & Mixed Expressives’, *Empirical Issues in Syntax and Semantics*, ed. by Olivier Bonami & Patricia Cabredo Hofherr: 123–41
- Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. 2019. ‘Expressive Markers in Online Teenage Talk’, *Nederlandse Taalkunde*, 23.3: 293–323 <<https://doi.org/10.5117/NEDTAA2018.3.003.HILT>>
- Kirschbaum, Ilja. 2002. ‘Schrecklich nett und voll

- verrückt: Muster der Adjektiv-Intensivierung im Deutschen' (Düsseldorf: Heinrich-Heine-Universität Düsseldorf)
- Macaulay, Ronald. 2006. 'Pure Grammaticalization: The Development of a Teenage Intensifier', *Language Variation and Change*, 18.03 <<https://doi.org/10.1017/S0954394506060133>>
- Philipp, J. Nathanel et al. forthcoming. 'The Role of Information in Modeling German Intensifiers'
- Quirk, Randolph et al. 1985. *A Comprehensive Grammar Of The English Language* (Longman)
- Reichelt, Susan, & Mercedes Durham. 2017. 'Adjective Intensification as a Means of Characterization: Portraying In-Group Membership and Britishness in *Buffy the Vampire Slayer*', *Journal of English Linguistics*, 45.1: 60–87 <<https://doi.org/10.1177/0075424216669747>>
- Runkehl, Jens, Peter Schlobinski, & Torsten Siever. 1998. 'Sprache Und Kommunikation Im Internet', *Muttersprache*, 108: 97–109
- Scheffler, Tatjana, Michael Richter, & Roeland Van Hout. 2023. 'Tracing and Classifying German Intensifiers via Information Theory', *Language Sciences*, 96: 101535 <<https://doi.org/10.1016/j.langsci.2022.101535>>
- Schmidt, Jessica. 2022. 'Do Intensifiers Lose Their Expressive Force over Time? A Corpus Linguistic Study', in *Particles in German, English, and Beyond*, Studies in Language Companion Series, ed. by Remus Gergel, Ingo Reich, & Augustin Speyer (Amsterdam: John Benjamins Publishing Company), CCXXIV, pp. 69–94 <<https://doi.org/10.1075/slcs.224>>
- Silverstein, Michael. 2003. 'Indexical Order and the Dialectics of Sociolinguistic Life', *Language & Communication*, 23.3-4: 193–229 <[https://doi.org/10.1016/S0271-5309\(03\)00013-2](https://doi.org/10.1016/S0271-5309(03)00013-2)>
- Stratton, James M. 2020. 'Adjective Intensifiers in German', *Journal of Germanic Linguistics*, 32.2 (Cambridge University Press): 183–215 <<https://doi.org/10.1017/S1470542719000163>>
- Tagliamonte, Sali A. 2008. 'So Different and Pretty Cool! Recycling Intensifiers in Toronto, Canada', *English Language and Linguistics*, 12: 361–94 <<https://doi.org/10.1017/S1360674308002669>>
- . 2016. *Teen Talk: The Language of Adolescents* (Cambridge: Cambridge University Press)
- Wyss, Eva L., & Barbara Hug. 2016. 'WhatsApp-Chats. Neue Formen der Turn-Koordination bei räumlich-visueller Begrenzung', in *Jugendsprache in Schule, Medien und Alltag*, ed. by Carmen Spiegel & Daniel Gysin (Frankfurt am Main: Peter Lang), pp. 259–74 <<https://doi.org/10.3726/978-3-653-04950-3>>

# Exploring Register Variation in Turkish Web Corpus

Selcen Erten

University of Turku, Finland

[seerte@utu.fi](mailto:seerte@utu.fi)

## Abstract

In linguistics, web registers are language varieties occurring on the web such as *news reports* and *editorials*. Most of the previous web register research has been done for Indo-European languages. Additionally, previous research has mainly focused on the restricted corpora with pre-determined registers. This article describes Turkish web registers on the web. The data is Turkish web register corpus which consists of 2601 web texts with a total number of 11 million of words. A taxonomy was adapted to register label these texts. The manual annotations of the texts were done with the adapted taxonomy, and the registers were defined accordingly. Text dispersion keyword analysis was used to generate the keywords of the registers and examine the basic linguistic characteristics of them. The results display the web registers existing for Turkish, and the linguistic characteristics associated with the *news report* and *editorial* registers.

**Keywords:** Turkish web registers, manual annotation, text dispersion keyword analysis.

## 1. Introduction

In linguistics, registers are language varieties written in a particular situation with pervasive linguistic features that serve important functions within that situation of use. (Biber, 1988; Biber & Conrad, 2019). Considering that the web is possibly the first source one resorts to when seeking information, it is important to understand registers on the web. Registers occurring on the web are called web registers. Some examples of them are *news reports* and *editorials*. Understanding web registers is crucial to be able to distinguish, e.g., facts from opinions and advertisements from informative texts.

Register studies have a relatively long history in linguistics (Biber, 1988). However, most of the previous research has been restricted to English and other Indo-European languages (Biber & Finegan, 1994; Conrad & Biber, 2001; Asencion-Delaney & Collentine, 2011; Berber-Sardinha et al., 2014; see, however, Kim & Biber, 1994; Jang, 1998; Ravid & Berman, 2009; Aksan & Aksan, 2015). Additionally, previous research has mainly focused on carefully curated restricted corpora, where the documents have been selected manually from established sources featuring pre-determined registers. This has led to a situation where registers are typically examined in discrete classes where texts are very good examples of their categories. The web, on the other hand, offers a very different perspective to register variation by including a wide and sometimes noisy range of documents (Biber & Egbert, 2018). There are no gatekeepers to ensure that documents follow the guiding principles of specific registers on the web. Further, not all documents have a single register or any register at all (Santini, 2007; Egbert et al., 2015). By taking the full variation into account, a much more complete understanding of web register variation can be gained than what the current studies based on restricted samples can offer.

## 2. Present Study

In the current study, web registers in Turkish are defined by using a web register taxonomy adapted from English (Egbert et al., 2015) and Finnish (Laippala et al., 2019) to Turkish. Although the pioneering register studies have been done in Indo-European languages, English being the most

studied, understanding language use on the internet will be restricted if the examination of registers is limited to a certain group of languages. For this reason, culturally and linguistically different languages need to be studied so that more understanding of web registers can be acquired not only for language research but also for the applications of web registers in the areas of media literacy and intercultural communication. Turkish is culturally and linguistically a very different language than the commonly studied Indo-European languages. Considering that people are regularly surrounded by media, and there is little incentive to employ an ‘off’ switch (Butler, 2020), it is vital that people know how to be able to judge what is useful and misleading information and when media can be trusted (Livingstone, 2018). Further, all cultures use language for different communicative purposes in different situations. Registers are based on pervasive patterns of linguistic variation across such situations (Biber & Conrad, 2019; Biber et al., 2020). Understanding Turkish web registers will help uncover misunderstandings and failures in intercultural communication.

In this study, registers displaying the features of *news reports* and *editorials* are specifically examined with text dispersion keyword analysis to see their basic linguistic characteristics. *News reports* are texts typically written by professionals to report on recent events while *editorials* are texts typically written by professionals on a news-related topic with a purpose to persuade the reader about opined points.

In the light of the aims of defining Turkish registers on the web and examining the linguistic characteristics of them, the following research questions are answered:

1. Which web registers exist for Turkish in terms of the defined categories by Egbert et al. (2015) and Laippala et al. (2019)?
2. What are the basic linguistic characteristics of *news report* and *editorial* registers?

## 3. Data and Methodology

### 3.1. Data

The data in the study is Turkish web register corpus. The corpus targets the full Turkish speaking web. It is based on a random sample of the web, originally computationally

collected by Common Crawl (commoncrawl.org) and cleaned and pre-processed within *Massively Multilingual Modelling of Registers in Web-scale Corpora* project run by TurkuNLP team at the University of Turku, Finland. Altogether, the corpus consists of 3767 unique web texts covering various domains in Turkish with a total of 21 million words.

Regarding the ethical issues in the phase of collecting data for the Turkish web register corpus, the guidelines published by the Finnish National Board on Research Integrity TENK and the Turkish Council of Higher Education on Scientific Research Directive were followed. Upon the completion of manual annotations, it was assured that there are no personal elements in the data collected and used for this research. The data is openly and freely accessible on the web.

### 3.2. Methodology

The taxonomy used to register-label the web texts were adapted from Egbert et al. (2015) and Laippala et al. (2019) to Turkish and its specificities. The benefit of using this taxonomy is that it allows to annotate registers in a wide range of different types of documents with no dependence on pre-determined categories. Manual annotations of each text in the corpus were completed with the adapted taxonomy on the annotation tool Prodigy. On Prodigy, the texts were first accepted or rejected. Accept means that the text was put in the data, and it was register-labelled. The document was rejected in the situations such as where the text consisted of only short list of items, the sentences did not form a coherent text, the amount of coherent text was very small compared to the junk text or the text was not in the target language (Biber & Egbert, 2018). In the data, around 35% of the texts were rejected for one or more of these reasons. Some of the accepted documents were annotated as hybrids, which means that they were given two labels or more although it was typically two. This occurred when a text featured characteristics of more than one register such as a marketing text followed by reviews (description with intent to sell + review). Compared to the single-category registers where each text had only one register, the hybrid texts were a few. Hybrid registers were not included in this study, and only 2601 accepted, single-category registers were examined.

In addition to the manual annotations of the Turkish web texts, keyword analysis method was used to examine the basic linguistic characteristics of *news reports* and *editorials*. The concept of keyness in text have been discussed in various ways (Scott, 1997; Bondi, 2010; Culpeper & Demmen, 2015), yet there are two fundamental approaches in corpus linguistic methods which determines the keyness in frequency (Scott & Tribble, 2006) and in dispersion (Egbert & Biber, 2019). In this study, text dispersion keyword analysis was used, as it is seen as the most suitable method for register studies with large corpora containing many texts. Text dispersion keyness uses the text, rather than the corpus, as the unit of observation. It is based on a word's dispersion across the texts of a corpus rather than its overall frequency in the corpus (ibid). This means that text dispersion keyness disregards word frequency entirely but generates keyword lists based on word dispersion across texts. Log-likelihood is used as it estimates probabilities more accurately even when the counts are low, and because the dispersion of the words across texts tend to follow a Zipfian distribution. The

requirements for text dispersion keyness are many texts in target and reference corpora as well as a special program. In frequency-based keyness, the most frequent words are general high-frequency words which are not particularly distinctive to the target corpus. The text dispersion method, on the other hand, identifies words which are much more strongly related to the target corpus than the reference corpus. In the current study, both the target and reference corpora were generated from the web text data. If, for example, texts of news reports were the target corpus, the reference corpus was all the other texts belonging to various registers minus news reports. As for the special program to acquire the text dispersion values, Python codes were utilised for the purpose.

## 4. Results

### 4.1. Web registers of Turkish

Based on the taxonomy adapted to Turkish, 9 main register categories and sub-registers falling under them were identified.

Below, the web registers defined for Turkish are displayed without the distinction between main or sub-registers. The total number of texts and number of words for each register are also as in the following:

Register	Number of texts	Number of words
Description with intent to sell	645	1,920,852
News report	556	1,502,494
Machine translated	329	1,819,394
Other-informational description	224	954,392
Description of a thing or person	124	524,602
Legal terms	105	593,928
Editorial	93	774,258
Review	66	240,271
Opinion blog	58	377,527
Narrative blog	52	325,364
Interactive discussion	50	595,558
Advice	46	193,699
Recipe	40	81,555
Other-informational persuasion	40	103,980
Sports report	30	66,565
Religious blog	29	323,391
Other-spoken	20	59,908
Other-how-to or instruction	22	62,427
Encyclopaedia article	18	93,131
Other-opinion	16	132,042
Lyrical	16	39,810
Interview	12	107,967
FAQ	6	27,030
Research article	4	19,649
Total	2601	10,939,794

Table 1: Registers identified in Turkish web register corpus with their numbers of texts and words.

As seen, there are *other* categories among the registers identified for Turkish. *Other* means that the text fell under one of the main categories, but it could not completely be annotated as one of the sub-categories of the main category.

Although this study does not focus on the *other* categories, they might in fact show language and culture-specific features.

## 4.2. Linguistic characteristics of news reports and editorials

Egbert & Biber (2019: 87) state that the top 100 keywords suffice to show the strengths of the text dispersion keywords.

The top 100 keywords for news reports and their values of keyness are displayed below:

Keyword	Translation	Keyness
1 dedi	s/he said	295,277
2 başkanı	chairman of	289,815
3 konuştu	s/he spoke	202,403
4 söyledi	s/he said	167,021
5 başkan	chairman	142,405
6 etti	s/he did	138,902
7 kullandı	s/he used	128,390
8 belediye	municipality	127,679
9 ifadelerini	expressions of	112,388
10 edildi	it was done	109,101
11 belirtti	s/he indicated	106,923
12 belirlen	... who indicated	104,918
13 bulundu	it was found	103,052
14 kaydetti	s/he noted	92,111
15 açıklamada	in the statement	89,773
16 belediyesi	municipality of	88,350
17 bin	thousand	84,626
18 belirterek	by indicating	83,929
19 koronavirüs	coronavirus	82,141
20 verdi	s/he gave	73,258
21 yapıldı	it was done	68,539
22 bakanı	minister of	67,920
23 katıldı	s/he participated	65,634
24 müdürü	director of	62,220
25 ifade	expression	60,086
26 covid	covid	59,936
27 ilçe	county	58,616
28 haber	news	57,432
29 yaptığı	...which s/he did	57,014
30 alındı	it was taken	55,330
31 chp	chp (republican party)	54,503
32 sözlerine	to the statements of	54,251
33 il	province	53,835
34 ardından	afterward	53,394
35 büyükşehir	metropolis	52,103
36 ekipleri	teams of	51,055
37 itfaiye	fire department	50,234
38 ilçesinde	in the county of	49,755
39 mustafa	mustafa	48,647
40 onaylanmamaktadır	it is not (being) approved	47,076
41 olay	incident	46,793
42 dile	to the tongue	46,544
43 öğrenildi	it was learnt	46,308
44 müdürlüğü	directorship of	46,241
45 mehmet	mehmet	46,197
46 yardımcısı	vice of	45,138
47 heyeti	board of	43,955
48 harflerle	with the letters	42,862
49 bildirildi	it was informed	41,520
50 '	'	41,228
51 şunları	those	41,168
52 belirtildi	it was stated	40,975
53 edinilen	...which was acquired	40,080
54 dr	dr (doctor)	39,899
55 açıkladı	s/he explained	38,371
56 milyon	million	38,370
57 verildi	it was given	37,999
58 hesabından	from the account of	37,422
59 aa	aa (anatolian agency)	36,433
60 parti	party	35,683
61 basın	press	35,259

62	soruşturma	investigation	35,121
63	vurgulayan	...who underlined	34,899
64	konusan	...who spoke	34,779
65	vurguladı	s/he underlined	34,761
66	19	19	34,721
67	polis	police	34,115
68	kullanılmayan	...which was/is not used	34,115
69	ekiplerinin	of the teams of	34,077
70	muhabirine	to the journalist of	34,077
71	vatandaşlar	citizens	33,988
72	ilişkin	related	33,782
73	söyleyen	...who told	33,602
74	milletvekili	congressman	33,444
75	katıldığı	...which s/he participated	33,444
76	inşallah	God willing	33,276
77	yaşındaki	in the age of	33,276
78	vali	governor	32,831
79	devam	continuation	32,804
80	gerçekleştirildi	it was fulfilled	32,630
81	gözetim	to the custody	32,321
82	tedbirleri	precautions of	31,850
83	değinen	...who mentioned	31,091
84	salonunda	in the hall of	31,068
85	salgını	epidemic of	31,068
86	toplantısında	in the meeting of	31,068
87	yüzde	per cent	29,993
88	başlatıldı	it was started	29,968
89	açıklamalarda	in the statements	29,771
90	saatlerinde	in the time of	29,732
91	kaydeden	...who noted	29,514
92	olayla	with the incident	29,511
93	jandarma	gendarme	29,404
94	yaralı	injured	29,404
95	recep	recep	28,778
96	düzenlenen	...which was organized	28,344
97	içişleri	internal affairs	28,150
98	yaralandı	s/he was injured	28,118
99	tarım	agriculture	27,618
100	önümüzde	ahead of us	27,617

Table 2: Top 100 text dispersion keywords of news reports and their values of keyness.

Closer inspection of the table shows that 53 keywords emerge as nouns, 33 keywords as verbs and 14 as *other*.

Among the nouns, most nouns are administration words such as *chairman*, *municipality*, *minister*, *director*, *board*, *citizen*, *congressman* and *governor*. There are other nouns falling under the themes of disaster (*covid*, *coronavirus*, *fire department*), legality (*investigation*, *police*, *custody and gendarme*), journalism (*aa*, *journalist*, *press*, *news*) and communication (*expression*, *statement*, *utterance*).

Among the verbs, 23 of them are finite verb forms (predicates) while 13 of them are non-finite. With one exception, all predicates have *past tense* + *3<sup>rd</sup> person singular* pattern, which seems to be the pattern for news reports that report what happened. In addition to *past tense* + *3<sup>rd</sup> person singular* pattern, half of the predicates have *passive voice*, which also emerges as a pattern in news reports where the action is important. Passive voice also emerges in non-finite verb forms of the keywords of news reports. There are three non-finite verb forms in Turkish, which are verbal nouns, participles and converbs (Göksel & Kerslake, 2005). In the keyness of news reports, all non-finite verb forms, with one exception, were found to be as participles: non-finite verb forms of relative clauses formed with *who* and *which*. When both predicates and non-finite verb forms are considered together, it is seen that there are many communication verbs such as *say*, *tell*, *note*, *underline*, *inform*, *state* and *explain* used in news reports.



When it comes to editorials, the striking thing featuring for the keyness of editorial texts is that seven different part-of-speech classes and *other*-category were identified:

42 nouns, 5 adverbials,  
13 discourse connectives, 4 postpositions,  
13 adjectives, 4 pronouns,  
7 verbs, 12 *other*-category words.

Top 100 text dispersion keywords for editorials and their values of keyness are as in the following:

Keyword	Translation	Keyness
1 iktidar	rulership	73,517
2 ama	but	50,719
3 meselesi	matter of	48,418
4 bile	even	47,603
5 üstelik	what's more	46,382
6 karşı	against	43,233
7 yok	nonexistent, no	40,817
8 devlet	state	39,631
9 siyasi	political	39,364
10 ne	what	38,337
11 işte	"işte" (discourse connective)	37,848
12 demokratik	democratic	37,693
13 değil	not	37,681
14 asıl	actual	36,939
15 mi	"mi"	35,729
16 aslında	in fact	35,048
17 çıkarları	benefits of	35,007
18 kapitalizmin	of capitalism	34,625
19 halk	public	34,458
20 devletin	of the state	34,180
21 erdoğan	erdoğan	34,018
22 erdoğanı	erdoğan-accusative	33,580
23 düpedüz	sheerly	33,579
24 devrimci	revolutionary	32,875
25 dedikleri	..which they say	32,874
26 oysa	though	32,623
27 önünde	in front of	32,418
28 yana	sideways	31,565
29 mı	"mı"	31,484
30 o	she/he/it	31,218
31 çünkü	because	31,118
32 akp	akp (ruling party)	30,647
33 dı	"dı"	30,238
34 peki	well then	30,116
35 artık	now/anymore	30,114
36 çıkmış	ensued/out of joint	30,061
37 türkiyeyi	Turkey-accusative	29,706
38 öyle	as such	29,207
39 biçimi	way of	28,643
40 cumhurbaşkanının	the president's	28,244
41 daha	more, yet	28,036
42 başkanlık	presidency	27,657
43 zaten	already, anyway	27,122
44 parti	party	26,912
45 propaganda	propaganda	26,864
46 ağustosta	in August	26,863
47 protesto	protest	26,851
48 sözde	so-called	26,850
49 ortaya	into the pot	26,848
50 hedef	target	26,840
51 ettiği	...which s/he/it did/does	26,833
52 azından	least	26,728
53 var	there is/are	26,436
54 diye	called, in case	26,124
55 vardı	there was/were	26,036
56 ona	to him/her/it	26,036
57 gibi	as, like	26,032
58 demokrasi	democracy	25,879
59 erdoğanın	erdoğan's	25,878
60 tarihsel	historical	25,878
61 değişen	changing	25,748

62 muhalif	opponent	25,720
63 siyasal	political	25,719
64 vardır	there is/must be	25,705
65 -dır	"-dır"	25,575
66 cumhurbaşkanı	president of	25,568
67 iki	two	25,502
68 tümü	all-accusative	25,492
69 ülkenin	of the country	25,470
70 böyle	like this	25,448
71 yol	way	25,447
72 müslüman	muslim	25,280
73 şöyle	as such	24,882
74 diyerek	by saying	24,847
75 neden	why, reason	24,709
76 tur	round	24,349
77 gerçi	actually	24,309
78 dini	religion-accusative	24,298
79 çıkar	benefit	24,153
80 yani	namely	24,149
81 bizim	our	24,126
82 savaş	war	23,944
83 hiç	nothing, none	23,881
84 ya	"ya" (discourse connective)	23,395
85 diyor	s/he says	23,210
86 kesimleri	parts of	23,207
87 yıl	year	23,102
88 gazeteci	journalist	22,906
89 yaşanan	...which is/was encountered	22,872
90 toplumsal	societal	22,856
91 vatan	homeland	22,854
92 işgal	invasion	22,854
93 iktidarı	rulership of	22,847
94 işin	of the matter	22,843
95 toplum	society	22,506
96 kendini	oneself-accusative	22,391
97 insan	human	22,155
98 mesele	matter	22,035
99 batı	west	21,966
100 esad	esad	21,931

Table 3: Top 100 text dispersion keywords for editorials and their values of keyness.

The number of frequencies of the words is not in the scope of text dispersion keyness or of the current study. Nevertheless, the results showed that there is a variety of part-of-speech classes for editorial texts especially compared to news reports.

Among the nouns, most nouns are governance-related and political words such as *rulership*, *state*, *presidency* and *democracy*. The rest of the words falls under the themes of strategy (political strategy word *propaganda*, military strategy word *invasion* and economic strategy word *capitalism*), society (*society*, *human*, *public*), direction (*middle*, *west*, *target*), and belief (*religion*, *Muslim*).

The adjectives were found to be the words showing clear opinion and stance of the editorial writers such as *political*, *democratic*, *revolutionary* and *so-called*.

Most of the conjunctions and discourse connectives such as *but*, *in fact*, *whereas* and *actually* have adversative function. Further, there are some that have the examples of additive (*even*, *what's more*), causal (*because*), corroborative (*in any case*), expansive (*in other words*) and organizational (*"işte"*) functions.

The linguistic features of news reports and editorials show that news reports have only two part-of-speech classes and the *other*-category while editorials have seven part-of-speech classes and the *other*-category. Out of these categories, while past tense and passive voice are very regular in the predicates used in news reports, none of them appear for the predicates of editorials. The use of adjectives shows differences in news and editorials, as well. For

editorials, the adjective use covers 12 % of the keywords, yet for news, it is only 1 %. Another distinctive feature between news and editorial texts is the use of conjunctions and discourse connectives. Out of the top 100 keywords, 13 are conjunctions and discourse connectives in editorial texts while in news reports, it is none.

Editorial writers seem to use a variety of language to persuade the reader while news report writers report the recent events with less variety of language. News reports in Turkish seem to report on the recent happenings with the *-DI* perfective suffix rather than the *-mİş* evidential/perfective suffix. It might be a linguistic feature specific to news reports with a purpose to look more factual and updated, as the *-DI* perfective suffix in Turkish is used when the person witnessed the happenings and *-mİş* is used when the person learnt it through outer sources. In editorials, any of the past tenses do not seem to be a typical use, but adjectives seem to exist unlike in news reports. This might indicate that the use of adjectives is useful for editorial writers to strengthen their personal opinions based on reality while news reporters rather focus on reporting what happened. The conjunctions and discourse connectives usage with various functions in editorials also indicate that editorial writers' informational persuasion is provided with tailored choices of conjunctions and discourse connectives. For news reporters, this does not seem to be the case for a purpose of persuasion. It might be possible to state that they rather aim to look more factual and updated, and they seem to do it with the *-DI* perfective suffix in Turkish.

## 5. Conclusion

Understanding different language varieties on the web is important in an era when most of the information one needs is acquired from the web. Having the data provided from the Turkish web, Turkish web register corpus has text samples on a large variety of registers with 24 different categories.

News reports and editorials are two typical web registers which have many samples on the Turkish web but have distinctive linguistic characteristics. Understanding the differences between these registers on the web is crucial to be able to differentiate facts from opinions. While acquiring information from the web, it is important to understand which features of the language are preferred and what the purpose of the texts are so that media literacy as well as inter-cultural communication are successfully accomplished.

## 6. References

- Asención-Delaney, Y., & Collentine, J. (2011). A Multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32, pp. 299–322.
- Aksan, Y. & Aksan, M. (2015). Multi-word units in informative and imaginative domains. In *The 16<sup>th</sup> International Conference of Turkish Linguistics*. Ankara: Middle East Technical University.
- Berber-Sardinha, T.; Kauffman, C. & Acunzo, C. M. (2014). Dimensions of register variation in Brazilian Portuguese. In T. Berber-Sardinha & M. Veirano-Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*, Philadelphia: John Benjamins, pp. 35–80.
- Biber, D. (1988). *Variation Across Speech and Writing*. The UK: Cambridge University Press.
- Biber, D. & Finegan, E. (1994). Multi-dimensional analyses of author's styles: Some case studies from the eighteenth century. In D. Ross and D. Bring (Eds.), *Research in Humanities Computing*. Oxford: University Press, pp. 3–17.
- Biber, D. & Egbert, J. (2018). *Register Variation Online*. Cambridge: Cambridge University Press.
- Biber, D. & Conrad, S. (2019). *Register, Genre, and Style*. 2<sup>nd</sup> ed. the UK: Cambridge University Press.
- Biber, D., Egbert, J. & Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16 (3), pp. 581–616.
- Bondi, M. (2010). Perspectives on keywords and keyness. In M. Bondi & M. Scott (Eds.), *Keyness in Texts*. John Benjamins, pp. 1–20.
- Butler, A. T. (2020). *Educating Media Literacy: The Need for Critical Media Literacy in Teacher Education*. Leiden and Boston: Brill Sense.
- Conrad, S. & Biber, D. (2001). *Variation in English: Multi-dimensional Studies*. Eastbourne: Pearson Education.
- Culpeper, J., & Demmen, J. (2015). Keywords. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of Corpus Linguistics*. Cambridge University Press, pp. 90–105.
- Egbert, J.; Biber, D. & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66 (9), pp. 1817–1831.
- Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14 (1), pp. 77–104.
- Göksel, A. & Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.
- Jang, S. C. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study*. Unpublished doctoral dissertation. University of Hawaii, Manoa.
- Kim, Y. J., Biber, D. (1994). A corpus-based analysis of register variation in Korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, pp. 157–181.
- Laippala, V.; Kyllönen, R.; Egbert, J.; Biber, D.; & Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the Northern European Association for Language Technology*. Turku, Finland, pp. 292–298.
- Livingstone, S. (2018, July 27). Media literacy-Everyone's favourite solution to the problems of regulation. Parenting for a Digital Future. Retrieved from <https://blogs.lse.ac.uk/mediapolicyproject/2018/05/08/media-literacy-everyones-favourite-solution-to-the-problems-of-regulation/>
- Ravid, D. & Berman, R. (2009). Developing linguistic register across text types: The case of modern Hebrew. *Pragmatics and Cognition*, 17 (1), pp. 108–145.
- Santini, M. (2007). Characterizing genres of web pages: *Genre hybridism and individualization*. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. Hawaii, pp. 71
- Scott, M. (1997). PC analysis of key words. *System*, 25 (2), pp. 233–245.
- Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam/Philadelphia: John Benjamins.



# Digital Corpus Linguistic Analysis of the Language on disability and inclusion in social media – in a German corpus of 2,559 Tweets on #disability and #inclusion between 1st of December – 31st of December 2020

## Short abstract

This presentation examines the digital language usage on disability and inclusion – edited by disabled and non-disabled people – in Social Media. For this examination, we use a small corpus of 2,559 Tweets with 61,249 tokens as a part of a big corpus of 214,926 total with 5,663,504 tokens. The whole corpus consists of tweets published in the time span of 2007-2023 under the hashtags 'inclusion' and 'disability', while the small corpus was published from the first day until the last day of December 2020 UTC. This linguistic study provides valuable insights into the lexicon on disability, inclusion, the co-occurrences of the lexical units by using AntConc. In addition, our research goal focuses on the classification of the Tweets via Sentiment Analysis (SentiStrength). The paper shows the potential and capacity of a quantitative Corpus Linguistic examination carried out on a German corpus from Social Media on disability, inclusion, discrimination and exclusion. The study provides not only a quantitative lexicon analysis with AntConc but also combines this with a Sentiment Analysis, which is a research desideratum in German Corpus Linguistics. Our paper describes also the potentials of a quantitative lexical and sentiment analysis with SentiStrength for language studies on the communication on disability and inclusion in Social Media accomplished with a critical reflection of methodological issues.

**#Keywords:** #DisabilityDiscourse, #DiscourseofInclusion, #SentimentAnalysis, #LexicalAnalysis, #DigitalDiscourseAnalysis

## 1. Extended Abstract

Computer-Mediated Communication (CMC) encompasses various forms of communication, which take place through digital devices and networks. The language used in CMC can vary depending on the platform, context, and participants involved. According to the Call for Paper on CMC, “specialized corpora of the language of CMC and social media are increasingly vital for the analysis of the ‘diversity in terms of speakers and settings’ Barbaresi (2019: 29-30) in digital contexts”. Our paper would like to contribute to this conference plaidoyer in terms of speakers’ as well as of platform’s diversity in CMC with a computer-driven lexical analysis of the German discourse on #disability and #inclusion contributed in a time period of one month (from the 1st of December until the 31st of December 2023). Hence, our study is interested in the software-based-monitoring and recognition of vocabulary associated with ‘disability’, ‘inclusion’, ‘discrimination’ and ‘exclusion’ on Twitter, a platform which is in particular important to minorities, including people with disability. Minorities, also people with disability want to raise awareness for their own life, self-chosen references to disability as well as to their own view of inclusion but also discrimination and exclusion in everyday life. The German discourses on #disability and #inclusion can be lexically classified and characterized on social media on the basis of a corpus consisting of 2559 German Tweets. For a linguistic classification, in this paper we carry out a software-based-lexical analysis with AntConc and SentiStrength, in particular of the substantives ‘disability’ and ‘inclusion’ and associated lexical entities, which also is prevalent for a CMC-based-linguistic study of minority languages used reporting on issues and agenda of people with disability, also but not only, by people with disability. As according to Hall (2019) discriminatory language towards people with disabilities is on the rise, we carry out this quantitative CMC-study to get insights into the representation of people with disabilities as well as of

inclusion on social media based on the examination and visualization of the language and communication used when discussing disability and inclusion on Twitter. A semantic classification of Tweets into the categories negative, neutral and positive with SentiStrength as part of a Sentiment Analysis will supplement our research (e.g. Kiritchenko et al, 2014; Dai et al. 2017; Palomino et al., 2020) our linguistic research on the lexicon carried out with AntConc. AntConc was developed by Anthony Lawrence (Waseda University/Japan), SentiStrength by Mike Thelwall (University of Wolverhampton/England). Both of them are free available for non-profit goals.

## 2. Overview of the Research in Linguistics on Social Media Discourses of minorities

The communication of inclusion and exclusion of minorities have been studied extensively in the social sciences, and discourse analysis is beginning to catch up. In recent years, linguistic studies have focused on discourses on refugees and migrants (e.g. Viola & Musolff, 2019), of people on grounds of gender (e.g. Gnau & Wyss, 2019; Paknahad & Baker, 2017; disability (e.g. Grue, 2014; Sties, 2013) and on mental health issues (e.g. Harvey, 2012) etc. in various countries and contexts. Many of these studies make use of data from digital media, itself an increasingly popular object of study in linguistics (e.g. Marx & Weidacher, 2020; Wright, 2020; Knuchel & Bubenhofer, 2023). These important studies have raised awareness for the analysis of minority issues from the point of view of Corpus Linguistics and Discourse Analysis, but the communication of people with disabilities and towards them in digital media has barely been focused so far. Herrera (2022) argues that social media analytics tools need to be designed to support inclusive public services for all, including persons with disabilities. Sinclair (2010) emphasizes the importance of paying attention to social barriers that inhibit communications inclusion, rather than just technological barriers. Zelena (2020) explores, how new media platforms become the platform of communal

loss for users of different ages, genders, social statuses, and diverse Internet usage habits and socialization. Finally, Pan et al. (2014) examines the role of community diversity in influencing perceived inclusion of newcomers in the online community and the influence of such perception on newcomers' engagement intention. This wide range of the corpus linguistic research on language on social media shows the lack of interest in studies in terms of the language usage in digital discourses on disability, and inclusion.

### **3. Lexical study and Sentiment Analysis of the Language Usage on inclusion and disability on Social Media as a Research Desideratum**

Despite of having raised the awareness for the necessity of studying minority discourses and inclusive communication on disability in digital media has barely been focused so far. In addition, a software-based monitoring of the semantics of the lexicon associated with inclusion in digital discourses on disability and inclusion has not been carried out so far on corpus consisting of Tweets. Given this desideratum, this presentation addresses itself to the corpus linguistic study of disability and inclusion in social media discourses on the basis of a corpus of 2559 Tweets. This remit includes the vocabulary on 'disability' and 'inclusion' by both members and non-members of the diverse group of people with disabilities. This small study is part of a study on the communication of inclusion related with disability in Social Media supported by the Bavarian Research Institute for Digital Transformation, funded by the Bavarian Ministry of Art and Science and led by Annamária Fábián. The whole project examines the linguistic and discursive aspects of references used for describing disability but also covers the communicative aspects of inclusion, discrimination and exclusion and provides an analysis on the digital communication of critical aims fostering the inclusion and/or countering the exclusion of people with disabilities. We would like to highlight for our study that an analysis of the communication of inclusion in social media pays attention to the diversity of communities present in social media channels as well as to social barriers that inhibit communicative inclusion. Moreover, people with disability, have been successfully engaged for more than 10 years on Social Media for inclusion through visibility. Overall, this paper is the first paper of our project with a special research issue from the point of view of digital Corpus Linguistics.

Our research design includes quantitative research methods, while pursuing following goals:

- (1) We observe the core communication and vocabulary in the Twitter discourse on disability and inclusion to get a first impression on the Semantic and the Sentiment of the communication in a digital discourse on disability.
- (2) We provide a lexical analysis including the analysis of collocations and N-Grams (Corpus-driven lexical Analysis)

on disability and inclusion in our Twitter corpus.

- (3) We classify the Tweets as part of our digital corpus in negative, neutral and positive (Sentiment Analysis).

From this reason, a team consisting of Corpus Linguists and Computational Scientists have gathered digital data and apply methods of both sciences. We would like to thank Prof. Dr. Jürgen Pfeffer (Technical University of Munich/Computational Social Sciences), who supported our project with the collection of big data.

### **4. Methods of Digital Corpus Linguistics**

This study will also bring methodological reflections on the affordances on this issue in Social Media. In addition, the project aims to gaining insights into effective digital linguistic methods (tools, software etc.) adaptable for the communicative analysis of data in digital media.

Dai et al. (2017) proposes a word embedding based clustering method for Twitter classification that achieves good accuracy without requiring labeled training data. Lui & Baldwin (2014) but also Heaton et al. (2023) evaluate off-the-shelf language identification systems for Twitter messages and their usability for linguistic analysis. Lui (2014) finds that simple voting over three specific systems consistently outperforms any specific system. Yang (2014) proposes a methodology for translating surveys into social media surveillance, which achieves better precision and recall than standard methods using lexicons or classifiers. While Yurchenko & Ugolnikova (2021) focus on linguistic methods in social media marketing, the paper highlights the relevance of linguistic methods in the digital age and their potential for improving social media communication monitoring accuracy. The quantitative background of this digital linguistic study is twofold:

- (1) We will carry out a lexical analysis of the Twitter corpus on #disability and #inclusion by using AntConc, a tool often used by digital linguists. We decided for AntConc as an 'equipment' as the adaptability of AntConc is useful for capturing and visualizing the lexical units and their collocates.
- (2) We will conduct a sentiment analysis with SentiStrength. SentiStrength is a sentiment classification tool which does not need proficiency in Machine Learning and can easily be used also by digital linguists without a background in computational linguistics. In addition, according to Palomino et al. (2020: 8), SentiStrength can be employed "to identify the polarity of tweets as positive, negative or neutral, though SentiStrength can also work as a binary classification tool – positive or negative."

## 5. A data-driven Semantic Study of ‘disability’ and ‘inclusion’ in a digital corpus on Twitter

### 5.1 Doing data-driven Semantic-Analysis with SentiStrength and AntConc – methodological considerations for Corpus Linguists

Before processing with our Corpus Linguistic Study with SentiStrength, we needed to prepare our corpus for working with this program from the Linguistic point of view as SentiStrength was developed to analyse shorter texts line by line especially for business purpose. First, we needed to remove all line breaks in a large corpus like ours on the hashtags *Inklusion* (‘inclusion’) and *Behinderung* (‘disability’) from 01/01/2007 to 31/03/2023 with 5,663,504 tokens for an overall analysis at sentence level. In addition, SentiStrength does not output the results in a separate file, but writes them to a txt UTF-8 corpus file, which slightly doubles in size as a result. These framework conditions imply that the program cannot analyse large corpora. This fact leads to our decision to reduce our corpus for this paper and provide a Sentiment analysis on the communication of only one month. For the analysis, however, we chose the month December of 2020, which was in the middle of the Covid lockdown in German-speaking countries which has an impact on the Sentiment Analysis in the corpus as ‘COVID’ is quite frequent<sup>1</sup>. This part of our large corpus consists of 2,559 tweets, 950 sentences, 61,249 tokens and 11,251 types.

The German sentiment strength dictionary file *EmotionLookupTable\_v5\_fullforms* for the program SentiStrength was provided by Sentistrength (<http://sentistrength.wlv.ac.uk>) and Hannes Pirker, Interaction Technologies Group at the Austrian Research Institute for Artificial Intelligence (OFAI) with additions from Elias Kyewski of the University of Duisburg.

SentiStrength performs the sentiment analysis using a sentiment strength dictionary, in which lexemes are assigned a sentiment rating. Positive sentiment ratings are marked with a scale of 1 to 5, negative ones with a scale -1 to -5. Each lexeme is rated with a maximum of 4 or -4, only repeated occurrences can result in a rating of 5 or -5 for a phrase. A neutral sentiment of a lexeme is marked with 0. In this paper, the positive numbers are always marked with a plus sign, i.e. the positive scale is +1 to +5.

In the case of sentences, the rating is always made up of a negative and a positive rating, e.g. -2/+3. These two ratings of a sentence result from the addition of the positive ratings and the addition of negative ratings. The sum is capped at +5 or -5. When the overall sentiment rating of a sentence is calculated, the maximum values which can result are +4 (=+5-1) or -4 (=+1-5).

While using SentiStrength, our first considerations were that this dictionary file *EmotionLookupTable\_v5\_fullforms*

is very extensive for negative words such as insults. We also considered that the negative ratings are sometimes inconsequent as serious verbal insults such as *Scheiße* (‘shit’, ‘fuck’ or ‘fucking’) are rated at -3, but *leider* (‘unfortunately’) at -4. As the consequence of this consideration, we decided to correct this: In our new sentiment strength dictionary file *EmotionLookupTable\_v6\_fullforms*, *Scheiße* (‘shit’, ‘fuck’ or ‘fucking’) is rated at -4, and *leider* (‘unfortunately’) at -3. Another observation on SentiStrength was that the sentiment strength dictionary v5 contains only few positive words. Positive foreign words and positive word formations, that are typically for German, are enormously underrepresented in the lexicon of SentiStrength. Especially in the German-speaking countries, non-partisan recognized political words which express a high grade with a high level of positivity (‘Hochwertwörter’) such as *gerecht* (‘just’) or *sozial* (‘social’) – also often occurring in corpora on social issues such as disability, and inclusion – are missing and, as a consequence of it, classified by SentiStrength as neutral (0). In this respect, the sentiment strength dictionary v5 had to be fundamentally revised for a sentiment analysis of public communication in the social and political sphere. In addition, we realized that highly discourse-relevant keywords for our study, which are associated with a positive semantical meaning, have not been included in the old sentiment strength dictionary file v5. Keywords in our study with positive meaning are such as *Inklusion* (‘inclusion’), *Teilhabe* (‘participation’), and *Barrierefreiheit* (‘accessibility’) and the adjective *barrierefrei* (‘accessible’). After having realized the poorly trained vocabulary of SentiStrength in German, we developed a register necessary for our Corpus Linguistic Analysis and accomplished the list with – from the point of view of our study of Computer-Mediated-Communication on disability and inclusion – ‘missing’ words. Therefore, we carried out a corpus-linguistic analysis of the Lexicon key to the discourse on disability and inclusion on Twitter along the Hashtags #Inklusion (‘inclusion’) and #Behinderung (‘disability’), which built the basis for detecting the key vocabulary in the corpus. We were able to develop a main register for the Sentiment Analysis with SentiStrength only after carrying out the detection of the main vocabulary by using AntConc. In this way, we could accomplish our register with the most important lexemes highly relevant to the discourse on disability and inclusion.

### 5.2 Findings of the corpus-driven analysis with AntConc and SentiStrength

A log-likelihood<sup>2</sup> analysis with the corpus linguistic program AntConc of the collocates of the #-words *Inklusion* (‘inclusion’)/*inklusiv* (‘inclusive’) and *Behinderung* (‘disability’)/*behindert* (‘disabled’) shows the lexicon mostly salient in the discourse:

<sup>1</sup> People with chronic disease and/or disability often used ‘COVID’ as a lexeme, also combined with a Hashtag, for protection by governmental regulations.

<sup>2</sup> Standard settings: threshold  $p < 0.05$  (3.84 with Bonferroni), effect measure size: MI, search window span from five words left to five words right.

Collocates of <i>inklusi*</i>	Freq LR	FreqL	FreqR	Likelihood
<b>Inklusion</b> (inclusion)	335	172	163	369.051
Hilfe (help, aid, assistance) <sup>3</sup>	347	11	336	247.531
Deutschland (germany)	366	21	345	238.594
News <sup>4</sup>	358	25	333	215.490
Berlin	322	37	285	169.196
<b>Teilhabe</b> (participation)	223	97	126	111.421
mit (with) <sup>5</sup>	301	164	137	80.026
<b>Barrierefreiheit</b> (accessibility)	149	77	72	68.312
Menschen (humans)	192	93	99	56.250
SARS	18	13	5	38.405
<b>barrierefrei</b> (accessible)	74	47	27	35.526
CoV	20	14	6	34.365
Behinderung (disability)	624	476	148	33.453
Pflege (care)	79	17	62	26.221
Menschenrecht <sup>6</sup> (human right)	29	6	23	20.427

Collocate of <i>behinder*</i>	Freq LR	FreqL	FreqR	Likelihood
Menschen (humans)	732	669	63	781.086
mit (with)	865	797	68	610.112
Deutschland (Germany)	375	23	352	436.349
Hilfe (help)	333	13	320	377.870
News	316	16	300	279.251
Tag <sup>7</sup> (day)	188	165	23	198.911
Berlin	261	27	234	178.352
Behinderung (disability)	144	61	83	162.125
internationalen <sup>8</sup> (international)	65	58	7	83.670

<sup>3</sup> The lexeme ‚Hilfe‘ (help) is mainly used by one of the mostly ‘visible’ actors around disability and inclusion, which is a professional organization. The productivity of this organization in terms of the production of Tweets has an impact on the evaluation of the entire corpus. Other frequently tweeting users – especially people with disabilities without institutional background – however do not use ‘help’ very often.

<sup>4</sup> see comment above

<sup>5</sup> This frequency is related to the frequent usage of the political correct reference ‘Menschen mit Behinderung’ (people with disability).

<sup>6</sup> in contrast of English this expression is a compositum

<sup>7</sup> As part of the phrase ‘Internationaler Tag der Menschen mit Behinderung’ (International Day of People with Disability) which is on the 3d of December and with this key part of our corpus.

Collocate of <i>behinder*</i>	Freq LR	FreqL	FreqR	Likelihood
der <sup>9</sup>	478	340	138	61.514
internationaler <sup>10</sup> (international)	42	36	6	60.523
internationale <sup>11</sup> (international)	39	35	4	51.498
Welttag (World Day)	35	30	5	48.299
von (of)	210	159	51	40.894
es (it, e.g. in <i>es braucht</i> = <i>it is necessary</i> , also there: <i>es gibt</i> = <i>there is</i> )	48	12	36	38.631
vielen (many)	5	2	3	35.567
Gesundheit (health)	31	19	12	35.339
<b>Inklusion</b> (inclusion)	701	171	530	31.376
ich (I)	36	5	31	31.161
Corona (COVID 19)	116	68	48	29.127
SARS	12	7	5	28.751
CoV	14	9	5	24.455
das <sup>12</sup>	80	22	58	23.811
Beschäftigung (employment)	21	19	2	23.027
veröffentlicht (published)	4	4	0	20.749
<b>Teilhabe</b> (participation)	113	63	50	19.797
Erinnerungen (memories)	8	6	2	19.710

The words in bold are the first missing words in SentiStrengths we added to our register to train the program sensible for our discourse on disability and inclusion. In terms of these lexical findings, we would like to point out that the corpus-linguistic program AntConc recognizes all German morphological forms as separate types. Due to the variety of forms of the German adjective inflection with up

<sup>8</sup> The lexeme ‚international‘ occurs in our corpus in many different forms as the German grammar system has a complex flexion system causing many different endings. This causes, however, to frequent appearance of the same word with different endings which are recognized as different findings by programs for processing with language data.

<sup>9</sup> ‚der‘ can be understood as a definite article in German (masculinum), for instance in the collocation ‘**der** internationale Tag’ (the international day), but also the pluralform with genitive, for instance in the collocation ‘der internationale Tag **der** Menschen mit Behinderung’ (The International Day of People with Disability)

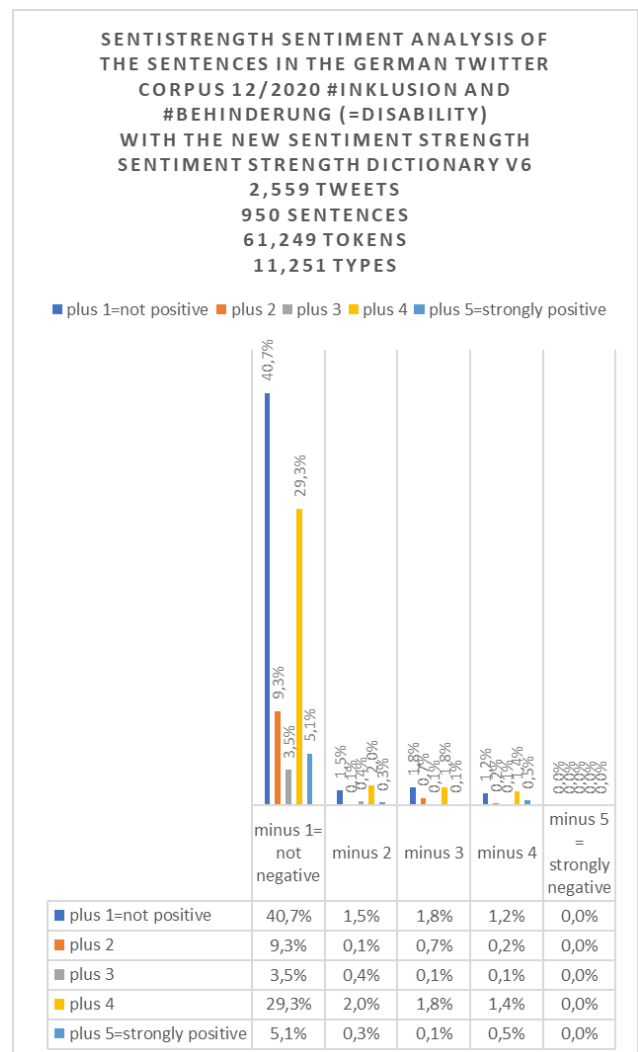
<sup>10</sup> see above, marked with 7

<sup>11</sup> see above, marked with 8

<sup>12</sup> definite article neuter

to 17 endings (including the Ø-ending and the combinations with the endings of comparative forms), the log-likelihood analysis for adjectives is often incorrect. That is why we also added the adjective *inklusiv* ('inclusive'), which has 87 tokens with eight morphological forms in the corpus, to the developed register. We used this findings for checking the language register of SentiStrength and preparing the conduction of a SentimentAnalysis by searching all of these words with SentiStrength, but however, we found out that not only the lexemes with positive ranking such as *Inklusion* ('inclusion')/ *inklusiv* ('inclusive') and *Behinderung* ('disability')/ *behindert* ('disabled') were not recognized by the program but also the most frequent words with a negative sentiment rating in the corpus such as *Exklusion* ('exclusion'), *exklusiv* ('exclusive'), *Diskriminierung* ('discrimination'), and *diskriminierend* ('discriminatory'). As a result of these findings, we trained SentiStrength by adding this vocabulary on the list of the finding to the evaluation list of the sentiment strength dictionary: We rated the positive words *Inklusion* ('inclusion')/ *inklusiv* ('inclusive')/ *Teilhabe* ('participation')/ *Barrierefreiheit* ('accessibility'), and *barrierefrei* ('accessible') with +4 and the negative words *Exklusion* ('exclusion'), *exklusiv* ('exclusive'), *Diskriminierung* ('discrimination'), and *diskriminierend* ('discriminatory') with -4. We decided for these rating values from the point of view of the discourse participants – mainly people with disabilities – as inclusion is essential to the users with disability, while discrimination and exclusion have an enormous negative impact on the lives of many humans with disability. In its default settings, the program SentiStrength will rate these lexemes with +4 or -4 for single occurrences and with the maximum rate +5 or -5 for multiple occurrences in a sentence. This rating values also help to give an insight into the polarized positions on inclusion, discrimination and exclusion in the analysed discourse.

The following chart shows the results of the sentiment analysis with the corrected sentiment strength dictionary file *EmotionLookupTable\_v6\_fullforms* and the corpus without the # characters:



The overall sentiment rating of a sentence is calculated from the positive and the negative sentiment rating. Overall, a neutral or a positive sentiment rating of the discourse on inclusion can be seen:

**40,7 %** of all 950 sentences are quite **neutral (+1-1=0)**, they have a neutral positive (+1) and a neutral negative (-1) sentiment rating.

**47,2 %** of all 950 sentences have a **positive sentiment without negative sentiments**:

29,3 % of all 950 sentences are **very positive (+4-1=+3)**, they have an highly positive (+4) and a neutral negative (-1) sentiment rating.

9,3 % of all 950 sentences are **slightly positive (+2-1=+1)**, they have a positive (+2) and a neutral negative (-1) sentiment rating.

5,1 % of all 950 sentences are **very very positive (+5-1=+4)**, they have an strongly positive (+5) and a neutral negative (-1) sentiment rating.

3,5 % of all 950 sentences are **positive (+3-1=+2)**, they have a very positive (+3) and a neutral negative (-1) sentiment rating.

**Only 4,5 %** of the sentences show a **negative sentiment rating**:

2,0 % of all 950 sentences are **positive (+4-2=2)**, they have a neutral positive (+1) and a very negative (-3) sentiment rating.

1,5 % of all 950 sentences are **slightly negative (+1-2=-1)**, they have a neutral positive (+1) and a negative (-2) sentiment rating.

1,2 % of all 950 sentences are **very negative (+1-4=-3)**, they have a neutral positive (+1) and a highly negative (-4) sentiment rating.

Some sentences are contradictory regarding their sentiment analysis, e.g.:

1,8 % of all 950 sentences are **confrontative and positive in the result (+4-2=+2)**, they have a highly positive (+4) and a negative (-2) sentiment rating.

These contradictory results of positive and negative sentiment ratings in a sentence are partly due to controversies in the discourse, but above all, they are attributed by the program SentiStrength to sentences with negations of positively rated lexemes, e.g. *keine* (= -2) *Inklusion* (= +4) ('no inclusion').

The corpus-linguistic AntConc analysis has shown that the collocates of the lexemes *Inklusion* ('inclusion') and *Behinderung* ('disability') in the German-language discourse on inclusion are either words that express a high grade ('Hochwertwörter'; e.g. in everyday language *Hilfe* 'help, aid, assistance', constitutional *Menschenrecht* 'human right', socio-political *Teilhabe* 'participation' and *Welttag* 'World Day') or persons (*Menschen* 'humans'), places (e.g. *Berlin* as a metaphor for the German federal government), jargon terms (e.g. social: *Beschäftigung* 'employment' or medical *Corona*) and function words (e.g. *mit* 'with'). The log-likelihood values in the collocation analysis have already given an indication that the inclusion discourse is not as confrontative in style as, for example, the covid discourse (cf. Trost 2023) is, which also depends on the discourse participants: in the discourse on inclusion people with disability, and in the discourse on COVID19 often also members and voters of the Radical-Right-Party (AfD). The results of the log-likelihood collocation analysis shows a positive or neutral framing of the keywords in the entire discourse. A log-likelihood analysis remains a statistical extrapolation, which is the first choice for an analysis of large corpora such as our entire corpus of 5,663,504 tokens. Even with large corpora, a log-likelihood analysis with AntConc does, in contrast to SentiStrength, not lead to a crash of corpus-linguistic tools due to overload. However, only a comprehensive sentiment analysis of each individual lexeme shows whether the log-likelihood analysis is correct for the sentiment classification of other non-keyword-lexemes. Furthermore, our sentiment analysis was able to show in our sub-corpus of a monthly excerpt of the discourse on disability and inclusion that it is positive or neutral on the lexematic level. More negative sentiments occur mainly associated with #Barrierefreiheit (accessibility) as people with disability and their families explore their experience with discrimination and exclusion requiring instead 'accessibility'. We could consider that the main communicative participants in the discourse on disability and inclusion are people with disability, their

families, their allies and their representatives but – very often – also politicians.

While Computational Social Sciences and in general Computational Sciences enable data-driven studies on big data with lexical items, Linguistics can provide a concise analysis of Computer-Mediated-Communication which can also support research in Social Science and Political Science. In reverse, methodological considerations from Digital Linguistics can contribute to the development of programs and tools for data-driven language processing. In addition, a sentiment analysis thus enables the sociolinguistic and the discourse-linguistic to validate log-likelihood values by a detailed analysis of the framing at the level of individual lexemes and sentences.

## 6. Conclusions

In terms of methodological impact, this study suggests that digital linguistic methods can be used to improve the accuracy of the language of social media monitoring but the methodological equipment developed by Computational Science needs to be adapted and sometimes also trained for Corpus Linguistic Studies. Our paper shows that methods of Digital Linguistics and Computational Science, also relevant for Computational Social Science, can be integrated in a research design for the linguistic analysis of digital discourses on disability and inclusion, including discriminatory phenomena such as discrimination and exclusion. For this reason, our study submits itself to the scientific tradition of Corpus Linguistics (e.g. Oussalah et al., 2016; Beißwenger, 2016; Clausen & Scheffler, 2019; Brooks & McEnrey, 2020; Scheffler et al., 2020; Heritage & Baker, 2022; Grieve & Woodfield, 2023 and Computational Social Science (e.g. Brantner & Pfeffer, 2018; Ralev & Pfeffer, 2022; Strathern et al.; 2022) but considers itself as part of the interdisciplinary studies focusing on digitization and communication. For a more concise study of social and political communication and language usage in Social Media, we plead for more interdisciplinary collaborations between Corpus Linguistics as well as Social and Political Science as well as Computational Social Science and, in general, Computational Science. Our descriptions in the methodological part of our study pointed out that SentiStrength is an important program with a high potential to Corpus Linguistics but, unfortunately, its' usability for Corpus Linguistic research studies is strictly limited which could be improved by interdisciplinary collaborations for developing tools and programs for language-based data-processing between Computational Science, Corpus Linguistics and Social as well as Political Science including a vocabulary-based training of programs and tools. One of the most prosperous methodological finding for Linguistics is, however, that an AntConc-analysis combined with an analysis with SentiStrength is eligible for providing valid insights into the semantics of a particular digital discourse.

## 7. References

- Barbaresi, A. (2019): The Vast and the Focused: On the need for thematic web and blog corpora. 7th Workshop on Challenges in the Management of Large Corpora (CMLC-7), 29–32.
- Beißwenger, M. (2017): Empirische Erforschung internetbasierter Kommunikation. (Empirische Linguistik/Empirical Linguistics 9). Berlin, Boston: De Gruyter.
- Brantner, C. & Pfeffer, J. (2018). Content Analysis of Twitter - Big data, big studies. In: The Routledge Handbook of Developments in Digital Journalism Studies. Abingdon, UK: Taylor & Francis, pp. 79–92.
- Brookes, G. & McEnery, A. (2020): Correlation, collocation and cohesion: A corpus-based critical analysis of violent jihadist discourse In: Discourse and Society. 31, 4, pp. 351–373.
- Clausen, Y. & Scheffler, T. (2020): A corpus-based analysis of meaning variations in German tag questions: Evidence from spoken and written conversational corpora. Corpus Linguistics and Linguistic Theory. DOI:10.1515/cllt-2019-0060
- Dai, Xiagfend & Marwan Bikdash; Meyer, B. (2017): From social media to public health surveillance: Word embedding based clustering method for twitter classification. DOI: 10.1109/SECON.2017.7925400
- Gnau, Birte C./Wyss, Eva L. (2019): Der #MeToo-Protest. Diskurswandel durch alternative Öffentlichkeit. In: Hauser, Stefan/Opilowski, Roman/Wyss, Eva L. (eds.): Alternative Öffentlichkeiten. Soziale Medien zwischen Partizipation, Sharing und Vergemeinschaftung. Bielefeld: transcript, pp. 131–165.
- Grieve, J. & Woodfield, H. (2023): The Language of Fake News. In Cambridge University Press. Online: <https://www.cambridge.org/core/elements/language-of-fake-news/7B37014A5C0768AEE806167E8ADD5897>, 20.04.2023
- Grue, J. (2014): Disability and Discourse Analysis. London: Routledge.
- Hall, P. (2019). Disability Hate Speech: Interrogating the Online/Offline Distinction. In: Lumsden, K., Harmer, E. (eds) Online Othering. Palgrave Studies in Cybercrime and Cybersecurity. Palgrave Macmillan, Cham. [https://doi.org/10.1007/978-3-030-12633-9\\_13](https://doi.org/10.1007/978-3-030-12633-9_13)
- Harvey, K. (2012): Disclosures of depression: using corpus linguistics methods to interrogate young people's online health concerns International Journal of Corpus Linguistics. 17(3), pp. 349–379.
- Heaton, D. & Clos, J. & Nichele, E. & Fischer, J. (2023): Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. PeerJ Computer Science 9:e1211 <https://doi.org/10.7717/peerj-cs.1211>
- Heritage, F. & Baker, P. Crime or culture? (2022): Representations of chemsex in the British press and magazines aimed at LGBTQ+ men. In: Critical Discourse Studies. 19, 4, pp. 435–453.
- Herrera, L. & Gjosaeter, Terje (2022): Community Segmentation and Inclusive Social Media Listening. Proceedings to the “International Conference on Information Systems for Crisis Response and Management”.
- Knuchel, D. & Bubenhofer, N. (2023): Machine Learning und Korpuspragmatik. Word Embeddings als Beispiel für einen kreativen Umgang mit NLP-Tools. In: Simon Meier-Vieracker, Lars Bülow, Konstanze Marx & Robert Mroczynski (eds.), Digitale Pragmatik (Digitale Linguistik), vol. 1, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 213–235. [https://doi.org/10.1007/978-3-662-65373-9\\_10](https://doi.org/10.1007/978-3-662-65373-9_10)
- Kiritchenko, S. & Zhu, X. & Mohammad, S. (2014): Sentiment analysis of short informal texts, Journal of Artificial Intelligence Research, Volume 50, August 2014, Pages 723-762.
- Lui, M. & Baldwin, T. (2014): Accurate Language Identification of Twitter Messages. In Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) @ EACL 2014, pp. 17–25, Gothenburg, Sweden, April 26-30 2014. Association for Computational Linguistics. DOI: 10.3115/v1/W14-1303 <https://aclanthology.org/W14-1303/>, 20.05.2022.
- Marx, K. & Weidacher, G. (2020): Internetlinguistik. Ein Lehr - und Arbeitsbuch. 2. aktualisierte und durchgesehene Auflage. Tübingen: Narr.
- Paknahad, M. & Baker, P. (2016): Resisting silence: Moments of empowerment in Iranian women's blogs. Gender and Language 11(1), pp.77-99. DOI: 10.1558/genl.22212
- Palomino, M. & Aditya Padmanabhan Varma & Gowriprasad Kuruba Bedala & Connelly, A. (2020): Investigating the Lack of Consensus Among Sentiment Analysis Tools. Lecture Notes in Computer Science. Springer Verlag. doi: 10.1007/978-3-030-66527-2\_5

- Oussalah, M, Escallier, B & Daher, D 2016, 'An automated system for grammatical analysis of Twitter messages. A learning task application', *Knowledge-Based Systems*, vol. 101, pp. 31–47  
<https://doi.org/10.1016/j.knosys.2016.02.015>
- Pan, Zhao & Yao-bin Lu, Sumeet Gupta (2014): How heterogeneous community engage newcomers? The effect of community diversity on newcomers' perception of inclusion: An empirical study in social media service. *Computers in Human Behavior*. Vol. 39, pp. 100-111. DOI:10.1016/j.chb.2014.05.034
- Ralev, R. & Pfeiffer, J. (2022). Hate Speech Classification in Bulgarian. *Fifth International Conference on Computational Linguistics in Bulgaria*. Sofia, Bulgaria, pp. 49–58.
- Sinclair, S. & Bramley, G. (2011): Beyond Virtual Inclusion – Communications Inclusion and Digital Divisions. *Social Policy and Society*, Vol. 10, Issue 1, pp. 1 – 11, DOI: <https://doi.org/10.1017/S1474746410000345>
- Strathern, W. & Ghawi, R. & Schönfeld, M. & Pfeiffer, J. (2022). Identifying Lexical Change in Negative Word-of-Mouth on Social Media. *Social Network Analysis and Mining* 12, 59.
- Scheffler, T. & Aktaş B. & Das, D. & Stede, M. (2019):. Annotating Shallow Discourse Relations in Twitter Conversations. In *Proceedings of DISRPT, workshop at NAACL, Minneapolis, USA*.
- Sties, N. (2013): Diskursive Produktion von Behinderung: Die marginalisierende Funktion von Personengruppenbezeichnungen. In: MEIBAUER, J. (eds.): *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion. Linguistische Untersuchungen 6*. URL: <http://geb.uni-giessen.de/geb/volltexte/2013/10121/>, pp. 194–222.
- Trost, I. (2023): Corona als Basis sprachlicher Argumentation an die eigene Nation und andere Nationen – vom Impfnationalismus bis hin zur Public Diplomacy. In: A. Salamurović: *Konzepte der NATION im europäischen Kontext im 21. Jahrhundert*. Stuttgart: Metzler 2023, pp. 311–327.
- Yang, C. & Srinivasan, P. (2014): Translating surveys to surveillance on social media: methodological challenges & solutions. *Proceedings of the 2014 ACM conference on Web Science*, pp. 4–12.  
<https://doi.org/10.1145/2615569.2615696>
- Yurchenko, O. & Ugolnikova, N. (2021): Linguistic Methods in Social Media Marketing. *Proceedings to the “International Conference on Computational Linguistics and Intelligent Systems”*. URL: <https://ceur-ws.org/Vol-2870/paper55.pdf>, 20.05.2022
- Viola, L. & Musolff, A. (2019): *Migration and Media. Discourses about identities in crisis*. Benjamins.
- Wright, D. (2020): The discursive construction of resistance to sex in an online community. *Discourse, Context & Media*, 36: 100402.
- Zelena, A. (2020): The Psychology of Inclusion on New Media Platforms and the Online Communication. *Acta Universitatis Sapientiae, Communicatio*, 7, pp. 54–67. DOI: 10.2478/auscom-2020-0005



# Workflows and Methods for Creating Structured Corpora of Multimodal Interaction

Anne Ferger, Andre Frank Krause, Karola Pitsch

University of Duisburg-Essen

anne.ferger@uni-due.de, karola.pitsch@uni-due.de, andre.krause@uni-due.de

## Abstract

Corpus analysis of computer mediated and/or multimodal interaction can draw on methods of written and spoken corpora, while also providing further information like gaze or walk annotations or sensor-based data like kinect or motion capture or robot log files. We propose a workflow leveraging the developments of both worlds while simultaneously focussing on standard formats and a sustainable way of research data management.

**Keywords:** multimodal interaction, corpus analysis, corpus workflow

## 1. Introduction

In recent times, research on social interaction in the fields of Interactional Linguistics and Conversation Analysis begins to explore ways of linking genuine qualitative approaches with quantitative perspectives, e.g. (Pitsch et al., 2014; Stivers, 2015; Kendrick and Holler, 2017; Rühlemann, 2018; Luginbühl et al., 2021). This is particularly the case in situations in which novel communication technologies – such as robotic systems, smart speakers etc. – are part of the interactional situation and which attempt to act as an autonomous technical co-participant. While the notion of ‘computer mediatedness’ might need some fundamental rethinking to appropriately grasp such constellations (for a start see e.g. Arminen et al. (2016)), in this paper we focus on workflows and methods for creating structured corpora of multimodal interaction which have been developed when preparing a corpus of human-robot-interaction for long-term storage and integration in a scientific data repository (see also Pitsch (2020), Ferger et al. (2023)).

While such corpus work can draw on well-established resources for transcribing and annotating data using timeline- and XML-based tools (such as PRAAT, ELAN etc.), there is less support for the subsequent steps. Specifically, it is currently not straight forward how to best perform quality checks in order to make transcriptions/annotations consistent across different sessions/recordings (anonymized-citation), how to use well-established procedures such as lemmatization, tokenization and part-of-speech tagging on interactional data, how to merge the transcriptions/annotations of individual sessions/recordings into a joint data set and how the resulting table resp. data frame might best be structured. Also, while standardized data models (see the ISO standard 24624:2016 ‘Transcription of spoken language’ Hedeland and Schmidt (2022)) – which are relevant for including data into a corpus infrastructure (such as e.g. ZuMult<sup>1</sup>) – provide analysis and query functionality for verbal data, multimodal annotations, sensor data or robotic logfiles are missing (see Rühlemann (2018)).

Against this background, we propose a workflow for handling the creation, harmonization of inconsistencies in an-

notation, transcription and metadata and analysis of multimodal corpora in this research field taking advantage of corpus linguistic tools and methods as well as leveraging specific multimodal information in the corpora. The workflow focusses on a sustainable and reusable way of handling the data. The scripts and tools for the workflow will be made accessible and distributed open source.

## 2. Corpus Data

We developed and tested our workflows on existing research data on human-robot-interaction (scenario: museum guide) which we are preparing for long-term storage and further use as open data within the framework of an institutional repository, e.g. (Pitsch, 2020). It comprises transcriptions and annotations in the ELAN-XML format (Max Planck Institute for Psycholinguistics, 2020), videos, metadata in XML format and various sensor-based data such as robot logfiles or kinect data.

The rich data material is particularly interesting with respect to ‘human-robot interaction’ and the multidimensionality of the interaction and contains different data formats such as videos, XML-based transcriptions and robot logfiles.

## 3. Workflow for Creating a Structured Corpus

To process the (already existing) corpus data of the human-robot-interaction scenario (sect. 2.), we developed a workflow (Fig.1) which starts (i) from timeline-based transcripts/annotations (here: created in ELAN), needs (ii) to apply methods for quality checks, normalization and harmonization, and (iii) to enrich the verbal data with lemmatization and part-of-speech tagging, and (iv) to gather the individual sessions into a joint corpus, and (v) to create different output formats, so that the data can be used for different purposes: as ISO-standard TEI for long-term archiving, as a dataframe to be analyzed with the computing environment R, and to be included in a corpus environment such as ZuMult (Fandrych et al., 2021) which provides a comfortable graphical user interface.

The overview in Fig.1 is focussed on the transcription and annotation data, and does not include video and other additional data also belonging to the corpus.

<sup>1</sup><https://zumult.ids-mannheim.de/>

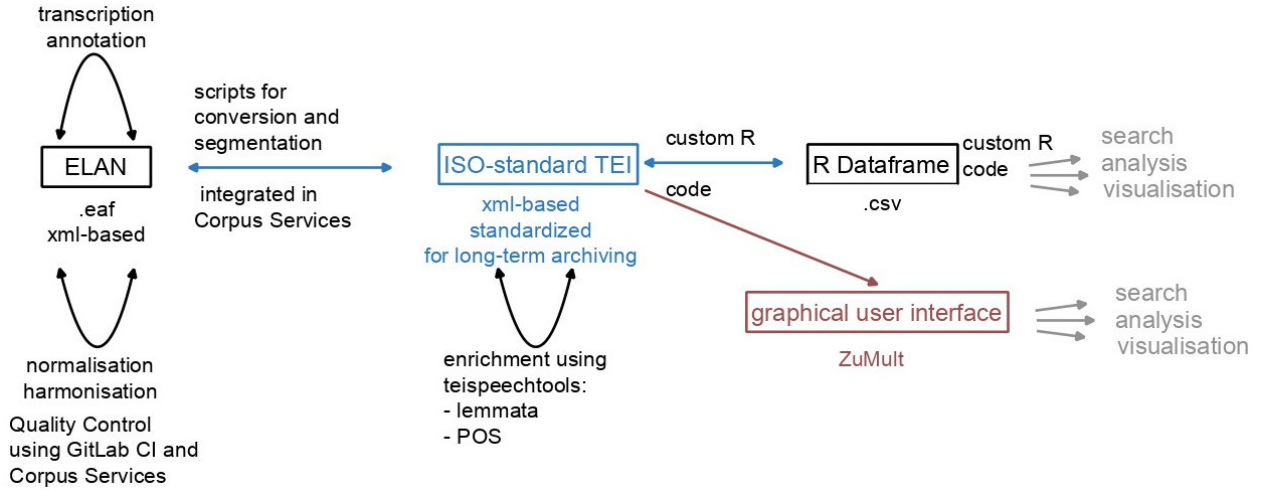


Figure 1: Workflow of corpus creation, harmonization and analysis.

The different data formats shown in Fig.1 are used and needed for different phases of the corpus processing: creation/harmonizing, segmentation/enrichment and visualization/search/analysis. Since the corpus should be analyzed and visualized already during the creation and while new files are added, the steps for the conversion into the other formats are continuously and automatically reapplied using GitLab CI (Ferber et al., 2023).

#### 4. Harmonizing the Source Data

In a first step, the original transcription/annotation files (produced with the timeline- and XML-based tool ELAN) needed to be harmonized, and inconsistencies needed to be fixed to obtain a machine readable corpus source (see Fig. 1). To create a sustainable workflow we developed a script-based GitLab CI setup (Ferber et al., 2023), which offers the benefit that it is automatically re-applied to a source file once new changes have been made. In particular, we aim at keeping the harmonization and fixing of inconsistencies in the source data itself (here: in the ELAN files). This has the benefit of having a one-and-only source file per transcribed/annotated session, to which also manual changes - which are always required to an important degree when ameliorating transcriptions/annotations and solving inconsistencies in corpus data - can be added. Furthermore, the GitLab CI setup makes it easier for new annotations to be added to the corpus and allows for an easier way of data publication and archiving in long-term repositories.

To carry out this task, the Corpus Services Framework (Ferber et al., 2020; Hedeland and Ferber, 2020) was used as it offers functionalities for automatically finding inconsistencies in transcription/annotation files, such as the ELANTranscriptionChecker, which checks characters in transcriptions according to standardized conventions, GAT 2 in our case. We added an automated conversion of ELAN files to EXMARaLDA files so that ELAN-data can be dealt with better and that this setup can be easily reusable with other corpora in future.

#### 5. Converting the source data into TEI and Further Enrichments

To prepare the corpus data for analysis we chose to convert it into a standard format for linguistic data, to allow for archiving of the source data as well as making the workflow interoperable and usable for further use cases. Therefore we chose the TEI<sup>2</sup> format following the ISO standard 24624:2016 ‘Transcription of spoken language’ (see e.g. (Hedeland and Schmidt, 2022)).

To convert the ELAN files into the TEI format we used scripts adapted from the EXMARaLDA software suite (Schmidt and Wörner, 2014) and the Corpus Services Framework. During this conversion the files are segmented following the transcription standard they were transcribed and annotated in, which is GAT 2 (Selting et al., 2009) for the transcription in this case. We added some multimodal annotations, such as gaze or nodding, to the TEI files as well. This goes beyond the current standard for spoken language, but it can still be modelled following its data format. We built this conversion process into the existing Corpus Services Framework (see above section 3.) so that it can easily be distributed and reused.

After the conversion into TEI with our custom conversion, we use the teisspeechtools library<sup>3</sup> which builds on Tree-Tagger<sup>4</sup> for part-of-speech and lemma tagging (Fisseni and Schmidt, 2019).

#### 6. Extending the data model to include multimodal annotations

As the TEI data model has originally been developed for use with verbal language, the data model does not provide any structural resources to include and describe multimodal annotations, such as e.g. gaze, gestures, bodily conduct, facial expressions etc. However, the corpus of human-robot-

<sup>2</sup><https://tei-c.org/>

<sup>3</sup><https://github.com/Exmaralda-Org/teisspeechtools>

<sup>4</sup><https://www.cis.lmu.de/~schmid/tools/TreeTagger/>

#	id	annotation	lemma	pos	type	starttime_ms	endtime_ms	duration	participant_id	utterance_id	file
1	w1	ja	ja	ADV	utterance_SPK_robmus_2015_01_001_W	22400	23700	1300	SPK_robmus_2015_01_001_W	u1	/media/data
2	w2	eins	eins	CARD	utterance_SPK_robmus_2015_01_001_W	48709	49316	607	SPK_robmus_2015_01_001_W	u2	/media/data
3	w3	ähm	ähm	ITJ	utterance_SPK_robmus_2015_01_001_W	84688	85942	1254	SPK_robmus_2015_01_001_W	u3	/media/data
4	w4	sorry	sorry	ITJ	utterance_SPK_robmus_2015_01_001_W	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	/media/data
5	w5	kannst	können	VMFIN	utterance_SPK_robmus_2015_01_001_W	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	/media/data
6	w6	du	du	PPER	utterance_SPK_robmus_2015_01_001_W	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	/media/data
7	w7	wieder[holen]	wieder[holen]	VVINF	utterance_SPK_robmus_2015_01_001_W	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	/media/data
8	w8	hm	hm	NE	utterance_SPK_robmus_2015_01_001_W	113095	114216	1121	SPK_robmus_2015_01_001_W	u5	/media/data
9	w9	tschüss	tschüss	NE	utterance_SPK_robmus_2015_01_001_W	198012	198863	851	SPK_robmus_2015_01_001_W	u6	/media/data
10	inc1	U	N/A	N/A	walk	N/A	8640	N/A	SPK_robmus_2015_01_001_W	inc1	/media/data
11	inc2	-	N/A	N/A	walk	8640	10490	1850	SPK_robmus_2015_01_001_W	inc2	/media/data
12	inc3	U	N/A	N/A	walk	10490	12650	2160	SPK_robmus_2015_01_001_W	inc3	/media/data
13	inc4	-	N/A	N/A	walk	12650	13990	1340	SPK_robmus_2015_01_001_W	inc4	/media/data
14	inc8	~	N/A	N/A	smile	13160	14310	1150	SPK_robmus_2015_01_001_W	inc8	/media/data
15	inc5	U	N/A	N/A	walk	13990	15810	1820	SPK_robmus_2015_01_001_W	inc5	/media/data
16	inc9	@	N/A	N/A	smile	14310	20640	6330	SPK_robmus_2015_01_001_W	inc9	/media/data
17	inc17	prep-G	N/A	N/A	act	17170	17460	290	SPK_robmus_2015_01_001_W	inc17	/media/data
18	inc18	peak-G	N/A	N/A	act	17465	18310	845	SPK_robmus_2015_01_001_W	inc18	/media/data
19	inc19	retr-G	N/A	N/A	act	18310	18810	500	SPK_robmus_2015_01_001_W	inc19	/media/data
20	inc10	~	N/A	N/A	smile	20640	21455	815	SPK_robmus_2015_01_001_W	inc10	/media/data

Figure 2: Dataframe created from TEI for further analysis.

```

<incident end="ts15" start="ts14" type="act" who="SPK_robmus_2015_01_001_W" xml:id="inc17">
  <desc xml:id="des8">prep-G</desc>
</incident>
<incident end="ts17" start="ts16" type="act" who="SPK_robmus_2015_01_001_W" xml:id="inc18">
  <desc xml:id="des9">peak-G</desc>
</incident>
<incident end="ts19" start="ts17" type="act" who="SPK_robmus_2015_01_001_W" xml:id="inc19">
  <desc xml:id="des10">retr-G</desc>
</incident>
<incident end="ts22" start="ts20" type="smile" who="SPK_robmus_2015_01_001_W" xml:id="inc10">
  <desc xml:id="des11"></desc>
</incident>
<annotationBlock end="ts25" start="ts23" who="SPK_robmus_2015_01_001_W" xml:id="aui">
  <u xml:id="u1">
    <w lemma="ja" pos="ADV" xml:id="w1">ja</w>
    <anchor synch="ts25"/>
  </u>
</annotationBlock>
<incident end="ts26" start="ts24" type="smile" who="SPK_robmus_2015_01_001_W" xml:id="inc11">
  <desc xml:id="des12"></desc>
</incident>
<incident end="ts28" start="ts26" type="smile" who="SPK_robmus_2015_01_001_W" xml:id="inc12">
  <desc xml:id="des13"></desc>
</incident>

```

Figure 3: Example TEI of multimodal annotation.

interaction, which we have prepared for integration in an institutional repository and data reuse (see section 2.), contains a range of multimodal annotations. Therefore, we needed to find a way to include these into the existing ISO TEI data model.

For verbal language, the TEI data model provides the elements ‘u’ for utterances consisting of ‘w’ for words with ‘pos’ and ‘lemma’ attributes, see Fig. 3 marked green.

To model the multimodal annotations, such as e.g. walking behavior or smile, we have used the TEI element ‘incident’, which following the TEI guidelines “marks any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication”(TEI Consortium, 2023)<sup>5</sup>. The element ‘incident’ is usually used for audible laughing transcribed in the verbal transcription. The ‘type’ attribute added to the ‘incident’ refers to the level of annotation, such as smile or act, see Fig. 3 marked yellow. The type allows for analysis on specific annotations of the same phenomenon in different settings, such as nod as opposed to smile. This extension will also be added to the Corpus Services Framework for further use.

<sup>5</sup><https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-incident.html>

## 7. Analyzing the Corpus of Multimodal Interaction

One benefit from using ISO-standard TEI as source format for the structured corpus consists in the possibility of easily creating different output formats depending on the required usage. In addition to prepare the corpus data for long-term archiving and for integration into a corpus framework that provides a graphical user interface for easy accessibility, one important goal has been to make the data easily analyzable with the computing environment R and its tools for quantification and visualization. Therefore, we have created an R dataframe as an output format (using custom R code<sup>6</sup>) that is also exported as a table in csv format from the ISO standard TEI files. While we consider the TEI file to be the the source data containing the most information in a standard, machine readable XML format, the csv table allows for easier analysis in different contexts.

The columns of the dataframe are inspired by current approaches in corpus pragmatics and linguistics (Rühlemann, 2020; Ehmer, 2021; Schuer, 2021; Rühlemann and Ptak, Under review). Furthermore, they also include the multimodal annotations and POS and lemma tags where applicable. All content of the dataframe columns is taken directly from the source TEI files and not normalized or changed further.

We are about to begin analyzing the dataframe using R and will briefly suggest two paths which are inspired by and adapted from Rühlemann (2020) and Rühlemann and Ptak (Under review) and are currently still under development. One path makes use of the visualization possibilities of R by creating an annotation density plot as shown in Fig. 4. It visualizes the timeline of the respective transcriptions and annotations per participant (x-axis) and shows for the tiers ‘verbal’, ‘smile’, ‘nod’, ‘act’ and ‘walk’ of participant ‘robmus\_2015\_01\_001\_W’ the moments in time at which some annotation is available (y-axis).

Such a visualization is helpful, on the one hand, for show-

<sup>6</sup><https://git.uni-due.de/mumocorp-open-access/elan-git-example>

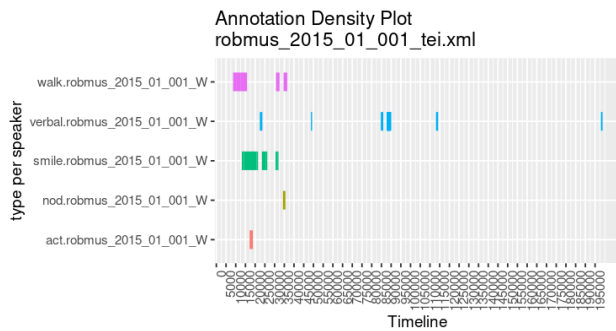


Figure 4: Example analysis of annotations on timeline.

ing at which moment in time certain transcribed/annotated elements exist on each tier and can be used during the work of harmonizing inconsistencies in the corpus. On the other hand, a density plot can also be first, visual access to the data identifying important fragments for analysis.

Another use case is the search of interesting phenomena for further analysis, such as different verbal and multimodal interactions happening simultaneously or overlapping for a certain time. A simple example would be to search for a smile annotation happening simultaneously to a verbal utterance, for which a result containing the respective elements matching the query is shown in Fig. 5.

id	annotation	lemma	pos	type	starttime_ms	endtime_ms	duration	participant_id
w1	ja	ja	ADV	verbal	22400	23700	1300	robmus_2015_01_001_W
inc11	@	N/A	N/A	smile	23465	25630	2165	robmus_2015_01_001_W

Figure 5: Example query result on dataframe.

## 8. Outlook

The development of visualizations and analysis as described above is still ongoing and will lead to further exploration of the corpus integrated into the institutional repository. Future work will deal with including sensor data and metadata in the workflow and in the analysis. This will enable richer searches and visualizations. We will apply and test the workflow on the data of different projects and data types and continue to enhance it this way.

The discussed code and scripts, including a new version of the Corpus Services code for further use is available at <https://git.uni-due.de/mumocorp-open-access/>.

## 9. Copyrights

Proceedings will be published under a Creative Commons Attribution 4.0 International license

## 10. Acknowledgements

This project was financed by the Volkswagenstiftung (grant number 90886, PI: Karola Pitsch).

## 11. Bibliographical References

Arminen, I., Licoppe, C., and Spagnolli, A. (2016). Respecifying mediated interaction. *Research on Language and Social Interaction*, 49(4):290–309.

Ehmer, O. (2021). act: Aligned Corpus Toolkit. R package version 1.2.2, <https://cran.r-project.org/package=act>.

Fandrych, C., Frick, E., Kaiser, J., Meißner, C., Portmann, A., Schmidt, T., Schwendemann, M., Wallner, F., and Wörner, K. (2021). Zumult: Neue zugangswege zu korpora gesprochener sprache. *Sprache in Politik und Gesellschaft: Perspektiven und Zugänge. Jahrbuch des Instituts für Deutsche Sprache*.

Ferger, A., Hedeland, H., Jettka, D., and Pirinen, T. (2020). Corpus Services (version 1.0), <https://doi.org/10.5281/zenodo.4725655>.

Ferger, A., Krause, A. F., and Pitsch, K. (2023). A continuous integration (CI) workflow for quality assurance checks for corpora of multimodal interaction. Accepted at CLARIN Annual Conference 2023.

Fisseni, B. and Schmidt, T. (2019). Clarin web services for tei-annotated transcripts of spoken language. In *Selected Papers from the CLARIN Annual Conference 2019. Leipzig, 30 September-2 October 2019*, pages 12–22. Linköping University Electronic Press.

Hedeland, H. and Ferger, A. (2020). Towards Continuous Quality Control for Spoken Language Corpora. *International Journal of Digital Curation*, 15(1).

Hedeland, H. and Schmidt, T. (2022). The tei-based iso standard ‘transcription of spoken language’ as an exchange format within clarin and beyond. In *CLARIN Annual Conference*, pages 34–45.

Kendrick, K. H. and Holler, J. (2017). Gaze direction signals response preference in conversation. *Research on Language and Social Interaction*, 50(1):12–32.

Luginbühl, M., Mundwiler, V., Kreuz, J., Müller-Feldmeth, D., and Hauser, S. (2021). Quantitative and qualitative approaches in conversation analysis: Methodological reflections on a study of argumentative group discussions. *Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion*, 22:179–236.

Max Planck Institute for Psycholinguistics, The Language Archive, N. (2020). ELAN (version 6.4) [computer software].

Pitsch, K., Vollmer, A.-L., Rohlfing, K. J., Fritsch, J., and Wrede, B. (2014). Tutoring in adult-child interaction: On the loop of the tutor’s action modification and the recipient’s gaze. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 15(1):55–98, June.

Pitsch, K. (2020). Answering a robot’s questions: Participation dynamics of adult-child-groups in encounters with a museum guide robot. Sep.

Rühlemann, C. and Ptak, A. (Under review). Reaching below the tip of the iceberg: A guide to the freiburg multimodal interaction corpus (fremic). *Open Linguistics*.

Rühlemann, C. (2018). *Corpus Linguistics for Pragmatics: A guide for research*. Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2019. | Series: Routledge corpus linguistics guides, 1 edition, October.

Rühlemann, C. (2020). *Visual Linguistics with R: A practical introduction to quantitative Interactional Linguistics*. John Benjamins Publishing Company, Amsterdam, July.

Schmidt, T. and Wörner, K. (2014). EXMARaLDA. In

- Jacques Durand, et al., editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.
- Schuer, T. (2021). ExmaraldaR, <https://github.com/TimoSchuer/ExmaraldaR>.
- Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., et al. (2009). Gesprächsanalytisches transkriptionssystem 2 (gat 2). *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*.
- Stivers, T. (2015). Coding social interaction: A heretical approach in conversation analysis? *Research on Language and Social Interaction*, 48(1):1–19.
- TEI Consortium, editor. (2023). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.6.0*. TEI Consortium, London.



# Balancing expert and peer-student identities in online discussion forums

Francisco Javier Fernández Polo

Department of English and German Philology

University of Santiago de Compostela

xabier.fernandez@usc.es

## Abstract

This paper analyses how students collaborating to improve a translation in online discussion forums construct credibility by projecting an expert image. The analysis focuses on the writing style of three prestige-prominent students, and how they manage to balance the conflicting goals of demonstrating expertise to legitimize their status as advice-givers and asserting their student identities to mitigate imposition. They present themselves as 1) knowledgeable and trustworthy, by using academic and specialized language, adopting a professorial role, citing reliable sources or claiming personal experience; but also 2) as sensitive towards other participants through displays of honesty, humility and in-group solidarity. Their distinct ways of balancing expertise and peer-solidarity arguably explains their relative prominence in the forums rendering their contributions more reliable and acceptable, consequently more worth reading by their colleagues, while also probably securing them better grades. The findings have pedagogical interest for the teaching of academic online discussion skills.

**Keywords:** discussion forums, expert, identity, peer-advice

## 1. Introduction

Issues of identity are central to the study of computer-mediated communication and social media discourse (Locher et al., 2015), where identity is constructed or “performed” by participants in interaction to further their discourse goals. The construction of an expert identity, in particular, plays a key role in many online communication contexts where “there is no pre-configuration of expertise” (Richardson, 2003), and has been especially well established in online peer-to-peer advice situations, notably in health-related online forums (Armstrong et al., 2012; Richardson, 2003; Rudolf von Rohr et al., 2019; Sillence, 2010). Unlike in institutionalized settings, such as doctor-patient interactions, where credibility and trust are automatically granted to the adviser, participants in peer-to-peer advice forums have to gain credibility through displays of expertise. As Richardson (2003) explains, “Participants who offer information and opinion cannot rely upon their reputation (...) the information offered must be formulated with a view to having it accepted as reliable by other participants.” (p. 174-175). Richardson lists several *warranting strategies* employed by non-experts to make their advice more “acceptable”: referring to reliable sources, citing one’s personal experience, referring to one’s own or a friend’s expertise on the matter and using specialized language, an implicit claim to expertise, therefore your credibility as advice-giver. However, giving advice is potentially impolite, especially in peer-to-peer contexts, because it presupposes an asymmetry in the status of the participants, which results in the possibility of the adviser coming across as imposing, vehement or rude. To downplay the inherent face-threat of advice, participants in these forums use various strategies, including a preference for non-directive expressions (Locher, 2013) and various positive and negative politeness devices, like hedges, humour and various forms of expression that construct the advice-giver as a friendly and approachable person (Harvey

& Koteyko, 2013).

Current research into advice discourse spans media, online and off-line, and various personal and professional contexts (Limberg & Locher, 2012), including academic settings such as office hours, where students get advice from teachers (Limberg, 2010; Waring & Hruska, 2012), as well as peer-tutoring, where students support other students in the learning process (Angouri, 2012; Waring, 2005, 2012). In these peer-tutoring sessions, attempts to construct an adviser-advisee relationship are often problematic and met with resistance on the part of the tutees (Waring, 2005). The root of the problem is that there is potential conflict between the student tutors’ expert and peer identities. Tutoring students’ natural response is to try to compensate the participant asymmetry created by the tutoring situation by making language choices that seek to downplay their role as advice givers (Angouri, 2012). While there are some accounts of how these tensions are resolved in face-to-face tutoring, we do not know how this is achieved in online peer-to-peer interaction, for example, in online discussions where students exchange advice to perform collaborative tasks online.

The aim of this paper is to investigate how students negotiate the potential conflict between their role as peers and their role as experts giving advice in a series of online discussion forum. Our aim is to gain insight into the different ways in which a small number of prominent students in these forums build an expert image to gain credibility and legitimize themselves as advice-givers, while, at the same time, strive to assert their student identities by coming over as approachable and solidary. We believe that these students’ ability to strike a balance between their dual identity as experts and student colleagues might account for their prominence in the social network of the participants in these forums. The analysis focuses on the similarities but also on the differences between these students’ online participatory styles.

## 2. Materials and methods

Materials consist of a selection of posts from the SUNCODAC corpus of academic forum discussions (Cal Varela & Fernández Polo, 2020). The context is a blended-learning undergraduate course in translation at a Spanish university. The total number of students enrolled in the course is around 150, about one third being exchange students with various lingua-cultural backgrounds. The working language is English, used as a lingua franca. The core of the forums are the suggestions made by participants in their *feedback posts* for the improvement of another student's (*forum moderator*) translation proposal of a set text, a form of peer-tutoring where student peers give each other advice, comparable to peer-advice, for example, in health-related forums (see above). In SUNCODAC, lecturers open each forum with a post describing the task and close it with a post where they summarize the main points of the debate and appraise and highlight participants' most significant contributions. All feedback posts are graded and count towards the students' final assessment.

The paper provides an in-depth, qualitative analysis of the strategies used by the three most "prestige-prominent" students participating in the forums organized during a one-semester course, to project an expert identity and gain credibility before their peers. Relative prestige in the forums' social network was established with Gephi (Bastian et al., 2009). Gephi is a tool for the study of social networks, to understand their structure and behavior, based on "relational data obtained from different resources, including content available on web pages, user interaction logs and social interaction information provided by users" (Wai & Thu, 2015), among others. Prestige measures in Gephi tally up the number of sending and receiving relations between different nodes in a network, in our case the number of times each participant cited and was cited by others, as well as the relative prestige of the "citing" nodes, a measure of "how well connected is a node to other well connected nodes".

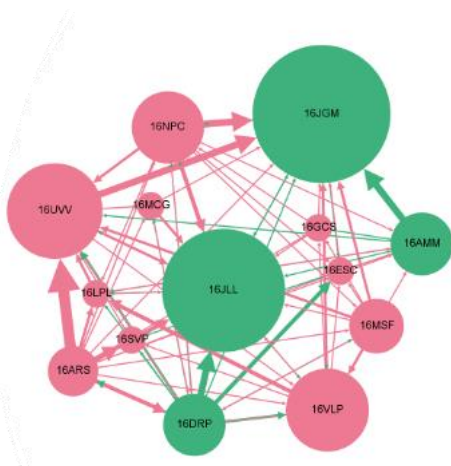


Figure 1. Participant prominence in SUNCODAC using Gephi.

The analyzed corpus consists of the 43 feedback posts (over 10,000 words) produced by these three students, one female (UVV) and two males (JGP and JLL), over the four-

month course period. Our aim is to describe both the variety of strategies they use to construct an expert identity for themselves, and the ways they manage to maintain an equilibrium between their conflicting identities as experts and student peers.

## 3. Findings

The three students use various strategies to construct their expertise into being in the forums and legitimize themselves as credible advisers.

- Citing sources. Citing a source adds trust to your claim. It is a way of shifting responsibility as far as trust is concerned. Obviously, sources are assumed to be reliable. This can be authoritative sources like dictionaries, encyclopaedias, mass media or, simply, general usage. Lecturers are also "citable" authoritative back-up: *As we have seen in class* , *passive constructions are very often in English but it that often in Spanish* [sic].
- Using specialized language. Jargon is strategically used to impress colleagues and lecturers: If you speak like an expert, it is to be assumed that you are an expert. In SUNCODAC, most of the specialized terminology used by students to claim expertise comes from the fields of linguistics and translation studies: *it just sounds better for me in sentence initial position; I decided to make a "cultural adaptation" for the Spanish reader*. In this last example, the very use of the inverted commas marks the expression "cultural adaptation" as an alien code, as the language of specialists. By using their language, the student is presenting him/herself as a credible connoisseur.
- Using formal and academic language. Students use formal words, academic vocabulary and grammar, and essay-like structuring devices like numbering, bulleted lists, etc. to give their posts an aura of sophistication and boost their claims as legitimate advice-givers: *I'd like to make a couple of remarks upon some details; The rarity of this word is probably due to its length*.
- Boasting encyclopaedic knowledge. We generally admire people who know a lot about different topics. Students may display their broad knowledge of different subjects, like geography, Renaissance art or Bible studies, like in the following example, to present themselves as knowledgeable, educated people, whose ideas are worthy of attention, and thus boost their proposals: *I checked the word "unigénito" and it actually has a strong connection with religion. Indeed, it appears explicitly in the Bible (John, 1 : 14)*.
- Behaving as a perceptive observer. Claiming to possess an up-to-date knowledge of the language and special critical skills as language observers may be adduced to legitimize you to tell others what is correct or incorrect in terms of language use: *Although "salir al campo" is not wrong, I think "saltar al campo" is more common and natural in our language*.
- Assessing other participants' work. Assessing or evaluating either the moderator's proposal or other students' suggestions is normally a lecturer's

prerogative. When exerted by a student, it becomes an implicit claim to expertise, presupposing the possession of the knowledge and skills that justifies your right to assess other students' work: *Your translation is **perfectly correct!**; **Awesome translation**, by the way.*

All these warranting strategies (Richardson 2003) were systematically tapped into by these three students to highlight their expert condition and boost their adviser competency. However, these strategies were also carefully balanced in their posts against other forms of expression that, this time, were intended to reinforce the interpersonal relationships with the group by helping mitigate the potentially face-threatening asymmetry inherent to the advice.

- Preference for non-directiveness. This was a tendency observed by Locher (2013) in peer-to-peer health forums, which is also recurrently found in our students' posts. When they make suggestions, they clearly avoid using imperatives and, more generally, any syntactic structure that mentions the advisee explicitly as a recipient of the advice. Non-directiveness is also reflected in the frequent use of hedged expressions intended to soften the imposition, to "downplay dogmatism" (Sillence 2010), e.g., *I know it is the perfect translation (...) but **I would maybe** translate it (...). I know it is a very free and adventurous translation but (...) the tone of the text **may fit** in some "free translations" (...)*
- Giving advice as personal narratives. Narratives may be used to display expertise without creating power imbalance, adding "to the construction of a non-threatening environment" (Kouper, 2010). Arguably, personal narratives reinforce solidarity with the addressee by constructing an identity of the poster as an equal, someone with whom readers share experiences and feelings of satisfaction, frustration, etc.: *I'd like to point out that **a difficult aspect of the translation for me** was to decide (...) **I was not sure** whether (...) . **Eventually, I chose** the first one.*
- Using informal language. The three students downplay authority in their posts by using language that make them appear as approachable (Angouri, 2012; Locher & Hoffmann, 2006), such as informal vocatives and salutations that contribute to relax the tone, or fuzzy expressions (*kind of, sort of, etc.*) that mitigate the stiffness of the academic and specialized language otherwise used to display expertise in different sections of their posts.
- Coming across as understanding and supportive. In general, the three students do a lot of facework in their posts, constantly trying to balance exhibitions of expertise with manifestations of friendship and solidarity. One way of doing this is by portraying themselves as well-wishing and supportive classmates, for example, when they excuse a partner's mistakes: *I think **you know and you are aware** that "Dutch" is not*

*German, but "los holandeses", and it was **obviously just an lapse!***

The three students demonstrate great dexterity in achieving a privileged position among their peers, by constructing authority through recurrent displays of expertise, while emphasizing their student identity and thus preserve a good relationship with their student mates. Their ability to balance these two conflicting goals may explain their prominent position in the forums' social network: their posts are the most frequently read and cited over the semester, and it is reasonable to conclude that there must be something in their writing style that accounts for this success. Actually, each of them has their own idiosyncratic way of achieving prominence in the discussions, by modulating the degree to which they heighten or downplay their expert and student identities in the forums.

### JGM

JGM constructs himself as a competent and legitimate advice-giver mostly, and paradoxically, by downplaying his expert condition, while emphasizing his student identity. In his posts, he strives to sound natural, unpretentious and close. Some of his suggestions are heavily hedged to counter the risky self-attribution of competence inherent to advice-giving, which would place him above his classmates, e.g., ***In my opinion, if I am not wrong**, the author of the text **might** have chosen this verb instead of another, due a sepcial reason (sic).* Additionally, he downplays his expertise by employing very little specialized jargon, while scattering informal language and orality features all over his posts, making him appear approachable and friendly, e.g. ***Congrats again 16UVV and kind regards to all!!).***

In his writing, there seems to be a premediated intention of creating an impression of improvised speech, with its high-involvement features (Chafe, 1982), including constant self-monitoring. He writes as he thinks, without much planning. Vocabulary is sometimes fuzzy and imprecise (*You have **done** it very well!*; *If we look up the meaning of splash, we **get** "salpicar, chapotear"*), and he does not seem to spend much time revising his text before posting it either. His writing, in general, is careless and contains many language issues, in grammar (*a little aspects*), phraseology (*to make word games*), spelling (*an other; sepcial*), haphazard punctuation, cohesion (e.g., there is a point 1 but no 2), etc.

High-involvement is also reflected in the (Chafe, 1982) numerous self-references and reader references in his posts, with either the forum moderator responsible for the draft translation or the group of students participating in the discussion being constantly addressed, directly (*you*) or through solidary, reader-inclusive *we* pronouns. Another way of making readers participate in the text is by posing them questions (*What do you think?*), showing consideration for the readers, by inviting them to express their opinion on the topic, an implicit and polite recognition of their expertise.



## UVV

UVV is the only female in the trio. In her posts, she comes across as a humble, respectful but also extremely conscientious, rigorous, and therefore reliable, advice-giver. She does a lot of facework, stressing the positive aspects of the moderator's proposals (*I think you're right; is correct; could be okay as well*), while downplaying the significance of the identified issues (*this is a minor point*) and the value of her own contribution. On one occasion, she announces her intention to make *only a couple of suggestions*, and then goes on to produce a thorough appraisal of the moderator's proposal, raising no less than seven points that need improving. Her criticism and suggestions tend to be strongly hedged to minimize the importance of her ideas, as illustrated in the following examples: *I'm not really sure about 'cocina campechana'; Perhaps 'adoran' is a bit strong here; I think we should go through 'la parpadelle'; I'd suggest using the expression 'dolce fare niente' here*, etc.

On the other hand, in her posts, UVV also uses some of the characteristic *warranting strategies* (Richardson, 2003) frequently used to demonstrate expertise and boost credibility in advice forums.

UVV's speciality is the citation of websources to back up her suggestions, of which she cites three times as many as the other two students. She makes it a point to support all her ideas with information and examples of usage from carefully selected sources – reference works, the media, etc. – to reinforce the value of her proposals.

She also uses a lot of technical vocabulary to highlight her expertise as a language expert (*connotation, article, agreement, persuasive text, plural noun, singular form*, etc.) therefore her competence as an advice-giver on language usage topics. She also makes frequent use of metatextual devices, like bullet points or numbered lists, to structure her posts and facilitate reading. Bulleted and numbered lists underscore her expert image by presenting her as a knowledgeable person who has a lot to say on the topic. They also contribute to portray her as an orderly painstaking writer, who carefully plans her text in advance, enhancing her image as a reader-friendly, considerate writer.

## JLL

JLL's strategy to build an expert image principally consists in making rhetorical choices that portray him as an authoritative and trustworthy advice giver in the forums.

Unlike JGM, his writing is streamlined, with no redundancies and very few grammar, spelling and punctuation mistakes. His texts are carefully crafted, rich in textual metadiscourse, with text structure, relationships and transitions between ideas clearly signalled, indicating a writer who is both in control and considerate towards readers. His posts seem to have been carefully planned and subsequently revised to ensure that everything is in order. He seems to have a predilection for precise, technical language (*structure, sentence initial position, subjunctive, collocation, inconsistency*), including specialized acronyms that are left unexplained (*As we can see in the DEL: "Reforzar una postura o una condición"*). His

writing style is academic, full of nominalizations (*the rarity of this word is probably due to its length*) and complex sentences. He is also capable of composing complex and well-structured arguments, demonstrating outstanding analytical and reasoning skills only exceptionally found in a second-year undergraduate. His writing style is likely to impress both colleagues and lecturers and will eventually garner respect for his ideas.

In his posts, JLL manages to build for himself an image of an educated, intelligent and self-confident person, who demonstrates extended encyclopaedic knowledge in a variety of topics, ranging from Bible studies to Italian Renaissance art. He projects a strong personal voice, presenting himself as someone who is constantly making decisions (e.g. *I consider that; I decided to*) and displaying critical skills (*an interesting collocation*). He takes up a professorial role when he assesses his colleagues' contributions (*I think there is nothing to be changed in this flawless translation*) or encourages them to think and share their views (*I would like to know your opinion*); he does not hesitate to bluntly criticize a colleague's proposals or even challenge ideas shared by the group (*We talked about the word 'auge' in class, and we agreed that it suggested that English cannot go further than where it is now. But that's not true, so I used 'ascenso' instead, because this word transmits the idea of a progression*). In general, his writing style is forthright and far less hedged than those of the other two students, transmitting confidence and authority.

However, from time to time, he also shows himself as an empathetic person (*The translation [...] is not easy*), capable of demonstrating humility (*I tried to keep*) and using humour (*I wouldn't say 'compartir una vida' because we only have one (I guess)*). And he seeks to relax tensions and reinforce the connections with the group by using informal, fuzzy language (*I would like to make a couple of remarks; it's kind of colloquial; apart from the 'cortex' thing that many of my partners have pointed out*).

## 4. Conclusions

The three students in this study employ multiple strategies to highlight expertise and build up their image as credible advisers: they present themselves as knowledgeable and trustworthy by using academic and specialized language, adopting a professorial role, citing reliable sources and quality examples, displaying encyclopaedic knowledge, claiming personal experience, etc. However, the analysis also reveals another, rather different, image of these students in the forums. They show sensitiveness towards other participants, including the forum moderator, through frequent displays of honesty, humility and in-group solidarity, in the form of reader-inclusive pronouns, disclaimers, self-confessions and humour, among others. Such duplicity arises from the conflict of identities that is enacted in these exchanges, where one must sound "credible, trustworthy and reliable" self (DeCapua & Dunham, 1993, pp. 519), without sounding haughty before their peers.

Each of the students has his/her own idiosyncratic way of

balancing these conflicting goals. JGM's advice-giving strategy consists in writing posts that sound very much like a friendly, informal conversation with the reader, emphasizing rapport, while downplaying expertise. UVV manages to balance expertise and solidarity, by presenting herself as a serious, hard-working and rigorous person, but also as a humble, respectful and well-wishing classmate. Of the three, JLL is the one who puts more emphasis on presenting himself as an expert: he projects a strong personal voice, shows independence of judgment, uses specialized language strategically to underscore his expertise, etc. All three approaches seem to be equally effective as self-promoting strategies: irrespective of their different writing styles, the three students enjoy a most prominent status in the group, receiving a lot of attention and credit from their classmates. The findings should be of practical relevance for the teaching of academic writing skills in computer-mediated settings.

## References

- Angouri, J. (2012). "Yes that's a good idea" Peer advice in academic discourse at a UK university. In H. Limberg & M. A. Locher (Eds.), *Advice in Discourse* (pp. 119–144). John Benjamins Pub. Co.
- Armstrong, N., Koteyko, N., & Powell, J. (2012). "Oh dear, should I really be saying that on here?": Issues of identity and authority in an online diabetes community. *Health (United Kingdom)*, 16(4), 347–365. <https://doi.org/10.1177/1363459311425514>
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), Article 1. <https://doi.org/10.1609/icwsm.v3i1.13937>
- Cal Varela, M., & Fernández Polo, F. J. (2020). Suncodac. A corpus of online forums in higher education. *Nexus*, 2020(02), 44–52.
- Chafe, W. (1982). Integration and involvement in speaking, writing and oral literature. In D. Tannen (Ed.), *Spoken and written language. Exploring orality and literacy*. (pp. 35–54). Ablex.
- DeCapua, A., & Dunham, J. (1993). Strategies in the discourse of advice. *Journal of Pragmatics*, 20, 519–531. [https://doi.org/10.1016/0378-2166\(93\)90014-G](https://doi.org/10.1016/0378-2166(93)90014-G)
- Harvey, K., & Koteyko, N. (2013). Electronic health communication. Peer-to-peer online interaction. In K. Harvey & N. Koteyko (Eds.), *Exploring Health Communication: Language in Action* (pp. 165–187). Routledge.
- Kouper, I. (2010). The Pragmatics of Peer Advice in a LiveJournal Community. In *Language@Internet* (Vol. 7, p. 1). [www.languageatinternet.de](http://www.languageatinternet.de),
- Limberg, H. (2010). *The Interactional Organization of Academic Talk: Office hour consultations*. John Benjamins. <https://doi.org/10.1075/pbns.198>
- Limberg, H., & Locher, M. A. (2012). *Advice in Discourse*. John Benjamins Publishing.
- Locher, M. A. (2013). Internet advice. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of Computer-Mediated Communication* (pp. 339–362). de Gruyter.
- Locher, M. A., Bolander, B., & Höhn, N. (2015). Introducing relational work in Facebook and discussion boards. *Pragmatics*, 25, 1–21. <https://doi.org/10.1075/prag.25.1.01loc>
- Richardson, K. P. (2003). Health risks on the internet: Establishing credibility on line. *Health, Risk and Society*, 5(2), 171–184. <https://doi.org/10.1080/1369857031000123948>
- Rudolf von Rohr, M.-T., Thunherr, F., & Locher, M. A. (2019). Linguistic Expert Creation in Online Health Practices. In P. Bou-Franch & P. Garcés-Conejos Blitvich (Eds.), *Analyzing Digital Discourse* (pp. 219–250). Palgrave-Macmillan. [https://doi.org/10.1007/978-3-319-92663-6\\_8](https://doi.org/10.1007/978-3-319-92663-6_8)
- Sillence, E. (2010). Seeking out very like-minded others: Exploring trust and advice issues in an online health support group. *International Journal of Web Based Communities*, 6(4), 376–394. <https://doi.org/10.1504/IJWBC.2010.035840>
- Wai, K. P., & Thu, E. E. (2015). 3. *Analyzing Social Network using Gephi*. 36–41. <https://meral.edu.mm/records/6529>
- Waring, H. Z. (2005). Peer Tutoring in a Graduate Writing Centre: Identity, Expertise, and Advice Resisting. *Applied Linguistics*, 26(2), 141–168. <https://doi.org/10.1093/applin/amh041>
- Waring, H. Z. (2012). Chapter 5. The advising sequence and its preference structures in graduate peer tutoring at an American university. In H. Limberg & M. A. Locher (Eds.), *Pragmatics & Beyond New Series* (Vol. 221, pp. 97–118). John Benjamins Publishing Company. <https://doi.org/10.1075/pbns.221.07war>
- Waring, H. Z., & Hruska, B. L. (2012). Problematic directives in pedagogical interaction. *Linguistics and Education*, 23(3), 289–300. <https://doi.org/10.1016/j.linged.2012.06.002>

# “Hebrew level: Bibist.”: Online Hebrew language corrections as a tool for “civilized” bashing

Shir Finkelstein & Hadar Netz

Program for Multilingual Education, School of Education, Tel Aviv University

E-mail: [shirmeiro@gmail.com](mailto:shirmeiro@gmail.com), [hadar.netz@gmail.com](mailto:hadar.netz@gmail.com)

## Abstract

Despite the potential for democratic engagement offered by online commenting, research suggests that political discourse within the Israeli online commenting sphere falls short of realizing its democratic potential. A notable example is the presence of language policing practices. Hebrew online comments often display non-standard language forms, occasionally prompting corrections from individuals who adhere to a standard language ideology. It is these instances of correction that serve as the focal point for the present study. Examining interactions containing language corrections drawn from four prominent Israeli Facebook news pages, we compare between interactions that follow posts related to political issues (the judicial reform/coup in Israel) and those that follow posts related to (mostly) non-political matters (celebrities in Israel and abroad). Findings indicate that language corrections are more prevalent in the political context compared to the non-political context. Qualitative analysis suggests that language corrections manifest as a form of supposedly “civilized” and sanctioned bashing. These language corrections are not driven by a genuine concern for the Hebrew language, but rather stem from the desire of (typically) left-wing correctors to establish their intellectual superiority over the (typically) right-wing individuals being corrected, thus contributing to the perpetuation of existing stereotypes prevalent in Israeli society.

**Keywords:** online commenting, language corrections, standard language ideology, political discourse

## 1. Introduction

Since its inception, the Internet has been regarded as a public sphere that encourages democratic participation, potentially eliminating social barriers and amplifying non-hegemonic voices (Dori-Hacohen & Shavit, 2013). This potential extends to linguistic diversity, encompassing not only various languages but also non-standard language varieties (Švelch & Sherman, 2018). However, research indicates that although the Internet is ostensibly democratic, many online platforms perpetuate the same hierarchical structures and social practices found in society (e.g., Dori-Hacohen & Shavit, 2013; Weizman & Dori-Hacohen, 2017). A notable example is the presence of language policing practices. Language policing exists not only in offline communities, such as educational institutions and academia, but is also prevalent in digital communities, such as Facebook (Švelch & Sherman, 2018) and Jodel (Heuman, 2020; Heuman, 2022). These practices often stem from a “standard language ideology” that advocates a prescriptive notion of “correct” language usage (Lippi-Green, 1997). In the realm of digital communication, this ideology has been termed “cyber-prescriptivism” (Schaffer, 2010).

## 2. Language Corrections

Individuals who adhere to a standard language ideology often engage in language corrections, aiming to rectify non-standard forms of language. In everyday conversations, language repair primarily serves the purpose of addressing interactional issues, such as improving intelligibility (Macbeth, 2004). Additionally, early studies in Conversation Analysis (Schegloff et al., 1977) highlight that in everyday conversations, speakers predominantly choose to self-correct, refraining from correcting others. In other contexts, such as educational institutions, the endorsement of a standard language ideology is commonly observed, not only among teachers who correct their

students’ language (e.g., Godley, 2007; Razfar, 2005) but also among students correcting their own language (Netz et al., 2018).

Finally, despite its potential for breaking down social barriers, the Internet has become a space where standard language ideologies can actually flourish. For instance, Švelch and Sherman (2018) conducted a study examining two Facebook pages, one in English and the other in Czech, both dedicated to the concept of “Grammar Nazi”. Originally a derogatory term referring to individuals who excessively police language, these “Grammar Nazi” Facebook communities have appropriated the term to signify language policing as a positive social practice. They collect and share instances of non-standard language use online, primarily for entertainment purposes. In another study investigating standard language ideology in digital communication, Heuman (2020) explored language corrections performed by users of the social network Jodel in Swedish. Heuman (2020) reveals that, in contrast to corrections in everyday conversations, non-standard language forms in Swedish digital communication on Jodel are primarily corrected by others rather than self-corrected. Like Švelch and Sherman (2018), Heuman (2020) also highlights the humorous tone associated with language corrections. However, the primary focus of Heuman’s (2020) study was on the various forms of these corrections, while her exploration of their pragmatic function in context was limited.

## 3. Online Political Discourse

The present study examines Hebrew language corrections within the context of Israeli political digital discourse. Previous research on Hebrew online political discourse has primarily focused on the Israeli online commenting sphere (e.g., Dori-Hacohen & Shavit, 2013; Weizman & Dori-Hacohen, 2017). These studies indicate that despite the democratic potential of online commenting, political

discourse in the Israeli arena is predominantly characterized by what Katriel (2004) termed as a “bashing style”. This type of discourse aims to establish boundaries between opposing right-wing and left-wing factions, thereby conveying a “radical pessimism about the possibility of political debate” (Dori-Hacohen & Shavit, 2013, p. 361).

For instance, Weizman and Dori-Hacohen's (2017) findings demonstrate that comments in response to political opinion editorials on the Israeli website NRG are typically “ethos-oriented,” employing highly emotional language such as aggressive or derogatory slurs targeting the columnist's personality or their political group affiliation, rather than engaging in “logos-oriented” challenges related to the argumentation itself. Moreover, the comments predominantly consist of “ad-personam” attacks, which directly target the columnist, rather than “ad-hominem” challenges questioning their credibility and professional authority (Weizman & Dori-Hacohen, 2017). Within this realm of bashing, leftists often depict rightists as uneducated, occasionally violent, or even fascist individuals, while rightists often portray leftists as elitist, naive, and bordering on delusional (Dori-Hacohen & Shavit, 2013, p. 370).

However, these studies have primarily focused on analyzing online comments that were written in direct response to opinion editorials, often overlooking the broader interactions that take place among the commenters themselves. The present study aims to fill this gap by investigating interactions among Facebook commenters, comparing between interactions that follow posts related to political issues (the judicial reform/coup in Israel) and those that follow posts related to (mostly) non-political matters (celebrities in Israel and abroad). Since online comments are composed by the general public and are typically unedited, they frequently contain non-standard language forms. Occasionally, individuals adhering to a standard language ideology take it upon themselves to correct these non-standard forms. It is precisely these instances of correction that serve as the central focus of the current study.

## 4. The Current Study

### 4.1 Research Questions, Data, and Method

This study seeks to answer the following questions:

- (1) What types of language forms are corrected by commenters on prominent Israeli Facebook news pages?
- (2) Does the context of the interaction (political vs. non-political) influence the frequency of language corrections on these Facebook pages?
- (3) What is the political affiliation of the correctors in political interactions?
- (4) Which identity categories are constructed through the language corrections?
- (5) How are the language corrections received (i.e., are they accepted by the person being corrected or do they generate antagonism)?

To address these questions, we conducted an analysis of two distinct sub-corpora of Facebook data: (1) a political sub-corpus consisting of comments following posts related to the judicial reform/coup in Israel, and (2) a (mostly) non-political sub-corpus of comments following posts related to celebrities in Israel and abroad. Each sub-corpus comprised a random sample of 150 threads, which encompassed both original posts and subsequent comments. These data were collected over a four-month period from March to June 2023, sourced from four major Israeli Facebook news pages: (1) *Ynet*, (2) *News 12*, (3) *News 13*, and (4) *Now 14*. After collecting a total of 300 threads (150 in each sub-corpus), we conducted a manual examination of the data to identify all comments that included language corrections. Altogether, we identified a total of 82 language corrections. For each correction found, we documented the initial comment that contained the use of non-standard language, the comment containing the correction, and any subsequent comments following the correction. Subsequently, we compared the frequency of language corrections in the two sub-corpora. We then carried out a qualitative discourse analysis of the data to gain deeper insights into the nature and dynamics of the language corrections and their implications in both political and non-political contexts.

### 4.2 Findings

As noted above, we identified a total of 82 language corrections. Most corrections were related to non-standard spelling. Less frequently, corrections were made regarding non-standard grammar and punctuation.

As hypothesized, a significant difference was found in the frequency of language corrections between the political and non-political sub-corpora. Specifically, out of the 82 language corrections, the political sub-corpus contained a total of 71 (87%) corrections, whereas the non-political sub-corpus only had 11 (13%) corrections. It is worth noting that certain threads had no instances of language corrections, while others exhibited multiple corrections within a single thread. Out of the 150 threads analyzed in the political sub-corpus, 48 threads included language corrections, whereas in the non-political sub-corpus, 6 out of the 150 threads included language corrections. This difference in the occurrence of language corrections was statistically significant:  $\chi^2(1, N=300) = 39.8374, p < .005$ . These findings support the notion that language corrections are more prevalent in the political context compared to the non-political context.

Interestingly, the few language corrections that were found in the non-political sub-corpus were, in essence, political in nature, as they were related to prominent disputes in Israeli society, including the political division between rightists and leftists and the religious-secular division. For instance, out of the 11 language corrections found in this sub-corpus, 3 were observed in an interaction related to the actress Alona Sa'ar, who, according to the original post, had suffered from depression. Notably, Alona Sa'ar is not only recognized for her acting career; she is also the daughter of Gideon Sa'ar, a politician who formerly served as a minister on behalf of the Likud party and later established

a new party called New Hope after an unsuccessful leadership bid against longtime leader Benjamin Netanyahu. In other words, despite the sub-corpus being intended as non-political and centered around celebrities, Israeli politics seeped in, and it was mainly within this context that language corrections were performed.

As for the political affiliation of the correctors, out of the 71 language corrections performed in the political sub-corpus, 63 (89%) corrections were performed by opponents of the judicial coup, whereas only 8 (11%) corrections were made by supporters who view it as a judicial reform. In other words, in the majority of cases, opponents (generally associated with leftist political views) were the ones correcting the language of proponents (typically associated with rightist views).

Moreover, qualitative discourse analysis revealed that language corrections were commonly accompanied by insults and mockery towards the person being corrected. Specifically, through the language corrections, those who were corrected (typically proponents of the judicial reform, and rightist in their political view) were often portrayed as unintelligent individuals lacking formal education and producing incoherent or even nonsensical written content. Given the derogatory tone, it is unsurprising that language corrections generated antagonism and heightened feelings of animosity between the two opposing camps. For instance, those who were corrected frequently responded with additional insults in return. Moreover, the reactions to language corrections often indicated that the corrections were perceived as condescending, with retorts like "You are not my language teacher."

In fact, there were very few language corrections that did not lead to antagonism. One such occurrence took place in the non-political sub-corpus, during an interaction about the singer Avi Aburomi. A mother commented on a post about Aburomi's success among youngsters, mentioning that she attended his show with her 12-year-old daughter and was amazed by his phenomenal voice. Her comment contained a word written in non-standard spelling, which prompted a language correction. However, the correction itself was an "exposed correction" (Heuman, 2020); i.e., the correct form was framed by an asterisk without any further comments. In response, the mother thanked the corrector and expressed that she had a feeling something was wrong with the spelling of that word. This example illustrates that when corrections are not part of a heated political debate, they do not necessarily evoke antagonism. In the rare instances where corrections are made without any apparent political context, they are more likely to be accepted without generating animosity or hostility.

## 5. Conclusions

In conclusion, our findings shed light on the nature of Hebrew online language corrections, which appear to manifest as a form of supposedly "civilized" and sanctioned bashing. It is evident that these corrections are primarily "ethos-oriented," characterized by emotionally charged and condescending comments, rather than "logos-oriented" challenges related to the argumentation itself. Notably, language corrections are prominently employed within the context of heated political debates. In such instances, corrections do not seem to arise from a genuine concern for the Hebrew language but rather stem from the correctors' desire to assert their intellectual superiority over those

being corrected.

In other words, language corrections are not about language but rather about attempts to demean and belittle others. By employing such language corrections in emotionally charged debates, individuals may further exacerbate tensions and create an atmosphere of hostility rather than promoting constructive discussions. Understanding these dynamics can be valuable in addressing the underlying issues and encouraging more respectful and constructive communication in various online forums and social media platforms.

## 6. References

- Dori-Hacohen, G., & Shavit, N. (2013). The cultural meanings of Israeli tokbek (talk-back online commenting) and their relevance to the online democratic public sphere. *International Journal of Electronic Governance*, 6(4), pp. 361--379.
- Godley, A. J., Carpenter, B. D., & Werner, C. A. (2007). "I'll speak in proper slang": Language ideologies in a daily editing activity. *Reading Research Quarterly*, 42(1), pp. 100--131.
- Heuman, A. (2020). Negotiations of language ideology on the Jodel app: Language policy in everyday online interaction. *Discourse, Context & Media*, 33, pp. 100353.
- Heuman, A. (2022). Trivializing language correctness in an online metalinguistic debate. *Language & Communication*, 82, pp. 52--63.
- Katriel, T. (2004). *Dialogic Moments: From Soul Talks to Talk Radio in Israeli Culture*. Wayne State University Press.
- Lippi-Green, R. (1997). *English With an Accent: Language, Ideology, and Discrimination in The United States*. Routledge.
- Macbeth, D. (2004). The relevance of repair for classroom correction. *Language in Society*, 33(5), pp. 703--736.
- Netz, H., Yitzhaki, D., & Lefstein, A. (2018). Language corrections and language ideologies in Israeli Hebrew-speaking classes. *Language and Education*, 32(4), 350-370.
- Razfar, A. (2005). Language ideologies in practice: Repair and classroom discourse. *Linguistics and Education*, 16(4), pp. 404--424.
- Schaffer, D. (2010). Old wine online: Prescriptive grammar blogs on the internet. *English Today*, 26(4), pp. 23--28.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), pp. 361--382.
- Švelch, J., & Sherman, T. (2018). "I see your garbage": Participatory practices and literacy privilege on "Grammar Nazi" Facebook pages in different sociolinguistic contexts. *New Media & Society*, 20(7), pp. 2391--2410.
- Weizman, E., & Dori-Hacohen, G. (2017). On-line commenting on opinion editorials: A cross-cultural examination of face work in the Washington Post (USA) and NRG (Israel). *Discourse, Context & Media*, 19, pp. 39--48.

# A Corpus study on the negotiation of pronominal address on talk pages of the German, French, and Italian Wikipedia

Carolina Flinz<sup>1</sup>, Eva Gredel<sup>2</sup>, Laura Herzberg<sup>3</sup>

<sup>1</sup>Università degli Studi di Milano, <sup>2</sup>University of Duisburg-Essen, <sup>3</sup>University of Mannheim  
E-mail: carolina.flinz@unimi.it, eva.gredel@uni-due.de, herzberg@uni-mannheim.de

## Abstract

The adequate use of social deixis is highly dependent on the situation and context and has therefore always been at the center of linguistic pragmatics. So far, principles of pronominal address have mainly been modelled with a focus on oral, co-present interaction. The use of pronominal address in computer-mediated communication with its translocal and partially anonymous contexts is still a research gap. In this context, it is particularly interesting that different digital platforms have developed specific customs or netiquettes regarding the appropriate use of address pronouns. This paper asks, from a contrastive perspective, how the appropriate use of address pronouns is negotiated on talk pages of the German, French, and Italian Wikipedia. The corpus study is based on the multilingual Wikipedia corpora of the Leibniz Institute for the German Language.

**Keywords:** Social Deixis, Corpus Linguistics, Wikipedia

## 1. Introduction<sup>1</sup>

The appropriate use of socio-deictic signs is highly dependent on the situation and context and has always been at the center of linguistic pragmatics (cf. Nübling et al. 2017: 205). However, principles of pronominal address have so far been mainly modelled with a focus on oral interaction where speakers are co-present (cf. Kretzenbacher 2010). The use of pronominal address in computer-mediated communication (CMC) with its translocal and (partially) anonymous contexts poses special challenges for writers and has been considered in only a few initial studies (cf. Gredel 2023). This paper aims to fill this research gap by analyzing meta-discourses on pronominal address in the CMC genre of Wikipedia talk pages. With the multilingual Wikipedia corpora of the Leibniz Institute for the German Language, digital language resources are used that allow a contrastive approach to this object of investigation.

The languages German, French, and Italian<sup>2</sup>, which are considered in this paper, each have a binary system of pronominal address containing an “intimate” pronoun (GER: *du*, FR: *tu*, IT: *tu*) and a more “formal” pronoun (German: *Sie*, FR: *vous*, IT: *Lei*). In oral face-to-face interaction, the selection of the appropriate pronoun in each communicative dyad is generally linked to variables such as social status, age, gender, and conversation situation of the interaction partners (cf. Nübling et al. 2017, 205). In CMC, these variables are not always apparent to writers, so the selection of appropriate pronouns must follow other principles. This corpus study focuses on this aspect through the analysis of meta-discourses.

Regarding CMC, it is interesting that different customs or netiquettes for the use of the appropriate address pronouns

have developed on various digital platforms (cf. Gredel 2023). On the multilingual Wikipedia, there are differences between the netiquettes of the considered language versions. However, there is no consensus on these netiquettes, and they are subject to controversial discussions. Based on the Wikipedia corpora of the Leibniz Institute for the German Language, this article explores whether and how writers negotiate the use of address pronouns on Wikipedia talk pages. It also analyses which aspects of the use of pronominal address are being discussed on talk pages of the German, French, and Italian Wikipedia.

## 2. Social Deixis

The concept of socio-deixis focused on here, which is often traced back to Fillmore (1975, 76), is characterized by Levinson as follows: “Social deixis concerns the encoding of social distinctions that are relative to participant-roles, particularly aspects of the social relationship holding between speaker and addressee(s)” (cf. Levinson 1983, 63). Central aspects of pronominal address, which in some cases pose communicative dilemmas for interaction partners, are the nature and timing of the transition from the distance form to the familiar form (cf. Mühlhäusler & Harré 1990, 142f.; cf. Simon 2003, 125). Based on corpus data, it can be shown that the unidirectionality of this transition from the distance form (in German: *Sie*) to the familiar form (in German: *du*) in CMC is not always given (cf. Gredel 2023). In specific situations or in the context of transitions from digital communication to oral interaction, the unidirectionality may be suspended in a communicative dyad, as shown in Example (1): User „Iste“ mentors new authors in Wikipedia. On his user talk page, he is asked by a new author for advice on editing Wikipedia pages and is

<sup>1</sup> The three authors have written the paper jointly. Carolina Flinz is responsible for the data and analyses of Italian, Eva Gredel for German, Laura Herzberg for France. Introduction (§1) and Conclusion and Outlook (§5) were written jointly.

<sup>2</sup> Social deixis can be expressed in Italian by the pronouns of second person singular *tu*, *ti* and plural *voi*, *vi*, or of third person singular *Lei*, *Le* and plural

*Loro* (cf. Milano 2015: 70). The ones used in contemporary Italian are *tu* and *Lei*. The use of *Loro* addressing more than one person is rare while the form *Voi* was used mainly in the past and during the fascist period. It has almost completely disappeared, even if it is being used nowadays, it's restricted to southern regional Italians and the ecclesiastic sphere (cf. Serianni 2000: 7).

addressed by this person with the formal pronoun *Sie*. In the example below, user “Iste” goes directly to the informal *du* by referring to the Wikipedia Netiquette (*WP:DU*), but marks it with an emoticon (*du ;-)*) to mitigate a potential face threatening act (FTA, cf. Brown/Levinson 2007: 60-66). At the same time, he offers the other author, whom he is addressing as a newbie (*Neuling*), to return to the pronominal form of address with *Sie* if he is very uncomfortable with the informal form (*du*) of address:

- (1) Zunächst einmal ist es hier in der Wikipedia üblich, dass die Benutzer sich untereinander duzen (WP:DU); wenn du ;- ) trotzdem gesiezt werden möchtest, dann sag einfach Bescheid. Grundsätzlich ist es durchaus möglich, auch als Neuling einen passablen Artikel zu schreiben, der nicht wieder gelöscht wird, wenn man sich vorher etwas eingelesen hat. (User Iste, 08.11.2011, WUD17/I65.44315)  
*First of all, it is usual here in Wikipedia for users to address each other by the pronoun du (WP:DU); if you ;- ) would still like to be addressed by Sie, then just let me know. Basically, it is possible for a newbie to write a decent article that will not be deleted again if you have read up a bit beforehand.*

In (1), the offer is made to suspend the unidirectionality of the transition from the formal to the informal pronoun in this communicative dyad – contrary to the Wikiquote.

A similar case is also found in the Italian corpus, in which the new user says that he has been used the informal form “tu” from the beginning and asks whether he operated correctly (2). The answer is as follows:

- (2) Per convenzione, wikipedia usa "di default" il tu per non distinguere gli utenti (ricordarsi chi vuole il lei e chi vuole il tu è impossibile, c'è solo da impazzirci) e anche per un discorso che, su wikipedia, non ci sono distinzioni di sorta, cioè un amministratore è importante tanto quanto un utente. Se vuoi che ti do del lei basta chiedere, ma nel lungo periodo dubito fortemente di ricordarmene (però farò uno sforzo). (User Aluong, 6 lug 2006, WUI15/A06.20577)  
*By convention, wikipedia uses 'by default' 'tu' so as not to distinguish between users (remembering who wants 'Lei' and who wants 'tu' is impossible, it's just going crazy) and also for a discourse that, on wikipedia, there are no distinctions whatsoever, i.e. an administrator is just as important as a user. If you want me to call you 'she', just ask, but in the long run I doubt very much that I will remember (but I will make an effort).*

The preference for the informal form is due to two reasons: first, it is practical, i.e. not to confuse and not to distinguish between users, since everyone, both administrators and users are important equals.

The cases in which a user doesn't want the informal form, are usually highly marked and are often thematized metalinguistically, as demonstrated in Example (3) as well.

- (3) Bonjour. Pour commencer, je n'apprécie pas d'être tutoyée quand je ne connais pas. [...]

Bien à vous. --chansonnette [causer avec dame éliane]  
 4 avril 2013 à 17:41 (CEST), WDF15/A31.85510)  
*Hello. First of all, I don't appreciate being on first-name terms when I don't know someone. [...]*  
*All the best.*

In (3), the importance of using the appropriate socio-deictic sign is marked in a discussion page posting. User “chansonnette” starts off her posting with an explicit change of topic, also marked linguistically by “first of all” (*pour commencer*), to openly show her displeasure about the chosen form of addressing. The factor of “unfamiliarity” between her and another user as well as the associated custom of falling back to the formal *you* in similar interactions, e.g., in oral face-to-face interaction, are used as “chansonnette”'s arguments for preferring the formal *you* variant. Interestingly, she also finishes her posting by using “bien à vous” (*all the best*), a very formal form of wishing farewell, again with *vous* underlining her preference for formal addressing signs.

Previous work has also described contexts in which address pronouns are used not reciprocally or symmetrically, but asymmetrically (cf. Clyne et al. 2003). These aspects of pronominal address have predominantly been discussed for oral interaction. This paper is one of the first to use corpus linguistic methods to investigate which rules and principles can be empirically reconstructed for pronominal forms of address in CMC.

Kretzbacher (2010: 3) names a total of four methodological approaches to metalinguistic comments on the topic of social deixis: These are focus group discussions, network interviews, participant observation of uncontrolled public interactions, and observation and questioning of German-language Internet forums. From our point of view, corpus linguistic approaches are particularly suitable because in this way metalinguistic comments from hundreds or thousands of internet users can be considered, whereas in studies with the above-mentioned methodological approaches far fewer informants could be considered (in Kretzenbacher 2010 is n = 190). In the following, the data and the method will now be described in detail.

### 3. Method and Data

The talk pages of Wikipedia share features of CMC genres such as a dialogic structure and an informal writing style with non-standard language (cf. Storrer 2017). There are two types of Wikipedia talk pages, whose data are considered in this study based on the multilingual corpora by the Leibniz Institute for the German Language: article talk pages, where authors negotiate online encyclopedic content and user talk pages, where the contributions of individual authors are discussed. These two types of talk pages will be considered for the study. The metadata for the corpora used are as follows, cf. Table 1:



Language	Article talk pages	User talk pages
GER	373,161,686 (wdd17)	309,390,966 (wud17)
FR	138,068,162 (wdf15)	374,390,445 (wuf15)
IT	52,070,465 (wdi15)	130,067,969 (wui15)

Table 1: Size of the corpora in tokens and corpus abbreviations<sup>3</sup> (DeReKo 2022 in COSMAS II 2022).

To be able to investigate meta-discourses and thus the negotiation of appropriate address pronouns, we use the following search strings when conducting queries in COSMAS II:

- GER: *&siezen* and *&duzen*
- FR: *vouvoyer* and *tutoyer*<sup>4</sup>
- IT: *dare del L/lei* and *dare del tu*<sup>5</sup>

Regarding the topic of social deixis and the relevant variables for choosing the appropriate address pronoun, it should be noted for the language data at hand that Wikipedia authors do not have to disclose their offline identity on Wikipedia. Against this background, guidelines have been developed for Wikipedia, which diverge depending on the language version. For example, in the German Wikipedia, it is recommended that authors generally use the informal *du* form of address (Wikipedia 2023a). In French, the usage of the formal *vous* form of address is greatly reflected upon. There are specific user boxes which can be implemented on a user page that indicate how a user wishes to be addressed (e.g., Wikipedia 2023b for *vouvoyer* boxes). Although there are also users who prefer the informal *tu*, the formal *vous* form of address

still plays a rather important role in Wikipedia user addressing. In several user surveys no consensus could be generated, so both forms of addressing continue to be used depending on a user's preference (Wikipedia 2023c). Investigating whether, and if so, to what extent users explicitly address these conflicting priorities will shed light on the use of the appropriate address pronouns in multilingual CMC environments.<sup>6</sup>

In the Italian Wikipedia there are no explicit guidelines, but the preference of the informal *tu* is deductible. First, there is often the focus on “welcome and inclusion” (cf. Wikipedia 2023d) and secondly, all help and guide pages are written addressing the reader with the informal *tu*. Finally, in the box which summarizes the principles of good communication, the informal *tu* is used (cf. Wikipedia 2023e). In the talk pages (*Bar*), the preferences are addressed as well as the possibility of a survey; however, there seems to be a consensus on the informal *tu*, by leveraging on the fact that Wikipedia is a project between colleagues, and it allows everyone to feel equal, regardless of their social or cultural status (cf. Wikipedia 2023f).

The targeted corpus study focuses on this issue, which will be examined in more detail for each language in Section 4.

#### 4. Negotiation of social deixis on Wikipedia talk pages

In the following, it will be quantitatively demonstrated to what extent corpus hits referring to a meta-discourse on social deixis can be found in the three languages under consideration. For the German language version, it can be shown that both corpora (wdd17 and wud17) contain hits for both search strings (*&siezen* and *&duzen*), cf. Table 2:

<sup>3</sup> The corpus abbreviations read as follows, *wdd17* is the Wikipedia corpus of German (*deutsch*) article talk (*Diskussion*) pages created from a 2017 Wikipedia dump; *wud17* represents the user discussion pages.

<sup>4</sup> All inflected forms were queried in a rather complex REG# (regular expression) search string:

#REG(^tuto(ie|nt|r(a|i(s|(en)?t)s)?i?(ez|ons)(ont)?s)?y(a(i(s|(en)?t)?nt|s(e|nt)s)?i(ez|ons)))?)?â(mes|t(es)?é(es)?|er|è|rent|i?(ez|ons)))\$) oder  
#REG(^vouvo(ie|nt|r(a|i(s|(en)?t)s)?i?(ez|ons)(ont)?s)?y(a(i(s|(en)?t)?nt|s(e|nt)s)?i(ez|ons)))?)?â(mes|t(es)?é(es)?|er|è|rent|i?(ez|ons)))\$).

<sup>5</sup> All inflected forms were queried in a rather complex REG# (regular expression) search string:

#REG(^d(â(nno)?|a(i|n((d|n)o|te)|r(à|a(i|nno)|e(bbe(ro)?|i|mm?o|st(e|i)|te)?|ò)|t(a|e|i|o)|v(a((m|n)o|te|i)?|i|o))?)|e(mmo|s(s(e(ro)?i(mo)?))|st(e|i)|tt(e(ro)?i))|i(a((m|n)?o|te)?|e(d(e(ro)?i))|o|ò)\$) /+w1 del /+w1 (tu oder lei oder Lei).

<sup>6</sup> The multilingualism of the CMC environment *Wikipedia* with its approximately 300 language versions and the numerous interlanguage links between them has two relevant dimensions: On the one hand, there are frequent citations of other language versions or posts in other languages on the discussion pages. In addition, many authors who speak foreign languages do not only contribute to the language version of their mother tongue, but also edit in several language versions at the same time.

Language	Wikipedia name space	Corpus abbr.: search term	Occurrences	pMW <sup>7</sup>	Texts
German	Talk page	wdd17: & siezen	322	0.86	208
		wdd17: & duzen	993	2.66	682
	User talk page	wud17: & siezen	395	1.21	290
		wud17: & duzen	2,052	6.29	1,659
French	Talk page	wdf15: vouvoyer	103	0.75	95
		wdf15: tutoyer	449	3.25	426
	User talk page	wuf15: vouvoyer	200	0.53	181
		wuf15: tutoyer	1,655	4.42	885
Italian	Talk page	wdi15: dare del L/lei	29	0.56	9
		wdi15: dare del tu	61	1.17	61
	User talk page	wui15: dare del L/lei	84	1.61	82
		wui15: dare del tu	372	7.14	308

Table 2: Results of the search queries in the Wikipedia corpora (DeReKo 2022 in COSMAS II 2022).

For the French language version, Table 2 shows that both search strings (*vouvoyer* and *tutoyer*) yield results which, however, differ in their frequency: For both Wikipedia name spaces, i. e. the article talk pages as well as the user talk pages, inflected forms of the informal address *tutoyer* are more frequently discussed than forms of the formal variant *vouvoyer*. It becomes clear that French Wikipedia authors debate the means of addressing with each other; in sum more often on their own talk pages than on the article talk pages.

The Italian language version contains hits for both search strings in both corpora, cf. Table 2. In particular *dare del tu* is more discussed than the formal form variant *dare del Lei/lei*.

## 5. Conclusion and outlook

In all three language versions of Wikipedia considered, there are indications that authors negotiate social deixis – and specifically the pronouns of address – in the sense of a meta-discourse. This contribution shows the extent to which there are differences between the three language versions of Wikipedia. When comparing all three languages, the frequencies of discussing socio-deictic signs meta-linguistically are significantly different between the

German, French and Italian Wikipedia language versions<sup>8</sup>. In both analysed Wikipedia subcorpora, i.e. the Wikipedia article talk pages on the one hand and the article talk pages on the other hand, a greater deal of discussions about addressing styles takes place on the user talk pages, with the informal *you* variant being discussed more frequently than formal *you* variant.

## 6. References

- Brown, Penelope/ Levinson, Stephen C. (2007): Gesichtsbedrohende Akte. In: Herrmann, Steffen/Krämer, Sybille/ Kuch, Hannes (eds.): Verletzende Worte. Die Grammatik sprachlicher Missachtung. Bielefeld: Transcript Verlag, 59–88.
- Clyne, Michael/Kretzenbacher, Heinz/Norby, Catrin/ Warren, Jane: Address in Some Western European Languages. In: Proceedings of the 2003 Conference of the Australian Linguistic Society, 1–10.
- Clyne, Michael/Norby, Catrin/Warren, Jane: Language and human relations: styles of address in contemporary language. Cambridge 2009.
- COSMAS I/II (2022): Corpus Search, Management and Analysis System, <http://www.ids-mannheim.de/cosmas2/>, ©1991-2022 Leibniz-Institut für Deutsche Sprache, Mannheim.
- DeReKo (2022): Deutsches Referenzkorpus / Archiv der

<sup>7</sup> The abbreviation *pMW* stands for *occurrences per million words*. It is a measure of relative occurrence frequencies that are also normalized to a common base (one million current word forms). This allows for comparing frequencies in corpora of different sizes. To calculate pMW values, we need to divide the raw frequency by the total number of words in the corpus and multiply the result by one million.

<sup>8</sup> This holds for testing between the three languages, with the chi-square statistic being 87.5197. The p-value is  $< 0.00001$ . The result is significant at  $p < .05$  for comparing together the frequencies of the formal *you* variant as well as the informal *you* variant between the three languages, with the chi-square statistic being 61.361. The p-value is  $< 0.00001$ . The result is significant at  $p < .05$ , cf. <https://www.socscistatistics.com/tests/chisquare2/default2.aspx>. For each language, the differences in frequencies between the two analysed corpus types, i.e. Wikipedia article talk pages and user talk pages, are significant for the formal *you* variant in German and French, e.g. for the formal *you* variant in German, *Sie*, the difference between the name spaces is significant with the chi-square statistic being 27.5725. The p-value is  $< 0.00001$ . The result is significant at  $p < .05$ ; French: The chi-square statistic is 7.6534. The p-value is .005667. The result is significant at  $p < .05$ ; not for Italian: The chi-square statistic is 0.4735. The p-value is .491403. The result is not significant at  $p < .05$ .

- Korpora geschriebener Gegenwartssprache 2022-I (Release from 08.03.2022).
- Da Milano, Federica (2015): Italian. In: Jungbluth, Konstanze/Da Milano, Federica (eds.): *Manual of Deixis in Romance Languages*. Berlin/Boston: De Gruyter, 59–74.
- Fillmore, Charles J.: *Santa Cruz lectures on deixis* 1971. Mimeo, Bloomington 1975.
- Gredel, Eva (2023): Siezt du noch oder duzt du schon? In: Korpusstudie zum Gebrauch und zur Aushandlung sozial-deiktischer Zeichen auf digitalen Plattformen. In: Meier-Vieracker, S./ Bülow, L./ Marx, K./ Mroczynski, R. (Hg.): *Digitale Pragmatik*. Berlin/ Heidelberg: Metzler, 39–57.
- Levinson, Stephen C.: *Pragmatics*. Cambridge 1983.
- Mühlhäusler, Peter/Harré, Rom: *Pronouns and people: The linguistic construction of social and personal identity*. Oxford/Cambridge 1990.
- Nübling, Damaris/ Dammel, Antje/ Duke, Janet/ Szczepaniak, Renata: *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*. Tübingen 2013.
- Nübling, Damaris/ Dammel, Antje/ Duke, Janet/ Szczepaniak, Renata: *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*. Tübingen 2017.
- Serianni, Luca (2000): Gli allocutivi di cortesia. In: *La Crusca per voi*. N. 20 aprile 2000.
- Simon, Horst J.: Für eine grammatische Kategorie >Respekt<. *Synchronie, Diachronie und Typologie der deutschen Anredepronomina*. Tübingen 2003.
- Storrer, A. (2017). Grammatical Variation in Gespräch, Text und internetbasierter Kommunikation. In M. Konopka & A. Wöllstein (Eds.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, Berlin/New York: e Gruyter, pp. 105–125.
- Wikipedia 2023a: Warum sich hier alle duzen. [https://de.wikipedia.org/wiki/Wikipedia:Warum\\_sich\\_hier\\_alle\\_duzen](https://de.wikipedia.org/wiki/Wikipedia:Warum_sich_hier_alle_duzen)
- Wikipedia 2023b: Utilisateur vouvoient. [https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Utilisateur\\_vouvoient](https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Utilisateur_vouvoient)
- Wikipedia 2023c: Vouvoyer ou tutoyer sur Wikipédia?. [https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Sondage/Vouvoyer\\_ou\\_tutoyer\\_sur\\_Wikip%C3%A9dia%3F](https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Sondage/Vouvoyer_ou_tutoyer_sur_Wikip%C3%A9dia%3F)
- Wikipedia 2023d: Aiuto:Pagina di discussione [https://it.wikipedia.org/wiki/Aiuto:Pagina di discussione](https://it.wikipedia.org/wiki/Aiuto:Pagina_di_discussione)
- Wikipedia 2023e: Wikipedia:Wikiquote <https://it.wikipedia.org/wiki/Wikipedia:Wikiquote>
- Wikipedia 2023f: Wikipedia:Bar/Discussioni [https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Per\\_favore:\\_Datemi\\_il\\_%22Lei%22](https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Per_favore:_Datemi_il_%22Lei%22)

# A Multivariate Register Perspective on Reddit: Exploring Lexicogrammatical Variation in Online Communities

**Frenken, Florian**

RWTH Aachen University  
Kármánstr. 17/19, 52062 Aachen, Germany  
florian.frenken@ifaar.rwth-aachen.de

## Abstract

Even though social media have shaped day-to-day communication for years, the internal linguistic variation associated with these emergent register contexts still remains largely unknown. To fill this gap, the present study evaluates a geometric multivariate approach for this domain by investigating patterns in the visualisation of forty-two lexicogrammatical features derived from systemic functional theory for thirty-three communities on Reddit. The results successfully demonstrate that these subreddits can be interpreted as subregisters of a yet hypothetical macro-register that align with contextual and thus functional differences. Accordingly, this study argues that investigating individual texts rather than broad feature correlation patterns would improve multidimensional analyses of subregisters in general and hybrid web registers specifically. This perspective can not only improve our understanding of language variation at lower levels of instantiation but also hopes to incentivise further research on platform-internal register variation in light of practical implications for context-informed automatic classifications of web documents in functional terms.

**Keywords:** register variation, geometric multivariate analysis, reddit

## 1. Introduction

Over the years, computer-mediated communication (CMC) has continued to play a central role in everyday discourse (Crystal, 2001, 17). As users learn to navigate this new environment, they face continuously evolving communicative contexts to which they must adapt. Having no obvious offline counterpart, social media, in particular, represent the emergent registers of today whose labels are “instantly recognized” (Berber Sardinha, 2018, 126), yet surprisingly, their linguistic characteristics are not well understood despite the wealth of data available (Titak and Roberson, 2013, 235). A particularly salient perspective in this regard is the growing internal variation within these online platforms. One of the most productive examples of this is Reddit, an American social news website where users can submit, rate, and discuss content on various user-created boards called subreddits. With their hierarchical text structure, the resulting threads have clearly conversational character but the interactions are asynchronous and not necessarily linear (Crystal, 2001, 130–151).

In light of the specific rules governing each subreddit, enforced by self-appointed moderators who can ban users and remove contributions, this study argues that these communities represent unique contextual variants of the site. It therefore explores whether subreddits, as user-curated categorisations of web content, are linguistically meaningful, i.e., sufficiently contextually and therefore functionally different that they constitute subregisters of Reddit, as identifiable by systematic clusters in the distribution of their lexicogrammatical features. This perspective can not only improve our understanding of linguistic differences at lower levels of instantiation but also further yet ongoing efforts of “anatomizing” the web (Kilgariff and Grefenstette, 2003, 345) since Reddit demonstrates the benefits of functional categorisation at a smaller scale, allowing users to find content and communities matching their particular interests.

In general, the notion of register refers to groups of texts showing systematic linguistic patterns that are functionally associated with specific situational contexts (Matthiessen, 2019). Previous research on internet registers has so far largely followed the multidimensional approach (MDA) by Biber (1988), consistently identifying dimensions of variation that demonstrate significant overlaps with Biber’s original factors as well as offline registers (Titak and Roberson, 2013; Berber Sardinha, 2014). However, they tend to keep the crucial step from variable contexts to systematic differences in language use rather vague, often operating on face validity of labels, which says little about the actual text types, especially online (Kilgariff and Grefenstette, 2003, 343), and thus hinders generalisation due to the web’s fluidity (Crystal, 2001, 14).

Against this background, Biber and Egbert (2018) enlisted coders to categorise texts based on predefined, perceptually salient, situational characteristics with only minor success due to the inherent hybridity of web registers they encountered. As such, the present work argues that this gap can be best addressed by conceptualising internet registers from a systemic functional perspective (Halliday, 1978), which combines top-down and bottom-up approaches by deriving features for the linguistic analysis from the contextual parameters of a text, independent of the current register landscape. In doing so, this study theorises Reddit as an example of a macro-register according to Matthiessen (2019), comprised of more specific instantiations in a continuum of subregister variation; after all, this notion is already implied in its organisation into subreddits.

Indeed, Liimatta (2019) found evidence of systematic linguistic differences between communities on Reddit despite a noticeable personal bias in the corpus design. However, the low factor loadings as well as variance explained nicely demonstrate that comparing average frequency scores hides more nuanced differences between these presumably more specific texts (Matthiessen, 2019, 20). To combat this

shortcoming and provide an alternative to Biber’s approach, Diwersy et al. (2014) developed the Geometric Multivariate Analysis (GMA) pipeline, a procedure for visualising differences between individual texts rather than broad feature correlation patterns. To interpret the higher-dimensional space of register variation, GMA uses Principal Component Analysis (PCA), which projects the data onto latent dimensions that capture as much of its original variance as possible. The distances in this subspace are meaningful with respect to the linguistic (dis)similarities of the initial feature vectors, revealing more delicate distribution patterns than aggregated group centroids (Neumann and Evert, 2021, 146).

## 2. Method

The corpus for this study was compiled as a subset of the ConvoKit (Chang et al., 2020) subreddit datasets based on categorisations in the r/ListOfSubreddits (2018) wiki. Of course, one community cannot realistically represent the entire userbase of Reddit; still, this list naturally emerged as a community effort and was not elicited according to a predefined schema (cf. Biber and Egbert, 2018), so the categories can be considered authentic, albeit removed from linguistic theory. Due to the transience of web registers, field, tenor, and mode parameters were used to decide which communities from the general content category to include (Halliday, 1978, 62). Where multiple options were feasible, the most popular and basic one took precedence, assuming a lower specificity of its features (Matthiessen, 2019, 30). Only the first 5000 posts before May 4th, 2018, were analysed because they got archived and could therefore be considered finished. To reduce noise, this paper reports on only five of the thirty-three selected subreddits as an example (see Figure 1).<sup>1</sup>

Since GMA regards each text individually, sampling issues do not run as high a risk of under-representing registers with more internal variation, like Berber Sardinha (2014, 86) cautions for MDA. At the same time, this means defining what exactly constitutes one text is a crucial theoretical consideration, especially for the web. The present study equates the notions of thread and text for three reasons. Firstly, the context of situation pertains to the entire thread, not only individual comments, so regarding them separately, like Titak and Roberson (2013, 242), seems arbitrary as they are not merely about a text (like for blogs), but actively co-create the thread and must therefore be considered part of it. Secondly, the producer-user distinction proposed by Berber Sardinha (2014, 83) appears unfounded, considering that every producer is, by definition, also a user (though not vice versa). Thirdly, the statistical measures of per-text frequencies for GMA have been shown to require a minimum of 100 words, or 10 sentences (Neumann and Evert, 2021, 149) – a threshold most threads fail to reach. Ultimately, following this approach resulted in a sample of 74,960 texts, with fewer in subreddits focusing on ancillary language use (e.g., r/DIY).

To extract even complex lexicogrammatical features, all texts were normalised in terms of formatting and tokens

tagged for their part of speech using the CLAWS tagger and C7 tagset (Garside and Smith, 1997). Though not specifically trained on “dirty” web data (Kilgariff and Grefenstette, 2003, 342), a cursory inspection showed no systematic errors that would disproportionately affect its accuracy, not least because Reddit seems unusually concerned with correctness compared to other social media (Crystal, 2001, 45). The ConvoKit corpora were indexed in verticalised format for automatic feature extraction with the CWB platform (Evert and Hardie, 2011) using a query script by Neumann and Evert (2021) whose linguistic operationalisations based on situational parameters enable generalisability. The feature catalogue was slightly adapted to count usernames as proper nouns to replace salutations. Due to high correlations, which may exaggerate effect sizes by measuring the same underlying structures, aggregate adjective counts were removed from consideration. Additionally, titles were disregarded in favour of contractions and URLs as characteristic features on the web. Three other additional features measuring emojis, edits and forms of address were too sparse to include. As a result, the final table consisted of 42 features (see Figure 2), all normalised as relative frequencies with respect to sensible units of measurement (Neumann and Evert, 2021, 150).

PCA relies on correlations to project these features onto new axes that capture their combined variance. Compared to the rather opaque semantic relationships modelled by embeddings (inaccessible here), its deterministic visualisation enables systematic interpretations grounded in register theory at the cost of being sensitive to scaling differences. The raw feature scores showed extreme variation and outliers, so log-transformed z-scores were used to deskew the distributions (Neumann and Evert, 2021, 151). Since higher-dimensional visualisations become increasingly harder to grasp and each PC explains significantly less variance, only the first four components were analysed. Together, they already explained 42.9% of the original data, comparable to Biber and Egbert (2018) and a significant improvement over 17%, achieved by Liimatta (2019) using MDA. Here, only the first two, still accounting for over 30% variance, are described. Due to its overly optimistic group-awareness, a tentative exploration of a Linear Discriminant Analysis, which can reveal more subtle variation (Neumann and Evert, 2021, 46), hid pronounced differences that emerged quite clearly in the PCA, so the study omitted this step of the GMA procedure. All calculations and visualisations were performed in the statistical programming language R (R Core Team, 2021).

## 3. Results

Figure 1 shows a scatter plot of the first two PCs for the five exemplary subreddits, grouped pairwise with PC1 on the y-axis and PC2 on the x-axis, scaled equally so as to be understood as different perspectives on the same three-dimensional space. Within this space, each dot, colour-coded for subreddit, represents one text whose position is determined by its score for the respective PCs. Their potential clustering is analysed based on a dumbbell plot of the feature loadings for PC1 and PC2 (Figure 2), which indicate their relative prominence. The quantitative focus of

<sup>1</sup>The compilation and analysis scripts for the full corpus are available at <https://osf.io/a7m9d/>.

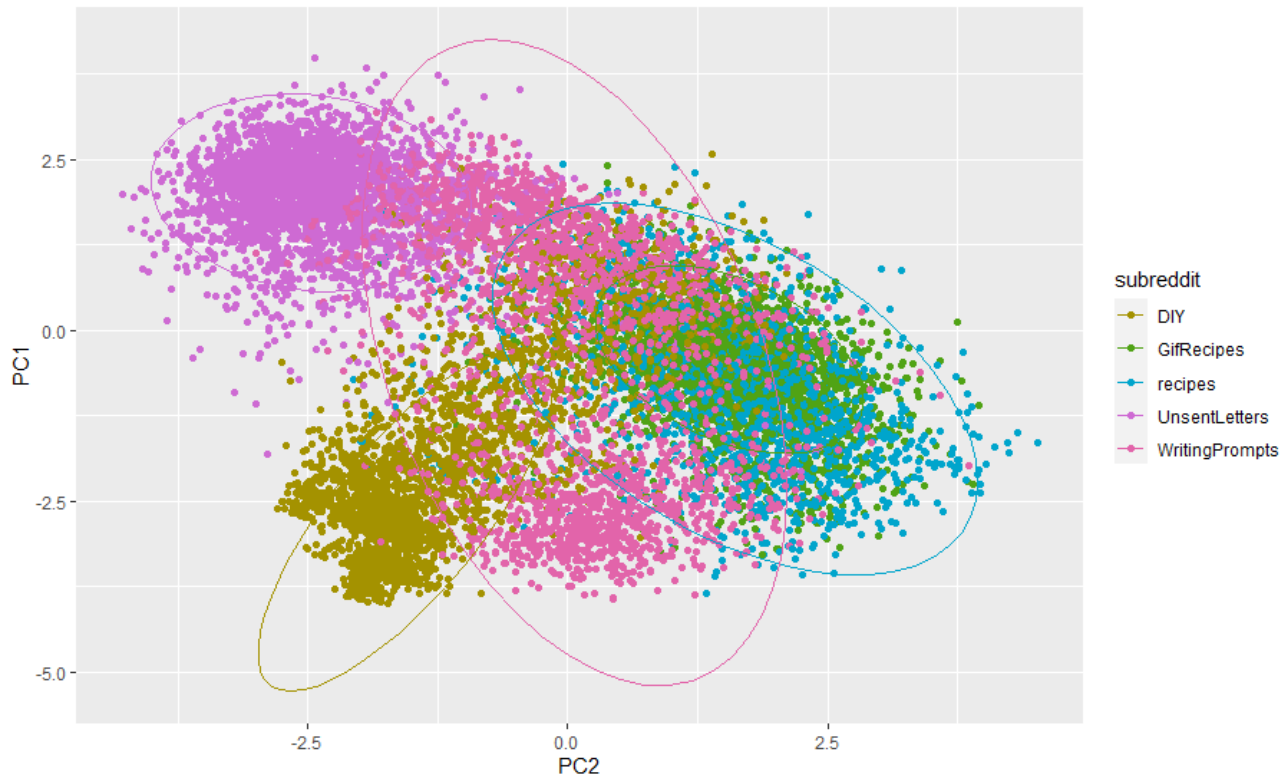


Figure 1: Scatter plot of text scores for PC1 and PC2

this study notwithstanding, the results are enriched with selected qualitative analyses to help ground abstract feature frequencies in their functional expression in concrete texts. To protect the pseudonymity of users, examples reference the unique ID of the post they belong to, which enables replicability but hopefully exacerbates user identification. *r/GifRecipes* is a media-sharing community, so its texts should contain features of ancillary language use, such as URLs and imperatives. Indeed, they strongly favour negative scores on PC1, which are associated with these indicators. Looking at concrete examples reveals that this results from posters including written versions of the recipes shown, which link to the source video and use imperatives to provide step-by-step instructions (e.g., 7jwqig). The list of ingredients, then, also explains the prominence of common nouns, engendering a high lexical density that moves the texts towards positive PC2 scores. As such, they often show strong similarities to their printed counterparts in terms of form and content, as Biber and Egbert (2018, 138) also find. Perhaps unsurprisingly, the functionally similar *r/recipes* also tends towards negative scores on PC1 and the positive side of PC2. Outliers are readily explained by posts that only link to a recipe (e.g., 7s1xcv), or questions (e.g., 7sir6i). In both cases, personal deixis from the comment section will start to dominate, indicating a higher involvement of users.

One would likewise expect *r/DIY* to be characterised by imperatives since submissions should include detailed instructions. However, this subreddit does not seem to be prototypically instructional but rather narrative. Unlike skills and hobbies, located on the side of conceptual writing

in Neumann and Evert (2021), where pronouns are infrequent, their (primarily possessive) forms frequently occur in theme position here because users are discussing their personal projects rather than writing a formal manual. That contractions also contribute positively to the first dimension supports this notion. Help requests, the other type of permissible content on *r/DIY*, contain first and second person pronouns, too, due to being more advisory rather than instructional, again indicating a more involved style (cf. Biber and Egbert, 2018, 57). In contrast, *r/WritingPrompts* generally favours pronoun usage across PCs where they attest a narrative goal orientation, which is consistent with creative writing from the International Corpus of English (Neumann and Evert, 2021, 153). For PC2 especially, there are also texts that demonstrate indicators more in line with Biber and Egbert’s (2018) literate-nominal dimension. It stands to reason that this variation is mainly attributable to differences in the field of discourse.

Still, this does not explain the prominent clusters of texts at the negative end of the first PC. Taking a closer look at concrete texts from *r/WritingPrompts* (e.g., 8efxuk), reveals that they contain moderation messages, often by bots, linking to helpful resources and using repeated imperatives with the thematic discourse marker *please* to instruct users how to avoid future rule violations. They are presumably constantly refined, striving for conciseness and intelligibility, which would explain their somewhat nominal style. As texts move along PC1 towards the positive end, these messages become less frequent, while the number of comments by actual users increases. The difference between the lower and upper cluster of texts, then, is the presence of contribu-

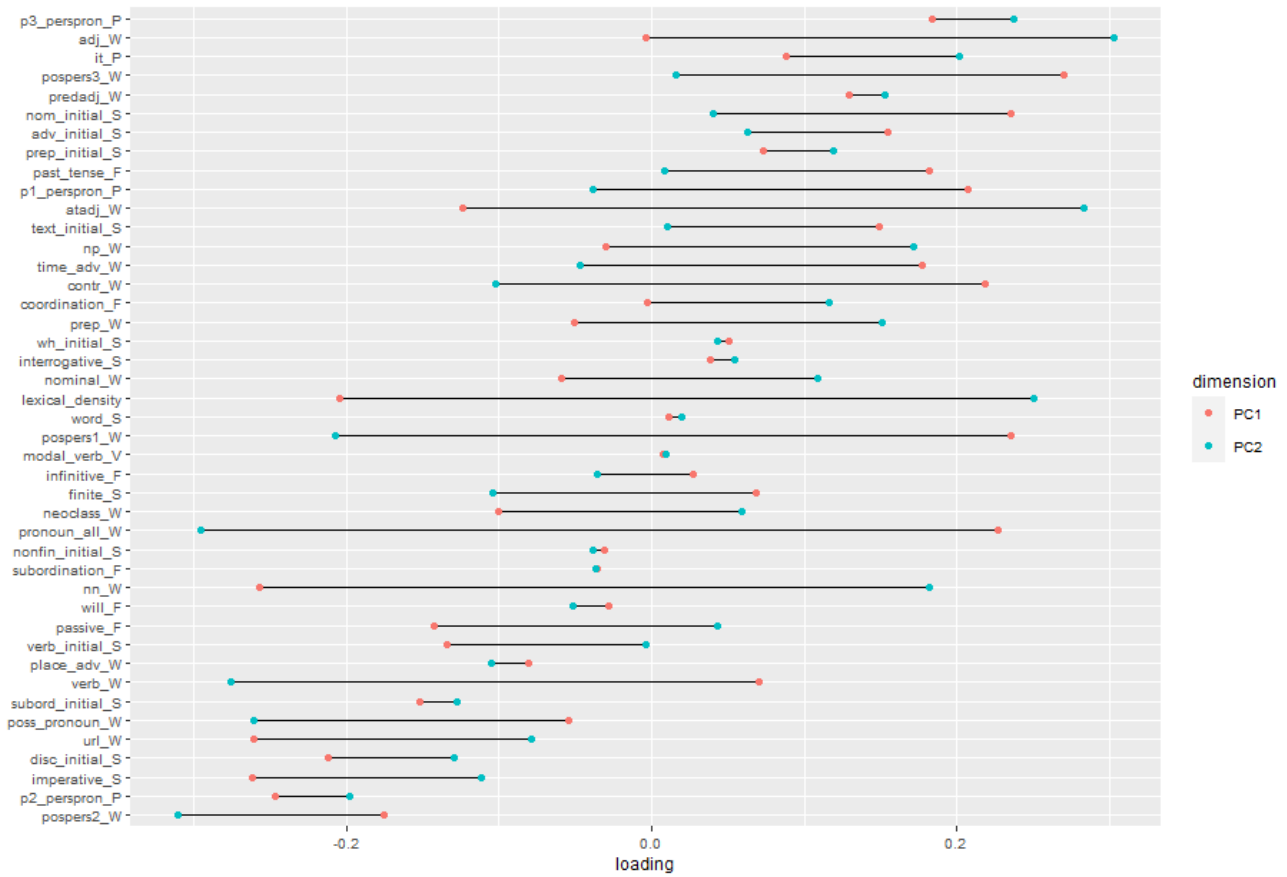


Figure 2: Dumbbell plot of feature loadings for PC1 and PC2

tions by humans. In that case, its distinct features will be gradually overshadowed by the narrative indicators, which are characteristic of the subreddit. In other words, the more interactions by actual users, the further the data points are pushed towards positive feature scores on PC1.

Looking at other selected texts in the bottom cluster of the first PC reveals that this phenomenon roughly occurs below scores of -2 and remains consistent across subreddits. The sub-groupings can be traced back to different kinds of rule violations, which suggests that they can be reliably derived from linguistic indicators alone. The prominent group of r/DIY texts around -3, for instance, predominantly seem to have been moderated because they consisted of only a single image (e.g., 8ajadx). This, then, also explains why only a few hundred texts from this subreddit were too short to include in the analysis compared to over half for most others. Instead of potentially indicating the type of content found in a community, or even specific rules (providing detailed instructions for a project presumably requires a certain number of words, after all), text length may therefore also hint at how actively the community is moderated. In any case, the presence of such messages adds another layer to the already somewhat opaque social relationships online as interactions need not occur exclusively between humans. Lastly, the userbase of r/ListOfSubreddits (2018) categorises r/UnsentLetters as a support community, but the subreddit’s rules expressly forbid unsolicited opinions or advice. Accordingly, its texts lack the indicators of prob-

lem solving in other advice documents, being characterised by first and third rather than second person pronouns (cf. Biber and Egbert, 2018, 128). Outliers on the negative end of PC1 and the positive end of PC2 seem to be primarily letters in other languages (e.g., 8b1vmy). The premise of unsent letters – personal messages that users were too afraid to post – explains this overlap with social letters in Neumann and Evert (2021, 151). At the same time, users frequently narratively reflect on past events in an informal manner, leading to even stronger PC1 loadings due to contractions, past tense, and time adverbs. For example, in the text with the highest positive score, the poster laments: “I wish I didn’t love him anymore. I wish I didn’t care about him anymore. I wish I didn’t need him.” (8h2hxj) Presumably due to the aforementioned rule, comments seem to be rare on this subreddit, so the features of such posts become more pronounced (or rather less blurred), which explains why its texts have such a prominent position, even in the full feature space.

## 4. Discussion

The results reveal that subreddits systematically cluster in terms of their linguistic features, suggesting that they can indeed be considered subregisters of Reddit. Conceptually related communities generally cover similar areas, attesting to a continuous space of variation due to the hybridity of web registers (Neumann and Evert, 2021, 152). Specifically, it seems that the majority of subreddits display fea-



tures of involvement, which is expected of a social media site for discussing specialised interests. The analysis has also demonstrated striking overlaps with offline registers, which valorises online registers as registers proper. Based on salient similarities with other web registers, one could argue, as Biber and Egbert (2018, 42) do for blogs, that Reddit represents a kind of microcosm of the web, viz. the web at large is reflected on a smaller scale within its communities. In a way, subreddits are dense accumulations of web content that also exists elsewhere, which attract interested users with easily understandable and searchable labels that are otherwise hard to find. By demonstrating that they can be differentiated linguistically via computational means, these findings pave the way toward automated functional web categorisation in informational retrieval.

A significant variable unique to the internet, and particularly public discussion forums, that emerged in this study but has so far been ignored in research of register variation on the web is moderation, which shapes the context of online communication not only socially and linguistically since moderators represent the de facto authority over the kind of language permissible in a given community. This has become abundantly clear by the separation of multiple subreddits into moderated and unmoderated texts on the first, most significant PCA dimension. A comparative investigation into the extent to which moderation solely occurs based on violations of conventionalised formal properties of contributions or if such measures also have a linguistic basis could prove valuable. The fact that certain subreddits evaluate submissions based on goal orientations with well-defined linguistic indicators (e.g., whether they entail a narrative element) certainly suggests so. This is especially relevant considering that many subreddits off-load moderation work to bots, a trend that has become increasingly relevant on the internet in recent years. In general, the issue of bots has likewise not yet received due attention in the field of internet linguistics despite important implications for the representativeness of register corpora and opportunities for variation research.

A detailed investigation of lexicogrammatical differences for selected subreddits is required to gain more systematic insights into the patterns of linguistic features engendered by community-specific rules revealed in this explorative study. Choosing the thread as the unit of analysis under the assumption that each of them constitutes a single asynchronous conversation and, by extension, one text, has had significant implications not only in terms of methodological possibilities but potentially also the results overall. Due to the tree-like structure of comment threads, it seems that contextual parameters often operate at lower levels of instantiation, either in local branches or perhaps at the level of individual contributions. This was reflected in the fact that the effects of ratings and other user interactions could not be properly accounted for as part of the tenor of discourse. Any future investigation of the sociolinguistic dynamics on the internet in systemic functional terms presupposes an extensive adaptation of the framework and its operationalisations by considering the characteristic features of CMC. At the heart of this endeavour lies a follow-up study that investigates text at some level below the thread.

## 5. References

- Berber Sardinha, T. (2014). 25 years later: Comparing internet and pre-internet registers. In Tony Berber-Sardinha et al., editors, *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*, pages 81–105. Benjamins, Amsterdam/Philadelphia.
- Berber Sardinha, T. (2018). Dimensions of variation across internet registers. *International Journal of Corpus Linguistics*, 23(2):125–157.
- Biber, D. and Egbert, J. (2018). *Register Variation Online*. Cambridge University Press.
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.
- Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., and Danescu-Niculescu-Mizil, C. (2020). ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60. Association for Computational Linguistics.
- Crystal, D. (2001). *Language and the Internet*. Cambridge University Press.
- Diwersy, S., Evert, S., and Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi et al., editors, *Aggregating Dialectology, Typology, and Register Analysis*, pages 174–204. *Linguae & Litterae*, Berlin.
- Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference 2011*, University of Birmingham.
- Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: Claws4. In Roger Garside, et al., editors, *Corpus Annotation: Linguistic Information From Computer Text Corpora*, pages 102–121. Longman, London.
- Halliday, M. A. K. (1978). *Language As Social Semiotic: The Social Interpretation of Language and Meaning*. Arnold, London.
- Kilgariff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.
- Liimatta, A. (2019). Exploring register variation on reddit: A multi-dimensional study of language use on a social media website. *Register Studies*, 1(2):269–295.
- Matthiessen, C. M. (2019). Register in systemic functional linguistics. *Register Studies*, 1(1):10–41.
- Neumann, S. and Evert, S. (2021). A register variation perspective on varieties of english. In Elena Seoane et al., editors, *Corpus Based Approaches to Register Variation*, pages 143–178. de Gruyter, Berlin.
- R Core Team, (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- r/ListOfSubreddits. (2018). List of subreddits. <https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits>. Accessed 2023-05-15.
- Titak, A. and Roberson, A. (2013). Dimensions of web registers: An exploratory multi-dimensional comparison. *Corpora*, 8(2):235–260.

# Collecting and de-identifying half a million WhatsApp messages

Prakhar Gupta, Lliana Doudot, Romain Loup, Aris Xanthos

University of Lausanne

{prakhar.gupta, lliana.doudot, romain.loup, aris.xanthos}@unil.ch

## Abstract

Instant messaging (IM) applications, especially WhatsApp, have become ubiquitous in contemporary computer-mediated communication practices. IM data have the potential to constitute a rich source of research material for corpus linguistics and cultural analytics, owing to their similarities with face-to-face conversations as well as their private nature. In this work, we outline the creation process of a large curated dataset of WhatsApp messages in French. The paper covers the protocol for collecting these messages as well as the de-identification process for removing sensitive information liable to identify the users in these messages. The de-identified dataset will ultimately be made available to researchers on request.

**Keywords:** WhatsApp, chats, instant messaging, IM, de-identification, corpus, French

## 1. Introduction

Colossal amounts of messages are being exchanged on a daily basis by users of instant messaging<sup>1</sup> (IM) applications such as WhatsApp or Facebook Messenger. These data are of particular interest to language and communication research thanks to features which make them relatively more similar to face-to-face conversations than several other forms of text-based computer-mediated communication (CMC) data (Ueberwasser and Stark, 2017). Due to their private nature, IM exchanges are also likely to offer a privileged viewpoint for studies focusing on the communication of socio-emotional content; for the very same reason, and in spite of their many advantages, IM datasets remain a scarce resource at the time of writing. The present paper reports on the current status of an ongoing effort to collect a large amount of WhatsApp messages and pre-process them in the perspective of sharing them with the scientific community and fostering research on this specific type of CMC data. It focuses in particular on challenges related to the necessity of de-identifying such data, the partly automated and partly manual workflow that was set up to perform this task in an accurate and efficient way, and the results that this method had obtained. The remainder of the paper is organized as follows: section 2 briefly reviews the existing work on IM corpus collection; in section 3 we report basic statistics about our corpus and we present the methodology of the project (concerning data collection and de-identification). In section 4 we present the way that we systematically evaluate our de-identification results and discuss the results we obtain; section 5 offers a brief conclusion and outlines future work directions.

## 2. Related work

In spite of the major role of IM in contemporary CMC communicative practices, IM corpora are vastly underrep-

resented among datasets available for research on CMC, in particular when compared to data documenting mass communication practices and retrieved from social media such as Twitter or from the web (notably discussion forums, blog posts, and comments). As an illustration of this claim, out of 28 corpora listed on the CMC Corpora section of the CLARIN website<sup>2</sup> at the time of writing, only 2 comprise material which qualifies as IM data as defined in section 1 above, although others comprise chat room or SMS data, such as the *SMS2Science* (Dürscheid and Stark, 2011) and *SoNaR* (Sanders, 2012) projects.

The work by Decker and Vandekerckhove (2017) is an early example of a large-scale attempt to collect IM data (along with other types of CMC data). The IM part of the corpus, which was produced between 2007 and 2013 by Flemish youth aged 13-20 on the MSN and Facebook messenger platforms, comprises about 1.3 million words. Arguably, the largest such project is *What's up, Switzerland?* (Ueberwasser and Stark, 2017), which collected more than 600 WhatsApp chats dating from 2010-2014 in a variety of languages spoken in Switzerland (mostly Swiss-German, German, French, Italian and Romansh). This corresponds to about 750K messages and more than 5.5 million tokens which have been made publicly available after their de-identification. At a smaller scale, Verheijen and Stoop (2016) collected 215 WhatsApp conversations (about 330K words) in Dutch, in the perspective of complementing the *SoNaR* corpus with IM data. Dorantes et al. (2018) report on collection of WhatsApp chats in Spanish gathered in Mexico City in 2017, which resulted in a set of 835 chats with more than 1300 informants and a total of about 750K tokens available for linguistic research.

Interestingly, all these data collection efforts have focused on other languages than English. However, aside from *What's up, Switzerland?*, French is not represented in these projects, which is a distinctive feature of the corpus we present. Also, the collections reviewed in this section comprise IM data produced between 2008 and 2017, and as such they do not document the most recent practices in

<sup>1</sup>Instant messaging is understood as a private, quasi-synchronous, and mainly text-based form of CMC, typically occurring between 2 or a relatively limited number of users through such platforms as WhatsApp and Facebook Messenger; in particular, it is distinct from SMS exchange from the point of view of synchronicity, and from chat room communication from the point of view of privacy.

<sup>2</sup><https://www.clarin.eu/resource-families/cmc-corpora>

IM and in particular WhatsApp messaging, which are constantly changing notably in relation to the evolution of the platforms themselves, or to important societal events such as the COVID pandemics (Seufert et al., 2022). Finally, it should be noted that while many more works have sought to gather and analyze sometimes very large amounts of IM data, most of them did not have an explicit goal of sharing the contents of messages with the research community, contrary to the works cited above and the present project.

### 3. Data collection and de-identification

#### 3.1. Data collection

In order to collect data, we host a website<sup>3</sup> for registering the consent of donors and users. This platform is available in the four Swiss national languages (German, French, Italian, and Romansh) as well as English. The platform is also used to collect information pertaining to the chat as well as personal information of the users.

As a preliminary step, prospective chat donors need to register their email-ID on the platform; in doing so, they commit to donating chats only after having requested other chat members' consent to do so. Then they are able to send any number of chats using the email-ID they previously registered. Each donated chat must be exported in plain text format from within WhatsApp (excluding media), then sent to one of the five different email-IDs we use to accommodate the different languages supported on the platform. Donors then receive an automated email reply redirecting them to a form on the website, where they should declare their consent for donating this specific chat under the project's terms and conditions, as well as enter the email address of each other chat member. An automated email is then sent to them, asking them to register their consent in a similar way. Optionally, the chat participants can indicate a list of sensitive words that they want to be redacted in the chat as well as answer a few basic questions about their profile (gender, age class, education level, language skills and use of CMC platforms). Chat donors also have the option to fill in important details about the chat such as the nature of the chat and relationship between the participants.

Each chat member can revoke their consent at anytime and even unilaterally request that the chat be deleted from the platform and our storage. Chats for which one or more users did not register their consent after several reminders are scheduled for deletion in this way at the end of corpus constitution. The chats are encrypted using the Fernet python library<sup>4</sup> which uses the AES 128 encryption standard (Daemen and Rijmen, 2002).

The chat collection campaign ran from August to October 2022. It was promoted by various means, including a press campaign, social media posts, and by word of mouth within the researchers' professional and private networks. In order to incentivize the donation of chats, gift cards with values ranging from 50 CHF to 200 CHF were awarded to a few of the participants selected via a lottery system.

By the end of the collection period, we have collected a total of 72 chats in French with the consent of all members.

We also collected a few chats in Swiss-German, Italian and English (or multiple languages). At this point, however, due to the relatively low number of chats in these languages, we focus our de-identification efforts on French. The French chats for which we have the consent of all members contain around 503K messages including system messages and indications of missing documents and media (image, audio, etc.). The largest chat contains around 177K messages while the smallest contains only 17. Most of the chats contain 2 members while the maximum number of members in a chat is 8. The total number of participants in these chats is 167.

#### 3.2. Data de-identification

We use a combination of automated processes and manual examination steps to detect message fragments that are liable to disclose sensitive or identifying information about the user(s). Similarly to the methodology used by Lungen et al. (2017) among others, these words and phrases are then "categorized", i.e. replaced with abstract placeholders, with the exception of first names; following the practice adopted in the *What's up, Switzerland?* project (Ueberwasser and Stark, 2017), first names are rather "pseudonymized", i.e. randomly replaced with other first names, in order to preserve the data readability. The following subsections describe the automated processes by which we attempt to capture different sensitive information categories as well as the subsequent manual examination steps which we use to improve the de-identification precision and recall.

##### 3.2.1. Automated de-identification processes

**First names and last names.** We use two NER (Named Entity Recognition)-based detection models to detect first names and last names in messages: a multilingual model (Tedeschi et al., 2021) and a French NER model fine-tuned on the CamemBERT model (Martin et al., 2020). All entities tagged as a "PER" (person) are then matched with a set of last names of the permanent resident population of Switzerland<sup>5</sup> as well as first names of the Swiss population by gender<sup>6</sup> with frequency 10 or more. We also include a list of 200 most popular baby names in the United States from 2000 to 2021<sup>7</sup> to introduce some more diversity in the set of first names. According to the matching, words are assigned as first names (also accounting for middle names) or last names. First names are replaced with another name of same gender (male, female or unisex) and beginning with a similar (vocalic vs consonantic) initial letter<sup>8</sup> in the entirety of the chat; last names are replaced with

<sup>5</sup><https://www.bfs.admin.ch/bfs/en/home/statistics/population/births-deaths/names-switzerland.assetdetail.23264628.html>

<sup>6</sup><https://www.bfs.admin.ch/bfs/en/home/statistics/population/births-deaths/names-switzerland.assetdetail.23045212.html>

<sup>7</sup><https://github.com/aruljohn/popular-baby-names>

<sup>8</sup>This is an important constraint, because in French a few words such as "de" (en. "of") are realized in a different way when preceding a vowel-initial word, e.g. "de Paul" ("of Paul") vs. "d'Olivia" ("of Olivia").

<sup>3</sup><https://whatsnew-switzerland.ch/>

<sup>4</sup><https://cryptography.io/en/latest/fernet/>

the “\_LAST\_NAME\_” placeholder.

It is worth noting that the system is implemented in such way that it can recognize variant spelling of names resulting from letter repetition, e.g. “Jooooohn”; in such cases, the replacement name is prepended with a specific marker, e.g. “Marco\_REPETITION\_”. Replacement first names also preserve the original case information of the replaced token.

**Numbers.** Any word containing more than 3 digits is replaced by the “\_NUMBER\_” placeholder. We use regular expressions (regex) to detect and replace such patterns. We also use regex to leave strings indicating date and time untouched wherever possible.

**URLs and email-IDs.** Using regex patterns, all URLs, whether partial or complete, and email-IDs are replaced with the “\_URL\_” and “\_EMAIL\_” placeholders respectively.

**Mentions.** All WhatsApp mentions of chat members using the “@” symbol are replaced with the “\_MENTION\_” placeholder.

**Commune names and street addresses.** All Swiss communes with less than 30,000 inhabitants and all Swiss street addresses are replaced with the “\_COMMUNE\_NAME\_” and “\_STREET\_ADDR\_” placeholders respectively. We set the population threshold to 30,000 so that people living in small communes cannot be traced back using the commune name. Commune names along with corresponding population counts were extracted from the website of the Swiss Confederation<sup>9</sup> whereas street addresses were downloaded from the Swiss Official directory of building addresses<sup>10</sup>.

**User-requested redactions.** In addition to the aforementioned replacements of sensitive information, we also allowed users to submit a list of words and phrases which they deemed sensitive and wanted to be redacted, e.g. nicknames, colloquial ways of naming communes, organization names, etc. Such words and phrases are replaced by the generic “\_MASKED\_TEXT\_” placeholder (unless they were previously replaced/redacted by one of the aforementioned, more specific processes).

**System messages and other media.** All system messages, including references to missing media and documents (which we explicitly asked donors not to export) are replaced by placeholders documenting the nature of the interaction. For example, when one of the members is made an admin, the system message is replaced by “\_ADMIN\_RIGHTS\_MESSAGE\_” or when a image was shared, the message is replaced by “\_IMAGE\_OMITTED\_”.

### 3.2.2. Manual de-identification steps

For each chat, the set of replacements identified by the automated steps described in the previous section is then manually reviewed in order to improve the resulting precision

<sup>9</sup><https://www.bfs.admin.ch/bfs/en/home/statistics/regional-statistics/regional-traits-key-figures/communes.html>

<sup>10</sup><https://www.swisstopo.admin.ch/en/geodata/official-geographic-directories/building-addresses.html>

and recall. The main error types targeted at this point are the following:

- false positives, e.g. common nouns incorrectly identified as names; this category also includes names of celebrities, historical or fictional figures, which do not leak any sensitive information about the users but greatly contribute to the readability of the data
- classification errors, e.g. first names or part of street addresses incorrectly tagged as last names
- granularity errors, such as first names which are actually hypocoristic variants of another first name in the chat, like “Ben” and “Benjamin” for example
- context-dependent errors, which concern the relatively rare occurrence of strings whose status as private information varies from one token to another (e.g. “Max” as a first name or as short for “maximum”).

In addition to these errors, we also attempt to identify false negatives, i.e. items that the automated processes failed to identify altogether—a problem which is both particularly challenging and crucial for privacy protection. To that effect, we mainly used the following two methods:

- out of those words which have not been marked for redaction, we manually review any word beginning with a capital letter and/or not listed in a large machine-readable lexicon (New et al., 2004)
- every word which differs from a redacted word by exactly 1 letter is manually reviewed, which enables us to capture a few instances of sensitive data which automated processes have missed because of typos or non-standard spelling.

Four human experts have been involved in these steps, the result of which is a final set of context-independent replacements for each chat. After these replacements have been automatically performed, the last step in the de-identification process is to deal manually with a few cases which could not be handled in a context-independent way.

## 4. Evaluation of data de-identification

In this section, we discuss the methodology designed to gauge the accuracy of our de-identification workflow and discuss the obtained results.

### 4.1. Evaluation methodology

The evaluation is based on a sample of 3000 snippets randomly drawn from all the French chats in the corpus. It is worth noting that the same sample will also be used for emotion annotation in a later stage of the project (see section 5 below); therefore the criteria used for building this sample are based on two distinct and sometimes conflicting goals (evaluating data de-identification and supporting emotion annotation). In order to compensate for the very large size differences between the chats and to ensure that even smaller ones are represented in the sample, the probability of selecting chat  $i$  is set proportional to  $\sqrt[3]{n_i}$ , where

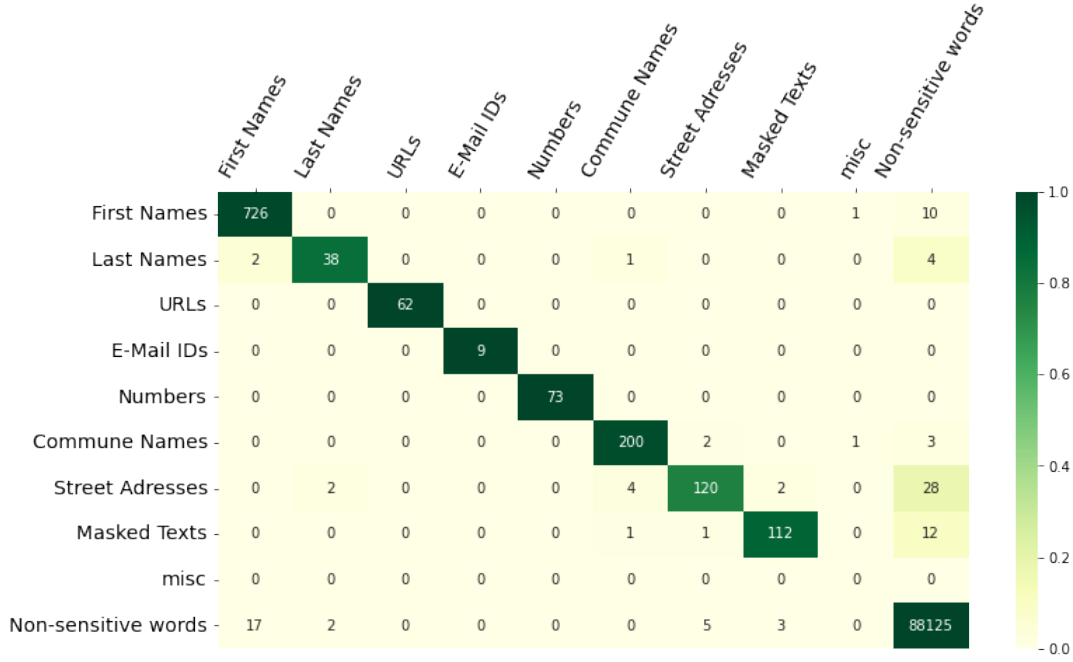


Figure 1: Confusion Matrix showing various errors in our de-identification process. The color-map is indicative of the degree of accuracy. The entry corresponding to  $i^{th}$  row and  $j^{th}$  column corresponds to number of times an entity  $i$  was classified as  $j$ . Thus, the diagonal entries in the matrix represent the correct cases.

Word / Phrase category	Counts	Precision	Recall	Redaction rate
First names	738	97.4%	98.4%	98.6%
Last names	46	89.5%	83.7%	90.2%
URLs	62	100%	100%	100%
Email-IDs	9	100%	100%	100%
Numbers	73	100%	100%	100%
Commune Names	207	97.1%	96.9%	98.3%
Street Addresses	157	93.8%	76.8%	82.2%
Other sensitive information (Masked Text)	126	95.7%	88.9%	90.5%
All sensitive information	1418	96.9%	94.6%	95.9%

Table 1: Precision, recall and redaction rates for different word/phrase categories after de-identification

$n_i$  is the number of messages in the chat. Several additional constraints pertaining to the emotion annotation goal are used to further restrict the random selection of snippets, notably the minimum number of users (2), minimum and maximum number of messages (2–5) and tokens (15–60), maximum proportion of redacted tokens (25%), and maximum duration of exchange (2 hours).

The final set of 3000 chat snippets randomly drawn based on these criteria contains 9994 messages, which corresponds to about 2% of the entire dataset. These have been manually reviewed by 3 human experts in order to determine the expected de-identification output (“gold standard” or “GS”) for each message. GS is then compared with the actual output of the automated and manual processing outlined in section 3.2.

#### 4.2. Error analysis and discussion

1180 messages of our sample contain sensitive information according to GS (11.8% of all messages in sample).

Out of these, 1080 (91.5%) are de-identified correctly using our protocol. In 39 among the remaining 100, sensitive information is actually redacted but with wrong replacements/placeholders. In the remaining 61 messages (5.16% of messages with sensitive information), at least some of the sensitive information has not been redacted completely. Precision and recall scores for each category can be found in Table 1. We also report the redaction rate for each category, i.e. the proportion of words/phrases in that category that have been redacted or replaced with a placeholder from that category or from some other category, thus preventing the leakage of sensitive information in any case.

To analyze the performance on a finer granularity level, the confusion matrix illustrating different error types is shown in Fig. 1.

The lowest recall and redaction rates are reported for the street address category. A close examination of these false negatives shows that most of them do not concern formal street addresses (which are generally well recognized), but

names of public places (restaurants, bars, buildings, etc.), which can be mapped to street addresses and have been treated as such in the GS for that reason. However, leaking the information that an anonymous chat member has been present in such a place at some point in the past appears as relatively benign when compared to leaking a member's street address for instance. Aside from this particular case, all categories have a redaction rate superior to 90%, and even close to 100% for the frequent categories of first names and commune names.<sup>11</sup>

## 5. Conclusion and future work

In this paper, we have reported on the current status of our work to collect and de-identify a large corpus of WhatsApp chats with the perspective of ultimately sharing them with the scientific community. The degree of privacy of such data makes them both particularly interesting for CMC research and particularly challenging to collect and properly de-identify, hence the emphasis on this aspect of our work in this paper. Our evaluation results show that a very large proportion of sensitive information is indeed redacted, in most cases with the correct categorization, thus striking a good balance between the sometimes conflicting goals of protecting participants' privacy and making the data as useful as possible for scientific purposes.

While the recall of our method is overall quite high, it is not perfect: roughly 1 piece of sensitive information out of 20 is being missed and thus left untouched. The impact of these missed items is less when they concern first names: due to the fact that these are pseudonymized, it is very hard to discriminate the occasional occurrence of a person real's first name from the vast majority of first names correctly replaced. The problem is more serious when e.g. a last name is missed, because then it can be immediately distinguished from corresponding placeholder. It was clear from the outset of the project that such risks could not be entirely eliminated, even though we were committed to de-identify the data as diligently as possible, using the most advanced technologies at our disposal. We made this explicit in our privacy policy, which users were requested to read as part of the consent process. That said, an additional layer of privacy protection stems from the fact that the data will not be made publicly available, but only shared on demand with people affiliated to a research institution, who will be requested to comply with strict license terms including using the data solely for scientific purposes, avoiding any disclosure of personal data, not sharing the data with third parties, etc. If researchers granted access to the data abide by these terms, the risk associated with the small proportion of recall errors in the de-identified chats seems tolerable to us.

De-identification is bound to become increasingly important and challenging in the future of CMC research and in particular research on IM, notably with the rise of non-textual information usage in this context. De-identifying

audio, image and video content requires completely different set of skills and methods than text, and a lot of work remains to be done in these directions. Even in the simpler case of text data, the amount of manual work currently involved in a thorough and accurate de-identification workflow is prohibitive when dealing with large amounts of data. We believe that future research in this area would benefit from reflecting on the possibility of having at least part of de-identification performed by donors themselves, using dedicated apps on their smartphones, prior to donation. This would obviously involve a time investment that some potential donors would not agree to make, but this may be counterbalanced by the gain in transparency on the user's part, thus helping establishing the trust relationship which is a core requirement of any project of this kind.

In a near future, the next step of the project will be to run a citizen science campaign to obtain emotion annotation for a sample of chat snippets. These annotations will then be used to train a machine learning algorithm and attempt to generalize the emotion annotation to the entire corpus. Our overarching goal is to dispose of a resource to study the evolution of emotion expression in IM data over time and to share this resource with interested CMC researchers in a way that does not compromise the privacy of our contributors.

## 6. Acknowledgements

This research received funding from the NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40\_180888. The authors thank Leyla Benkai and Andrea Grütter for their collaboration, Elisabeth Stark, Simone Ueberwasser and the *What's up, Switzerland?* team for the experience and resources they shared with us, as well as all the participants who consented to donating their chats to our project.

## 7. References

- Daemen, J. and Rijmen, V. (2002). *The Design of Rijndael*. Springer-Verlag, Berlin, Heidelberg.
- Decker, B. D. and Vandekerckhove, R. (2017). Global features of online communication in local flemish: Social and medium-related determinants. *Folia Linguistica*, 51(1):253–281.
- Dorantes, A., Sierra, G., Donohue Pérez, Tlahulia, Y., Bel-Enguix, G., and Jasso Rosales, M. (2018). Sociolinguistic Corpus of WhatsApp Chats in Spanish among College Students. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 1–6, Melbourne, Australia, July. Association for Computational Linguistics.
- Dürscheid, C. and Stark, E. (2011). sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In Crispin Thurlow et al., editors, *Digital Discourse: Language in the New Media*. Oxford University Press, October.
- Lüngen, H., Beißwenger, M., Herzberg, L., and Pichler, C. (2017). Anonymisation of the dortmund chat corpus 2.1. *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*, 3-

<sup>11</sup>The 100% precision reported for numbers is due to the fact that we never counted them as false positives. Indeed we adopt a strict policy which says that sequences of 3 or more digits should be systematically de-identified, regardless of whether it consists of personal information or not.

- 4 October 2017, Eurac Research, Italy, pages 21 – 24. Bolzano, first edition edition.
- Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.
- New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2 : A new french lexical database. *Behavior Research Methods, Instruments, Computers*, 36:516–524.
- Sanders, E. (2012). Collecting and analysing chats and tweets in sonar. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, pages 2253–2256, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Seufert, A., Poignée, F., Hossfeld, T., and Seufert, M. (2022). Pandemic in the digital age: analyzing whatsapp communication behavior before, during, and after the covid-19 lockdown. *Humanities and Social Sciences Communications*, 140(9).
- Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R. (2021). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Ueberwasser, S. and Stark, E. (2017). What’s up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online*, 84(5):105–126.
- Verheijen, L. and Stoop, W. (2016). Collecting facebook posts and whatsapp chats. In Petr Sojka, et al., editors, *Text, Speech, and Dialogue*, pages 249–258, Cham. Springer International Publishing.



# Acquiring, Analyzing, and Understanding Multimodal TikTok Short Video Data: The Case of Online Sex Worker Visibility Management

Teemu Helenius  
University of Turku  
E-mail: taohel@utu.fi

## Abstract

This paper introduces some key questions that affect how multimodal short video data on TikTok can be accessed, acquired, and analysed. The accompanying research ethical questions will also be highlighted. The issue of data collection is approached in terms of TikTok's platform features that most readily affect how videos are made visible and available to users: audio centrality of content and its delivery via the For You Page (FYP) recommender algorithm. The specific context of data gathering, and multimodal discourse analysis are connected to the visibility management of sex workers on TikTok as affected by the platform's content and visibility moderation. The paper presents work-in-progress approaches to data gathering for building multimodal corpora, multimodal discourse analysis, and research ethics of TikTok videos. Additionally, some early findings of the analysis on online sex worker visibility management on TikTok are presented.

**Keywords:** multimodality, TikTok, platform affordances, content moderation, visibility, social media

## Introduction

The social media platform TikTok presents an interesting opportunity and challenge for multimodal analysis. The highly multimodally complex TikTok short videos incorporate audio, still image, moving image, written text, and hypertext in multiple forms to create very dense and contextually bound multimodality. This paper introduces some of the key technical features of TikTok that structure communication on the platform. They in turn influence both how multimodal short video data can be gathered and analysed as well as how the visibility produced impinges on research ethics questions. These issues are connected to the themes of a larger research project's work in progress first article, where I examine the visibility management practices of sex workers in their promotional videos when they interact with content and visibility management by the platform. Some early findings of these forms of managing visibility are also presented.

## TikTok Platform Affordances and Visibility Management

Social media platform affordances cover possibilities and constraints to users in terms of the technological and socio-cultural constitution of the platform that provides new dynamics, communication formats, and interactivity between users (boyd, 2010: 46-47; Bucher & Helmond 2017: 239; Evans et. al, 2017: 37). Two key features of TikTok contribute to making multimodality on the platform distinct from other social media: audio-centric content creation and the content recommender algorithm of the For You Page. (Kaye, Zeng, and Wikström, 2022).

TikTok videos and the resulting trends and memes that circulate via the creativity of the users often centre on specific audio clips that include speech, music, sounds and

more (Zulli and Zulli, 2020). These clips are often also accompanied by specific sequences of embodied communication that connect with the prescribed audio. The ease of finding, re-using and remixing audio clips is embedded into the platform's user interface which further exemplifies the audio-centricity of communication. For example, through the *Use this sound – feature* users can instantly create a new video based on the audio clip of the video they are currently watching.

Further emphasizing creativity as the basis of activity on TikTok, Kaye, Zeng, and Wikström (2022: 12-14) conceptualize the creation of short videos on TikTok by applying the ideas of *vernacular creativity*, *social creativity*, *distributed creativity*, and *circumscribed creativity* from different scholars. I present these forms of creativity here briefly as they are relevant for understanding all types of TikTok videos:

1. *Vernacular creativity* (adapted from Burgess 2006): affective and platform- specific communication styles of users that are particularly visible as a focus on everyday and mundane content creation.
2. *Social creativity* (adapted from Glăveanu 2020): mutual shaping of creativity between social, material, and cultural assemblages that are particularly visible in features of the platform that facilitate interaction between users for creativity such as the *Use this sound – feature*.
3. *Distributed creativity* (adapted from Sawyer and DeZutter 2009): groups of individuals creating a collaborative end-product with no ownership or responsibility of the result specifically attributed to any of them.
4. *Circumscribed creativity* (adapted from Kaye, Chen, and Zeng 2021): creativity as it is facilitated and constrained by the features of short video creation on TikTok as well as

platform policies such as platform governance constraining through moderation.

The videos on TikTok are mainly discovered by users through a user-specific recommendation algorithm on the For You Page (FYP) of TikTok, which fosters new types of sociality, communication, and identity building practices (Abidin, 2021). The highly public and popular FYP makes TikTok videos very susceptible to spread “virally” to unforeseen audiences of form and scale (e.g. Boffone, 2021: 6). Users attempt to perceive how the algorithmic process turns human communication into data and attempt to structure their multimodal communication to fit this process (e.g., Burgess *et. al.*, 2022: 55-56, 86). Through how TikTok is built to deliver videos on the FYP, the visibility of messages and people’s bodies emerges as a key affordance to understand multimodal communication on TikTok. The visibility to be managed by both the users and the platform concerns the relative ease of finding users and videos on TikTok, and embodied ways of being in making oneself look and see and be seen and looked by others (Evans *et. al.*, 2017: 42; Jones, 2020: 24-25).

## Sexuality and Sex Workers on Social Media

Questions of the visibility of bodies and messages take a heightened turn when the focus is on sexuality and sexual expression on social media. Sexual expression on social media is ubiquitous, but simultaneously highly regulated by platforms through “community guidelines”, content moderation practices, and de-platforming (Tiidenberg and van der Nagel, 2020; Are and Briggs, 2023). The visibility of sexuality on social media platforms is at its core a question of how sexuality as a force that shapes sociality and society is presented (Paasonen *et.al.*, 2023). I explore these points of contestation through sex worker promotional content creation, which highlights new tensions in the relationship of users and platforms.

Online sex work (OSW) is typically defined as including the production and sale of erotic or sexual content online that was either produced earlier or is provided to the audience as a live broadcast (Easterbrook-Smith, 2022). Online sex workers use social media for building a following through promotional work and creating communities and relations with prospective audiences (Easterbrook-Smith, 2022). The linkages created across social media platforms through promotional work are essential for the work in the adult entertainment industry to be economically viable (Are and Briggs, 2023). At the current stage of research, I define the promotional content quite broadly, but the core elements include implicit or even explicit sexual innuendo and double-entendres. Sex workers face a constant threat of invisibility posed by platforms to remove and reduce the visibility of sex worker content coupled with inconsistent and unclear specifications of what sexual activity, nudity, and

solicitation is and is not (Are and Briggs, 2023). Sex workers must structure multimodal communication to be understood by prospective audiences correctly, evade algorithmic detection models, and skirt the community guidelines guiding the human moderators in their work.

## Moderating Content and Visibility on Social Media

A key contributing factor to the precarious state of existence for sex workers on TikTok is content and visibility moderation. Platforms shaped by human and algorithmic practices control how information is exchanged, and user activity directed by deciding on what to show and not show to different users (Zeng and Kaye, 2022). Moderation practices are based on “community guidelines” documents directed to users, and on the platform’s internal guidelines to moderators (Gillespie, 2018). TikTok’s guidelines for sexual expression ban displays of nudity and implied nudity or sex acts as well as solicitation of sexual activity or videos that glorify such solicitation (TikTok, 2023a).

Based on both the out-facing and internal guidelines, automated and human moderation practices are used by platforms in coordination, in a process where “problematic” content is either removed, reduced in visibility, or escalated for further review to human moderators (e.g., Zeng and Kaye, 2022; Gillespie, 2018). There is a trend towards the use of visibility moderation practices that in effect reduce the visibility or reach of “problematic” content (e.g. Zeng and Kaye 2022; Savolainen, 2022, Are, 2022). Such practices are often opaque to users who may not be aware that their visibility has been altered. TikTok in effect admits to this practice by stating that certain types of content are ineligible to be recommended by the FYP-algorithm (TikTok, 2023b).

The audio and algorithm centricity of TikTok’s platform affordances coupled with the platform’s intent on moderating which videos can be seen have major effects on how visibility is produced. **These effects also extend to questions of data gathering, data analysis, and how research impacts the visibility of precarious groups such as sex workers.** I thus propose that the following questions be asked with the platform affordances in mind:

- How to access and acquire multimodal TikTok short video data?
- How to analyse and understand multimodal TikTok short videos in general, and especially in terms of how online sex workers manage visibility?
- What are the implications of studying groups whose visibility on platforms is at risk and may be further threatened by research?

The next section provides some work-in-progress approaches to answer these questions and provide solutions.

## How To Access and Acquire Tiktok Short Videos?

The contested status of sex worker promotional videos, and the audio and algorithm centric nature of TikTok pose a challenge for data gathering and analysis. It is generally necessary to interact with the content recommendation algorithm and various other technological platform features for manual multimodal data gathering. Any interactions with the algorithm also feed into its recommendations for the future, thus creating at least a partial bias based on the researcher's past interaction with the platform. The methods for data gathering presented here are experimental in nature and intended to mitigate at least some of the challenges that TikTok's technological structure presents. I present them here for further iteration and discussion based on the questions raised by my current approaches.

I have adopted a digital ethnography inspired methodology in the early stage of data gathering that is based on previous explorations into TikTok (Abidin, 2021; Schellewald, 2021). The methods used are more focused on spending time in the app and gaining an understanding of how content creation and multimodal communication take shape than on collecting a specific number of data samples. I have so far chosen to use 1-hour increments in data gathering with focus on the different data gathering methods outlined in this section. These sessions include taking and then reviewing field notes, screen grabbing relevant data for further analysis, experimenting with the platform's functions and recommender algorithm, and collecting information on hashtags and audio clips.

The key issue of algorithmic feedback loops that TikTok and its trends are prone to perpetuate by design can be alleviated via a more longitudinal ethnographic approach that I intend to incorporate by systematically repeating observation sessions (Schellewald, 2021: 1440-1441). Further on the course of this article's research I will also conduct supplementary interviews both to better understand multimodal practices of online sex workers as well as to give them a voice in the research design. The digital ethnography data gathering process is preceded with a walkthrough of the TikTok platform for increased context.

The walkthrough method devised by Light et.al. (2018) can be used to critically engage with the affordances of the platform, which include technological features as well as the vision, operating model, and different governing modes of the platform. The *vision* of the app concerns what the app or platform is supposed to do and by extension implies how it can be used and by whom. The *operating model* concerns the business strategy and revenue sources of the platform with underlying political and economic interests. The *governing modes* concern how the app provider seeks to manage and regulate user activity to sustain their operating model and fulfil their vision. Material for the walkthrough may be found within the app itself or as documents and guidelines produced particularly by the company operating the platform (i.e., ByteDance for TikTok). The analysis of

this expected environment of use is followed by a technical walkthrough in which I used the desktop version and the mobile app of TikTok for three key stages: registration and entry; everyday use; and app suspension, closure, and leaving. My walkthrough analysis of TikTok was particularly centred on aspects that affect multimodal communication and visibility of users on the platform.

The walkthrough provides a basis for the next step in the data gathering process, where different options for eliciting data are explored. Based on the walkthrough I devised four different options for data gathering that were and can be used in conjunction. The experiments suggest that the *Use this sound* – repository may emerge as the key method. The methods also reflect the ways of video discovery on TikTok in general and are outlined below:

- Feeding the algorithm
- Hashtags
- Use this sound -repository
- Identifying specific creators

My experiments in data gathering suggest that the focus should be directed to the key audio and algorithm centric qualities of TikTok in facilitating data gathering. The process relies on first finding a suitable path to the relevant content through *feeding the FYP algorithm* with inputs (i.e., liking, saving, repeatedly watching specific videos) to amass a suitable pool of videos to build upon for further data gathering. I chose to first feed the algorithm with inputs on videos that are related to dating and relationships to navigate a path to online sex worker videos. Using the FYP algorithm can however only ever catch a glimpse of the reality of what is happening on TikTok.

The process can be supplemented by identifying and analyzing the use of *specific hashtags used in the relevant content*. Following the feeding of the algorithm to amass suitable data, I then observed a further number of videos to both document the used hashtags and to loosely categorize these hashtags thematically. The intent is to find hashtags that may typically hold online sex worker content or other hashtags that host such content less overtly.

The hashtags often connect with specific audio clips that users can embed within their videos. These audio clips can be found, reused, and remixed by users through *The use this sound -repository* that acts as a further tool to understand TikTok multimodality. It offers an audio-centric way to gather data on specific audio clips. In the current stage of the experiment, I was for example able to identify a trend called “the flashing background”, where creators used items that reflect light to showcase their bodies in ways that would not normally be in line with TikTok's content and moderation policies. This trend was often presented with a specific accompanying hashtag and audio clip which both enhanced its visibility by making the videos searchable.

The last supplementary method relies on identifying and focusing on specific creators on the platform for data or

using a creator's profile as a tool to access a wider variety of TikTok trends across a longer span of time. The video data is screen recorded from the TikTok mobile app to best preserve the intended format of the material for the building of a corpus for multimodal analysis.

## Multimodal Analysis of Tiktok Short Videos

The multimodal discourse analysis of the videos is done by adopting the transcription model devised by Baldry & Thibault (2005: 165-249). The transcription model is theoretically based on a systemic-functional-linguistics (e.g., Halliday, 1978) understanding of language use guided by metafunctions divided into *experiential* (demonstrating e.g. events, places, things, people), *interpersonal* (demonstrating relations between viewers and the world), *textural* (construction of an overall composition of a balanced text), and *logical* (construction of narratives and sequences of events) meaning making relations within the multimodal text (Baldry & Thibault 2005: 226).

The framework enables analysis of video-based data and the inclusion of the modalities possible for meaning making on TikTok. However, since the framework is not intended for analysis of social media data, I will present some iterations that concern TikTok's multimodal features specifically. The framework directs higher level analysis to time, visual frame, visual image, kinesic action, and soundtrack in connection with the metafunctional interpretation of the communication (Baldry & Thibault 2005: 174). They are further elaborated in the adapted framework by examining how visibility is produced via the *expression form* and *content form* of the videos. Baldry & Thibault (2005: 226) specify how the expression form and content form are linked to the metafunctional interpretations in their work as shown below.

*The expression form* concerns the display of invariants and their transformations in time in the delimited optic array (i.e. display of variations and repetitions in the optic qualities of the video transmitted on the screen). The metafunctions on the expression form manifest as:

- *Experiential*: Display on the screen of transformations, substitutions, nullifications of structure + visual kinaesthesia based on camera movement that produce a changing optic array
- *Interpersonal – orientational*: Field of view and movement of the camera as the optic array of the viewer + simulation of head-body movement in orientation to viewer
- *Textural*: Deletions, accretions, slippage of texture in the optic array
- *Logical -transitional*: Visual transitions as based on camera movement (e.g. pan, zoom, dolly shot), and based on video editing (e.g. cut, wipe, merge, dissolve) in post-production

*The content form* concerns depiction of events in the depicted world that the viewer sees on the screen. The metafunctions on the content form manifest as:

- *Experiential*: Depiction / perception of objects and events in the form of volumes and vectors in depicted world + movement of observer in depicted world
- *Interpersonal – orientational*: Use of colour, modalisation, camera angles to orient the viewer to the depicted world and to adopt an evaluative stance towards it; the creation of social-interpersonal relations between viewer and the depicted world
- *Textural*: Compositional principles of wholeness, balance, the relations of part to the whole
- *Logical -transitional*: Shot as single run of camera with no displacement in time or place of depicted scene + nesting of shots in higher-order units; dependency relations between shots

Focusing on the expression form enables the analysis of how OSW-creators combine TikTok platform features such as the extensive editing, audio creation and manipulation, and visual filtering tools to create promotional videos that circumvent content and visibility moderation. Focusing on the content form in turn enables the analysis of the meaning making by OSW-creators in terms of how they create sexualized and un-sexualized narratives and scenes in their videos. Combining these forms of displaying visual imagery and depicting the world makes it possible to analyse how visibility is managed in relation to the viewer and the platform. The focus is on how the video creator's status as creating OSW-content is made sensible to the viewers within the content guidelines enforced by TikTok.

The question of how visibility is managed in relation to the For You Page – algorithms and the governance of visibility by the platforms requires the incorporation of further functions of the platform into the analysis. The functions encompass the *like*, *comment*, *save*, and *share* functions that directly influence how users can interact with videos. When a viewer interacts with these functions, they all affect how the FYP- algorithm delivers future videos to the specific user and how likely it is to be shown to other users. As such these functions form a further component of videos are understood. Two further features for affording visibility are hashtags, and the audio clip repository. These features affect how videos can be found by users and the platform as well as how they can be re-used by other users.

## Present Insights into Visibility Management by Online Sex Workers

The analysis of the videos collected so far for this research has revealed some emerging trends of content creation that are outlined briefly below:

- Creators combine kinesic action of facial expressions and visual framing of close distance to viewers with spoken audio that simulate flirting, dating, and an imagined romantic relationship between creator and viewer.
- Creators construct humorous dialogues of everyday life scenes between them and the person filming which feature sexual innuendo and jokes.
- Creators produce imitations of sexual activity with kinesic action and audio, where the visual framing leaves out parts of bodies sensitive to moderation.
- Creators produce audio clips specifically designed to be re-used by other OSW-creators for their content creation.
- Creators produce videos that specifically refer or allude to their OSW content creation practices and how they can not show everything on TikTok to highlight the content found elsewhere.
- Creators take well known genres of audio-based pornographic material and adapt them for TikTok while evading visual and text-based moderation.
- Creators employ various forms of cuts and editing of visual content that encourage repeated views of their videos which in turn feed visibility to the algorithm. These cuts also limit the time that body parts sensitive to moderation are visible.
- Creators recontextualize already popular audio meme templates with sexual innuendo which are then potentially visible to larger audiences but also to increased scrutiny of moderation.
- Creators intentionally misattribute audio clips and often refrain from using any hashtags to curate their visibility to a more specific group of viewers while making their videos less easily searchable and visible on TikTok in general.
- Creators use text overlaid on videos to contextualize the videos as sexually suggestive but carefully refrain from direct referrals to sex or choose to use strings of emojis and intentionally misspelled words to evade moderation.
- OSW-creator videos regularly feature high numbers of views, likes, comments, and shares, which suggests that creators can adapt their promotional videos in ways that reach viewers despite TikTok outlining to not recommend videos with sexual content on the FYP.

The present findings point to a wide range of uses of TikTok functions for video creation and editing by OSW-creators to manage their visibility both towards viewers and the platform's governance. The creators' use of the platform appears to be highly in concordance with the four forms of creativity identified by Kaye, Zeng, and Wikström (2022). Creators appear to be quite skilled in utilizing mundane, everyday scenarios and making them into promotional material through contextualization with sexual meanings. The creators engage in use of the interactivity features of TikTok such as creating content based on existing audio clips and then modifying them, as well as responding directly to user comments with new videos, or

creating audio clips themselves that call for user participation (e.g., staring challenges facilitating the use of the *Duet-function* that enables users to response to a video with a video of their own played alongside the original). Creators also remix a wide variety of audio templates with little regard to how they were originally conceived, which may in some cases result in a change of the generally perceived meaning of e.g. a specific audio meme template. This process muddles the idea of ownership and responsibility over production of video content.

Finally, creators actively use the features provided for short-video creation and enhancing visibility on the platform in ways that are not in concordance with the content guidelines and policies outlined by TikTok. They can do this precisely because TikTok enables such a highly remixable, editable, and permutable multimodal meaning making. Through careful and strategic use of the multimodal meaning-making afforded by TikTok, OSW-creators can overcome many of the constraints to visibility that come in the form of content and visibility moderation. Their polysemous videos do not necessarily rely on exposure of the body or text content that may be caught by text based or visual moderation algorithms. Simultaneously, they contribute greatly to various creative processes on the platform that would not look the same without their presence. The early findings of this research suggest that the platform governance models on TikTok target a low hanging fruit of a high amount of exposure of the body or explicit nudity. TikTok can police bodies but not ideas or creativity for it fundamentally relies on their circulation.

The present analysis also suggests that to iterate this research framework further, comments could be scrutinized in more detail to better ascertain that viewers also understand these videos "correctly". Additionally, the importance of the profile page for OSW-creators should be examined since it acts as the gateway to the online sex worker content these creators are making elsewhere. The findings will also be supplemented by interviews with the creators of videos to better understand what factors they consider in managing visibilities.

## Research Ethics Regarding Visibility

The analysis of sex worker promotional videos from the chosen research perspective brings with it a host of questions about visibility of these users and their communication practices. Acquiring informed consent for content that is reproduced in research is a sensible baseline. However, this only partially solves issues. I propose that it is also highly important to better understand the viewpoints of the creators that are in a precarious position through giving them a voice within the research structure. This can be carried out through supplementary interviews that also help to better capture the intent of the creators as opposed to only imposing the researcher's point of view on the videos. This is also particularly important because I do not

face the risks these creators do and stand to benefit from conducting the research.

The results of this research also raise questions of what happens when we gain a better understanding of these promotional practices that attempt to circumvent moderation? Will the platform operators of TikTok use this knowledge to put these creators at further risk of having their visibility reduced or taken away? Or is it possible through research to also make visible how sex workers act as drivers of communication on a more general level on TikTok? I have suggested that the creative uptake of TikTok's features by these users highlight the centrality of the multimodal creation features also for circumventing moderation. Altering the features is likely not something TikTok would desire. Overall, even if sex worker promotional video creation is mostly an exercise in increasing visibility and driving traffic to their respective "home platforms", it is important to remember that not all publicity is good publicity.

## References

- Abidin, C. (2021) Mapping Internet Celebrity on TikTok: Exploring Attention Economies and Visibility Labours. *Cultural Science* 12 (1), pp. 77--103.
- Are, C. 2022. The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram. *Feminist media studies* 22(8), pp. 2002--2019.
- Baldry, A, Thibault, P.J. (2006) *Multimodal Transcription and Text Analysis: A Multimedia Toolkit and Coursebook with associated on-line course*. Oakville, CT: Equinox Publishing.
- Boffone, T. (2022) Introduction: The Rise of TikTok in US Culture. In (Ed.) Boffone, T. *TikTok Cultures in the United States*. Milton: Taylor and Francis. pp. 1--13.
- Bucher, T, Helmond, A. (2017) The Affordances of Social Media Platforms. In (Eds.) Burgess, J., Marwick, A., Poell, T. *The SAGE Handbook of Social Media*. 55 City Road: SAGE Publications Ltd. pp. 233--253.
- Burgess, J. (2006) Hearing ordinary voices : Cultural studies, vernacular creativity and digital storytelling. *Continuum*, 20 (2), pp. 201--214.
- Burgess, J, Albury K., McCosker A., Wilken R. (2022) *Everyday Data Cultures*. Cambridge, UK: Polity Press.
- Carolina Are, Briggs P. (2023) The Emotional and Financial Impact of De-Platforming on Creators at the Margins. *Social media + society* 9.
- danah boyd. (2010) Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. In (Ed.) Papachrissi, Z. *Networked Self: Identity, Community, and Culture on Social Network Sites*. New York: Routledge. pp. 39--58.
- Easterbrook-Smith, G. (2022) OnlyFans as Gig-Economy Work: a Nexus of Precarity and Stigma. *Porn studies* (Abingdon, UK), pp. 1--16.
- Evans, S.K., Pearce, K.E., Vitak, J., Treem, J.W. (2017) Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research. *Journal of Computer-Mediated Communication* 22, pp. 35--52.
- Gillespie, T. (2018) *Custodians of the Internet : Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.
- Glăveanu, V.P. (2020) A sociocultural theory of creativity : Bridging the social, the material, and the psychological. *Review of General Psychology*, 24 (4), pp. 335--354.
- Halliday, M.A.K. (1978) *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold (Publishers) Ltd.
- Jones, R.H. (2020) Towards an embodied visual semiotics: Negotiating the right to look. In Diémoz, F., Dürscheid, C., Thurlow, C. (Eds.), *Visualizing Digital Discourse: Interactional, Institutional and Ideological Perspectives*. Berlin: De Gruyter Mouton.
- Kaye, D.B.V., Chen, X., Zeng, J. (2021) The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok. *Mobile Media & Communication*, 9 (2), pp. 229--253.
- Kaye, D.B.V., Zeng, J., Wikström P. (2022) *TikTok: Creativity and Culture in Short Video*. Cambridge: Polity Press.
- Light, B., Burgess J., Duguay, S.. (2018) The Walkthrough Method: An Approach to the Study of Apps. *New media & society* 20(3), pp. 881--900.
- Paasonen, S., Sundén, J., Tiidenberg, K., Vihlman, M. (2023) "About Sex, Open-Mindedness, and Cinnamon Buns: Exploring Sexual Social Media." *Social media + society* 9(1): 205630512211473--.
- Savolainen, L. (2022) The Shadow Banning Controversy: Perceived Governance and Algorithmic Folklore. *Media, culture & society* 44(6), pp. 1091--1109.
- Sawyer, K., DeZutter, S. (2009) Distributed creativity : How collective creations emerge from collaboration. *Psychology of Aesthetics, Creativity, and the Arts*, 3, pp.81--92.
- Schellewald, A. (2021) Communicative Forms on TikTok: Perspectives From Digital Ethnography. *International journal of communication* 15, pp. 1437--1457. DOAJ Directory of Open Access Journals.
- Tiidenberg, K, van der Nagel, E. (2020) *Sex and Social Media*. Bingley, England: Emerald Publishing.
- TikTok. (2023a) Sensitive and Mature Themes. <https://www.tiktok.com/community-guidelines/en/sensitive-mature-themes/>
- TikTok. (2023b) For You feed Eligibility Standards. <https://www.tiktok.com/community-guidelines/en/fyf-standards/>
- Zeng, Jing, and D. Bondy Valdovinos Kaye. (2022) From Content Moderation to Visibility Moderation: A Case Study of Platform Governance on TikTok. *Policy and internet* 14(1): pp. 79--95.
- Zulli, D, Zulli, D.J. (2020) Extending the Internet Meme: Conceptualizing Technological Mimesis and Imitation Publics on the TikTok Platform. *New media & society*. SAGE Premier. pp. 1--19.

# MigrTwit Corpora. (Im)migration Tweets of French Politics.

Sangwan Jeon

Université de Lille

E-mail: sangwn.jeon@univ-lille.fr

## Abstract

Since the early 2010s, French politicians have steadily utilized Twitter as a communication tool. Political tweets about immigration are prolifically produced by Marine Le Pen (former leader of the French far-right populist Party Rassemblement National). Their biased social representation of migrants, immigrants, and asylum seekers seems to be instilled in manifold voters through online public debate and media. To characterize political immigration discourse on Twitter, we developed the diachronic bilingual corpus of political tweets posted throughout the last 12 years, from 2011 to 2022. The whole MigrTwit corpus consists of three subcorpora, for a total of 23869 tweets, with 703016 words. The constitution of the French MigrTwit corpora enabled us to study the evolution of immigration discourse in comparative and corpus-based approaches.

**Keywords:** Corpus linguistics, Twitter, Political discourse, Critical discourse analysis, Immigration discourse, Online hate speech

## 1. Introduction

Immigration issues continuously constitute the core of the political agenda of far-right populist parties in Europe and in the United States, increasingly appealing to manifold voters (Wodak, 2021; Aït Abdeslam, 2021). In France, anti-immigration discourses have been tremendously produced by the far-right Party Rassemblement National<sup>1</sup> (henceforth RN). International and national events at various levels seem to fuel anti-immigration arguments (Wodak, 2021; Pietrandrea & Battaglia, 2022). In this era of Web 2.0, electoral campaigns and results are affected by disinformation and hate speech spread throughout social media platforms (Badouard, 2017; Allcott & Gentzkow, 2017). Since the early 2010s, French politicians have steadily utilized Twitter as a communication tool. Manipulative discourse should be considered to detect abusive language in political discourse (Macagno, 2022). The shortness of messages and the decontextualized feature of political tweets may contribute to the “imperfect and biased way our mechanisms of information processing work” (Hart, 2013). Within the framework of the research project OLINDiNUM (Observatoire LINGuistique du DIscours NUMérique [Linguistic Observatory of Online Debate]), the MigrTwit corpus has been developed, annotated, and analyzed to characterize political immigration discourses on Twitter in collaboration with Elena Battaglia, Guido Blandino, Paola Pietrandrea, and with the participation of Adelina Stoian. The MigrTwit corpus consists of 23869 tweets posted between January 2011 and June 2022, from 51 French and British political figures and parties. It has three components, i.e., the corpus of French right-wing political migr-tweets, the corpus of British right-wing political migr-tweets (cf. Blandino,

2023), and the corpus of French left-wing political migr-tweets. The whole corpus is published and downloadable<sup>2</sup>. In this paper<sup>3</sup>, focusing on the French MigrTwit corpus, I will argue the methodological framework in Section 2. In Section 3.1, I will discuss the production rate and the frequency of migr-tweets of French right-wing politics. In Sections 3.2 and 3.3, I will briefly discuss how the topic of immigration is framed through collocational and topic analyses of tweets.

## 2. Methods

Within the theoretical framework of Critical Discourse Analysis (Reisigl and Wodak, 2015; Van Dijk, 1991; 1980), we proceed to the investigation of the collocations of the migr-lexicon, and the identification of topics, i.e., hashtags of migr-tweets. Hart argued that anti-immigration discourse is a manipulative discourse in the sense that its “argument acts [namely topoi<sup>4</sup>] may automatically yield decisions in favor of discrimination” (2013: 204). We formulate the hypothesis according to which the semantic connotation of migr-lexicon results from the topoi established “in perceived truth-status as a consequence of the frequency with which it is repeated” (Hart, 2013: 204). In other words, the pejoration has been aggravated by the spread of biased social representation of migrants, asylum seekers, and immigrants through right-wing politicians’ tweets containing words derived from the Latin root *-migr-* of *migrare* (henceforth migr-tweets). Since political tweets accelerate based on the political conjuncture, statistical analysis is needed to measure their visibility and frequency depending on time. Analyzing predications reveals the “discursive qualification of social actors, objects, phenomena, events/processes and actions (more or less positively or negatively)” (Reisigl & Wodak, 2014:95).

<sup>1</sup> Formerly the *Front National* till 2018 was founded by Jean-Marie Le Pen in 1972.

<sup>2</sup> Detailed information, such as the list of the selected Twitter accounts, is provided with the published versions of the corpus. You can use the links in the header of Table 1 to reach the site.

<sup>3</sup> As part of my doctoral research in preparation *Le discours de haine sur le web. Analyse linguistique et sociolinguistique* [Hate

*speech on the web. Linguistic and sociolinguistic analyses*].

<sup>4</sup> Following Hart’s work (2013:201), the notion of topoi is defined as “formal and content-related warrants which connect premises with conclusions ([Wodak,] 2001:75)” whose a key feature “is that they are ‘common-sense’ reasoning schemes typical for specific issues (Van Dijk, 2000b: 98)”.



Linguistic constructions such as collocations, which attribute, among others, to this qualification, were automatically annotated *ad hoc* by means of the corpus analysis platform Sketch Engine with the function of Word Sketch Difference. The contrastive approach may shed light on the fallacious side of topoi of the immigration discourse, in particular, “persuasive definition (implicit modification of the meaning of words)” and “quasi-definition (unshared or not commonly accepted inferences from the use of a word taken for granted)” (Macagno, 2022: 72).

## 2.1 Data collection

Concerning the FrRMigr-Twit corpus (henceforth FrR corpus), Pietrandrea and Battaglia (2022) created its first version to analyze migr-tweets<sup>5</sup> ( $n=5689$ ) of selected 10 French right-wing and far-right politicians. Furthermore, I have opted for the contrastive approach to point out formal and linguistic characteristics of political (far) right-wing politicians’ migr-tweets. 39 French political figures and parties were selected, based on four non-mutually exclusive criteria: high number of migr-tweets, political affiliation, political careers, that is, members of the European Parliament, and Presidential candidate between 2011 and 2022. To extend the FrR corpus using the Twitter API<sup>6</sup>, Battaglia and I carried out data collection at the beginning of the fourth quarter of 2021. The collection of tweets ( $n=11761$ ) and metadata from 16 (far) right-wing political figures and parties ended in July 2022. I extracted migr-tweets ( $n=5636$ ) of 23 (far) left-wing political figures and parties during March and April 2023 to create the FrLMigr-Twit corpus (henceforth FrL corpus). Table 1 illustrates essential information about these two corpora constituting the French MigrTwit corpus.

	FrLMigr-Twit	FrRMigr-Twit
Migr-tweets	5636	11761
Average number of migr-tweets per account	296.50	853.27
Twitter accounts	23	16
words	169818	358491

Table 1: FrMigr-Twit corpora

## 3. Results

### 3.1 Frequency analysis

Concerning the raw frequency of results, right-wing

<sup>5</sup> 5689 tweets were retrieved through Europresse.com before the Twitter API v2 Academic Research was launched in 2021.

<sup>6</sup> The Academic Research service has been deprecated since May 2023.

<sup>7</sup> The average number of migr-tweets is a sum of the average numbers of every single year. Since not every account did produce migr-tweets for all of the dozen years, the average number of

politicians produced more migr-tweets than left-wing politicians (Table 1). Overall, approximately 558 more migr-tweets per account have been posted by right-wing politics over the last dozen years, i.e., 853.27 versus 296.5 migr-tweets per account<sup>7</sup>. Moreover, we observed the considerable increases in Marine Le Pen’s and the other right-wing politicians’ migr-tweet productions in 2015 and 2018 (Figure 1). The peak year of 2015 for Marine Le Pen’s migr-tweet production is also a year of its highest growth (1123% increase). Incidentally, 2015 corresponds to the year of the highest growth for both FrR and FrL corpora, but their peak year is 2018.

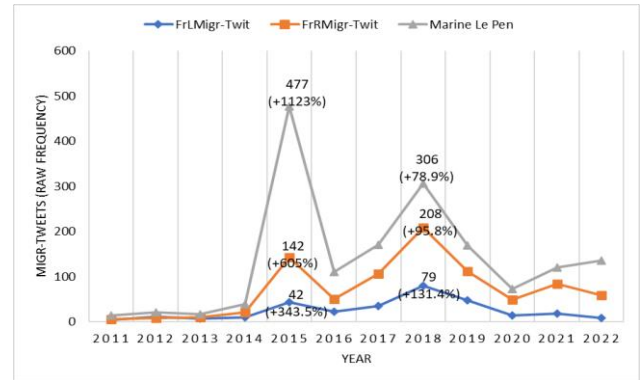


Figure 1: Average number of migr-tweets per account

We also monitored the monthly production of all tweets ( $n=212977$ ) of selected accounts ( $n=32$ ) to calculate the annual proportion of migr-tweets<sup>8</sup> (Table 2). The visibility of migr-tweets is calculated by dividing the total amount of retweets, likes, replies, and quotes by the number of migr-tweets (Table 3). Despite the relative low frequency of left-wing politics, their migr-tweets were more visible than those of right-wing politics in 2011.

	11	12	13	14	15	16	17	18	19	20	21	22
FrL	0.3	0.4	0.2	0.6	2.5	1.5	2.2	6.5	3.6	1.3	1.9	0.9
FrR	0.7	0.6	0.5	1.5	5.1	2.8	3.1	7.4	5.3	2.8	3.9	3.8

Table 2: Proportion of migr-tweets

The topic of immigration seems to have become viral on Twitter for both groups since 2012, albeit 2013 is the only year where the virality drops for both groups. However, since 2013, right-wing politicians’ migr-tweets have become more viral, and their virality has kept increasing until 2020.

migr-tweets is greater than the quotient of the total number of migr-tweets and a total number of Twitter accounts.

<sup>8</sup> The number of accounts ( $n=16$ ) is paired to calculate the proportion of migr-tweets of each group. Based on their productivity of migr-tweets, 16 left-wing political accounts were selected.

	FrL	FrR
11	7.0 (+3.0)	4.0
12	60.4	69.8 (+9.4)
13	8.7	39.1 (+30.4)
14	28.0	87.8 (+59.8)
15	25.5	134.8 (+109.3)
16	54.2	297.7 (+243.5)
17	253.8	325.5 (+71.7)
18	235.0	341.2 (+106.2)
19	195.3	527.2 (+331.1)
20	276.7	958.4 (+681.7)
21	295.2	908.3 (+613.1)
22	424.8	1001.3 (+576.5)

Table 3: Virality of migr-tweets

The first year of the highest increase (i.e., 2015) will be further investigated in Section 3.3 with a closer investigation of hashtags of June 2015 because the monthly frequency of migr-tweets shows two different peak months for each group, i.e., June for the FrL corpus and September for the FrR corpus (Figure 2).

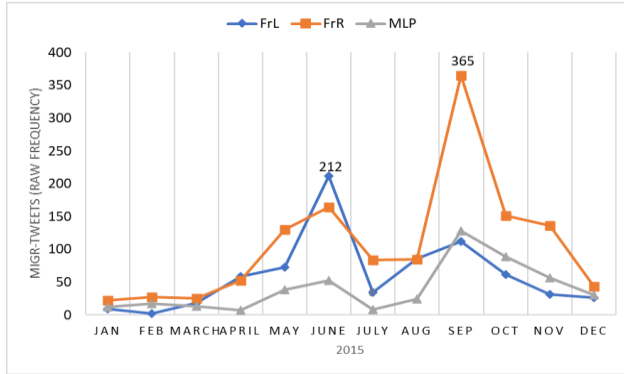


Figure 2: Number of migr-tweets of 2015 by month

### 3.2 Collocational analysis

Until 2014, a low frequency of MIGRANT<sup>9</sup> was observed through the FrR corpus, even with no occurrence in 2012. Following the 2015 refugee movement, migr-tweets containing MIGRANT increased through both subcorpora. Within the FrR corpus, the topos of abuse<sup>10</sup> is depicted from 2015 by modifiers, such as *economic*, *clandestine*, and *illegal*, as in (1). The explicitly threat-connoting modifier, i.e., *terrorist* occurred once in 2016 in the tweet (2) referring to the Daesh terrorist attack posted by Marion

Maréchal, retweeted 921 times.

- (1) *We must block the flow of #migrants essentially composed, contrary to what is said, of **economic migrants**. #RadioClassique ;2015-06-17:@ECiotti;57;17;24;0*<sup>11</sup>
- (2) *Remind yourself, we were told that it was false. Another **terrorist migrant** arrested today. How many others?; 2016-08-05:@MarionMarechal; 921;132;789;38*

Since 2019, ethnic modifiers used to depict crimes reported as committed by “a migrant” coming from Muslim ethnic groups have become typical<sup>12</sup> collocates, e.g., *Eritrean*, *Sudanese*, *Afghanistan*, *Algerian*, etc. Conversely, ethnic modifiers do not constitute the typical collocates of MIGRANT at the side of the FrL corpus. The inherent features of MIGRANT are rather observed with modifiers like *vulnerable*, *precarious*, and *climatic*. The humanitarian topos is triggered through the left-wing politicians’ migr-tweets as in (3), which are targeted by the far-right populist politicians. To give an example, when it comes to *minor migrants*, the humanitarian act of protecting children has been steadily pulled away as in (4) and (5).

- (3) Posted in February 2018 by Manon Aubry, member of the far-left populist party La France Insoumise:  
*RT@afpfr: Justice rules against the prefect and suspends the return of **minor migrants** to Italy #AFP; 2018-02-24:@ManonAubryFr;123;0;0;0*
- (4) Posted in May 2019 by Jordan Bardella, member of the RN:  
*We have been witnessing, for the past few months, in the Bourgogne Franche Comté region, the TRIPLING of “isolated **minor migrants**”, and these “**minor #migrant**” are as minor as I am I’m an archbishop! #Yonne #Le26MaiVotezRn ; 2019-05-21:@J\_Bardella; 114;28;186;5*
- (5) Posted in June 2022 by Marine Le Pen, former leader of RN:  
*At the #StadeDeFrance, it was the surge of hordes of **ultra-violent migrant minors** from the Porte de la Chapelle, these were raids perpetrated by city gangs acting with impunity. The government concealed the seriousness of the facts! #Legislative2022; 2022-06-05:@MLP\_officiel;183;20;425;4*

### 3.3 Topic analysis of hashtags

2015 was the transition year in terms of virality, productivity, and frequency of migr-tweets. The distribution of hashtags in migr-tweets posted during June and September showed that the highest increase in migr-tweet production in both subcorpora is not interrelated in 2015. Figure 3 illustrates the distribution of hashtags in June 2015. Regarding the FrL corpus, 66 out of 207 hashtags refer to the live issue of the violent evacuation of

<sup>9</sup> By writing the term in capital letters, I refer to its lemmatic status.

<sup>10</sup> “[T]he topos of abuse can be expressed as: “if a right or an offer for help is abused, the right should be changed, or the help should be withdrawn, or measures against the abuse should be taken” (Wodak, 2001: 77)” (Hart, 2013).

<sup>11</sup> The tweet example format is structured as follows: translation in English; posting date (year-month-day); user name; retweet; reply; like; quote.

<sup>12</sup> The typicality is measured utilizing the logDice score calculated by the analyzing tool Sketch Engine.

migrants from the Halle Pajol in the 18th district of Paris on 8 March 2015. In addition, 14 hashtags refer to the topic of ASYLUM, i.e., #réfugié (*refugees*), #asile (*asylum*), #Waterloomoral. However, none of the right-wing politicians' migr-tweets contain hashtags related to the topic of ASYLUM, even though 51 hashtags refer to the keyword of MIGRANT. Furthermore, the generic noun IMMIGRATION is more frequent in right-wing politicians' migr-tweets than in left-wing politicians' migr-tweets.

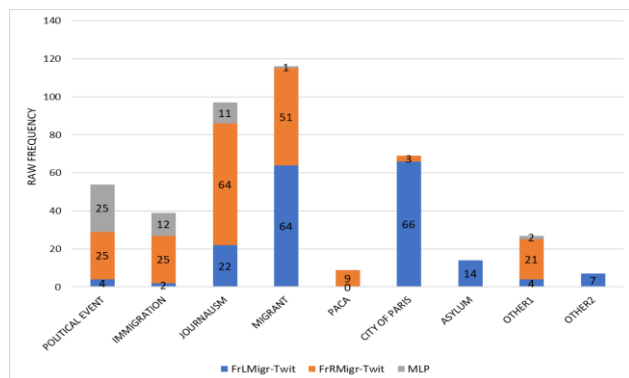


Figure 3: Topics of hashtags in migr-tweets of June 2015

The live issue of the evacuation of migrants is absent in right-wing politics' migr-tweets. Their hashtags refer to communication contexts such as the POLITICAL EVENT (e.g., #ConfMLP) and JOURNALISM (e.g., #BFMPolitique). 25 migr-tweets containing #ConfMLP were posted by Marine Le Pen on June 10, 2015. Hashtagging enables several "related" tweets to be linked together as in (6) and (7).

- (6) "Last February, a small town in #Burgundy was forced to take 60 migrants in order to unclog #Calais." #ConfMLP ; [2015-06-10:@MLP\\_officiel;59;7;26;0](https://twitter.com/MLP_officiel/status/597260)
- (7) "With his peopling policy, @manuelvalls wants to disseminate #immigration in our countryside in order to make it less visible" #ConfMLP; [2015-06-10:@MLP\\_officiel;103;9;34;0](https://twitter.com/MLP_officiel/status/1039340)

#### 4. Discussion

The analysis of the monthly distribution of hashtags indicates what topics are selected and how they are organized according to specific events or current topics because "[t]opics not only suggest what information is most important in the text, but also what is most important 'in the world'" (Van Dijk, 1991: 74). A closer investigation of hashtags of June 2015 showed that specific events put forward differ between right-wing and left-wing politicians. I suggest that even though MIGRANT is the most likely to be "what is the center or focus of the information" (Van Dijk, 1980: 98) as being highlighted by means of hashtagging and reiteration, the topic of MIGRANT is exploited by the far-right populist politicians to undermine

their political opponents.

#### 5. Conclusion

Comparatively exploiting the FrMigr-Twit corpora, the statistical analysis showed that right-wing political migr-tweets have become more productive and viral since 2015. By analyzing the collocations of MIGRANT, we demonstrated that topoi of anti-immigration discourse (i.e., topoi of abuse and danger) are gradually conveyed through right-wing political migr-tweets defeating the humanitarian topoi. Twitter structure (i.e., hashtag, brevity of messages, decontextualization) and the frequency contribute to the semantic connotation of the migr-lexicon.

#### 6. References

- Ait Abdeslam, A. (2021). Muslims and Immigrants in the Populist Discourse of the French Party Rassemblement National and Its Leader on Twitter, *Journal of Muslim Minority Affairs*, 41:1, 46-61.
- Allcott, H., Gentzkow, M. (2017). Social media and fake news in the 2016 election, *Journal of Economic Perspectives*, 31:2, 211-236.
- Badouard, R. (2017). *Le désenchantement de l'internet : Désinformation, rumeur et propagande*, FYP éditions.
- Battaglia, E., Blandino, G., Jeon, S., Pietrandrea, P. (2022). MIGR-TWIT Corpus. Migration Tweets of right and far-right politics in Europe [Data set]. Zenodo, <https://doi.org/10.5281/zenodo.7347479>
- Blandino, G. (2023). 10 years of public debate on immigration: combining topic modeling and corpus linguistics to examine the British (far-)right discourse on Twitter, MA thesis. University of Wolverhampton.
- Hart, C. (2013). Argumentation meets adapted cognition: Manipulation in media discourse on immigration, *Journal of Pragmatics*, 59, 200-209.
- Macagno, F. (2022). Argumentation profiles and the manipulation of common ground. The arguments of populist leaders on Twitter, *Journal of Pragmatics*, 191, 67-82.
- Pietrandrea, P., Battaglia, E. (2022). "Migrants and the EU". The diachronic construction of ad hoc categories in French far-right discourse, *Journal of Pragmatics*, 192, 139-157.
- Pietrandrea, P., Jeon, S. (2023). MIGR-TWIT CORPORA. Migration Tweets of French Left-wing Politics. [Data set]. Zenodo, <https://doi.org/10.5281/zenodo.7871602>
- Reisigl, M., Wodak, R. (2015). The Discourse-Historical Approach (DHA), In: Wodak, R., Meyer, M. (Eds.), *Methods of Critical Discourse Studies*, 87-121, SAGE.
- Scott, K. (2015). The pragmatics of hashtags: Inference and conversational style on Twitter, *Journal of Pragmatics*, 81, 8-20.
- Van Dijk, T-A., (1991). *Racism and the Press*, Routledge.
- Van Dijk, T-A., (1980). *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Lawrence Erlbaum Associates, Inc., Publishers.
- Wodak, R. (2021). *The Politics of Fear. The Shameless Normalization of Far-Right Discourse*, 2<sup>nd</sup> ed., SAGE.

# Anonymization of Persons in Videos of Authentic Social Interaction: Machine Learning Model Selection and Parameter Optimization.

André Frank Krause, Anne Ferger, Karola Pitsch

University of Duisburg-Essen

andre.krause@uni-due.de, anne.ferger@uni-due.de, karola.pitsch@uni-due.de

## Abstract

Automatic anonymization of persons in video recordings requires robust detection of face and head areas. Machine learning-based face and posture detectors provide bounding boxes of face and head regions, but specific parameters need to be optimized to maximize the number of correctly anonymized persons and minimize manual annotation and verification efforts. Three different, state-of-the-art ML models (RetinaFace Detector (RFD), Dual-Shot Face Detector (DSFD) and Yolo7-Pose Detector (Y7PD)) were evaluated regarding their suitability for face- and head-region anonymization. Results on our specific anonymization test dataset show that RFD slightly outperforms DSFD if recall (maximizing anonymization) is favored over precision (minimizing false positive face detections). Y7PD yields an even better recall, but at the cost of comparatively low precision. Besides anonymization, collected detector outputs can provide useful data for multimodal interaction research, like body-posture trajectories and face locations.

**Keywords:** anonymization, face detection, parameter optimization

## 1. Introduction: Anonymizing Video Data

Empirical research on human social interaction requires the researcher to respect ethical and legal issues of data collection and management (Roth et al., 2018). In particular, when creating video corpora of authentic multimodal communication, researchers need to be concerned with the protection of personal data both on the auditory and visual level (Rubinstein and Hartzog, 2016). While there is a long tradition of anonymizing or pseudonymizing transcripts of the spoken word (Kretzer, 2013), there is little information on how to best do this for video data. In particular, participants in a study who have been anonymized should not be re-identifiable by other humans or computer algorithms.

On the textual level, relevant personal data to be protected concerns in particular the names of persons, places, streets, federal states, and institutions, the professional and educational background of participants, as well as time information and more indirect context information (e.g. (Kretzer, 2013)). Anonymization of voice constitutes another urgent topic exhibiting conflicting objectives (e.g., privacy vs. utility for linguistic and interactional research (Srivastava et al., 2022)). In video recordings of social interaction, the visual level also needs consideration. For example, persons can potentially be deanonymized using facial recognition or highly specific biometric identifiers like the iris pattern (Daugman, 2006).

In this paper, we focus on anonymizing the visual level of video data. Standard image and video processing tools offer filters that can be applied to the entire video area. But this basic approach can make relevant setting information invisible. Alternatively, manual annotation and anonymization of face regions is a highly time-consuming task, setting a hurdle for correctly anonymizing data from authentic social interaction.

Therefore, we explore the potential of current feature detection and machine learning (ML) techniques for the issue of automatically anonymizing persons in videos or images. We evaluate different models for face- and body-posture

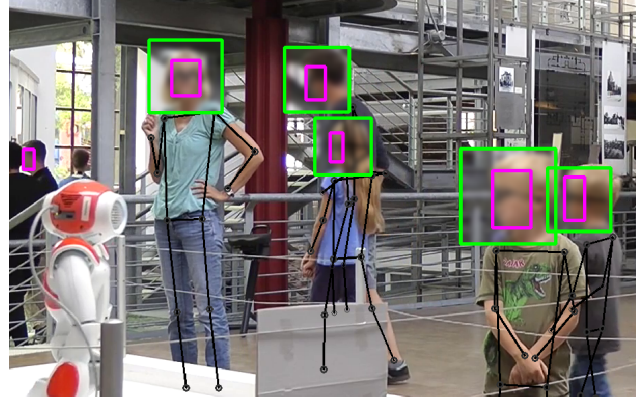


Figure 1: Sample frame from an anonymized video. The example shows the result of head-region anonymization using a blurring filter. Magenta-colored boxes show detected faces, and green boxes show the head region, estimated using key points from pose tracking (black dots and lines).

detection and optimize their parameters to maximize the number of correctly anonymized persons.

We provide an open-source toolkit <sup>1</sup> and workflow for anonymizing video recordings. The toolkit extends a web application for anonymizing still images (Krause et al., 2023). It has been developed within the data-reuse project "MuMoCorp" <sup>2</sup> to anonymize and release an existing multimodal corpus of human-robot social interaction (Pitsch, 2020) to the scientific community.

## 2. Video Anonymization Workflow

In this section, we describe a video anonymization workflow that combines automated analysis and manual control (Fig. 2).

<sup>1</sup><https://git.uni-due.de/mumocorp-open-access/anonymization>

<sup>2</sup><https://www.uni-due.de/kowi/mukom/mumocorp>



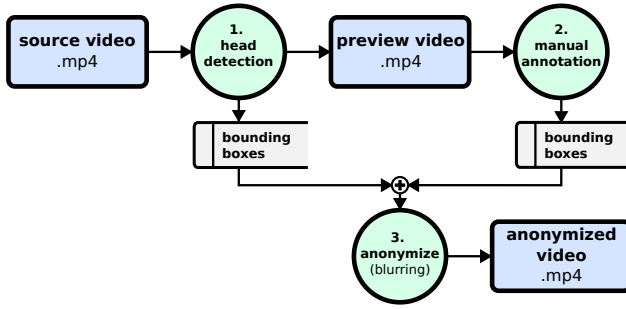


Figure 2: Video anonymization workflow, combining automated detection and manual control: 1. Automatic detection; 2. Manual annotation of missed detections and 3. Combining automatic and manual annotations to create the final, anonymized video.

It consists of three main steps:

1. The original video file is analyzed with ML algorithms for detecting the participants’ heads, faces, and body posture. These detections are rendered into a preview video, and bounding box data is stored in a separate file for later reuse.
2. The preview video is manually controlled by a human annotator. Missing bounding boxes are added using a video annotation tool. In our workflow, we use ”DIVE Desktop” (Dawkins et al., 2017).
3. The final pass uses both automatically detected and manually annotated bounding boxes to generate the anonymized video using, e.g., a blur filter.

This workflow has the additional advantage that non-facial regions can be manually marked and anonymized, e.g., with name tags or body decorations.

### 3. Automatic Detection of Face and Head Regions

In this next section, we will describe in greater detail the automation part of the workflow presented in section and explain which machine learning models are used for detecting face and head regions and how parameters have been optimized.

Generally, different options exist for automatically detecting a human face in a still image or video frame. An algorithm can attempt to detect a head or face region using facial features (e.g., eyes, mouth, nose, and eyebrows) and their spatial relationships (Omer et al., 2019; Payal and Goyani, 2020), or it could attempt to track a human’s body posture and infer the approximate head location and area.

Combining both face and pose tracking, almost all faces in a video can be automatically anonymized. Yet, in some video frames, the detection of the face or head area might still fail, e.g., due to motion blur or intermittent occlusions. Such detection gaps can be filled using a basic tracking algorithm that extrapolates the position of the face and head areas in the next frame based on the previous position and the movement velocity of the people in the video. This

tracking approach works best for videos recorded with a static camera, as is the case with our multimodal corpus of human-robot social interaction (Pitsch, 2020).

### 3.1. Face Detection

The first approach to detecting a person’s face employs deep-learning-based face detectors. We evaluated two state-of-the-art face detection models well suited for our anonymization task, the Dual Shot Face Detector (DSFD) and the RetinaFace Detector (RFD). DSFD was developed for challenging face detection situations, including bad lighting conditions, reflections, unusual makeup, blurry faces, and unusual face orientations (Li et al., 2019). RFD implements robust and fast single-shot face detection. Besides bounding box face detection, RetinaFace can provide facial landmarks and a robust 3D face reconstruction (Deng et al., 2020).

Both face detectors provide excellent detection rates but still occasionally fail to detect a face. For example, we observed that the models work reliably for fully visible faces but might fail for faces that are either partially occluded or only partially visible from diagonally behind.

Depending on specific anonymization requirements and the video material, one of the detectors (or future face detectors) could be dynamically selected based on detection confidence (see section 3.3.) or all used together with pose estimation for a potentially higher combined recall.

Both face-detectors and posture-based head region detection were systematically evaluated and their parameters optimized, as detailed in the next sections.

### 3.2. Head Detection using Pose Estimation

As a second approach, ML-based human pose estimation (Wang et al., 2022) can help to further reduce the chance of non-anonymized face areas (false negative pixels, see fig. 4), because a pose estimator holistically recognizes a human in a video frame, even if the head region might be occluded to a larger extent.

The position and size of the head region can be estimated using detected human-body key points. The center of the head region is calculated as the average of the detected nose, eye, and ear positions. The size of the head region is estimated by calculating the torso length (the average of the distance between the left shoulder and left hip and the right shoulder and right hip key points) and scaling it with a constant value.

### 3.3. Detection Parameters

The sensitivity of the face- and pose-detection methods can be adjusted and fine-tuned for the specific task requirements. The two main parameters offered by both methods are:

1. Confidence threshold: ML-based methods often provide a confidence value for detected objects. Reducing the confidence threshold might include more detections, thereby increasing the number of detected faces (true positive rate), but typically will also increase the number of spurious detections, i.e., the false positive rate. For person anonymization, where the cost of a

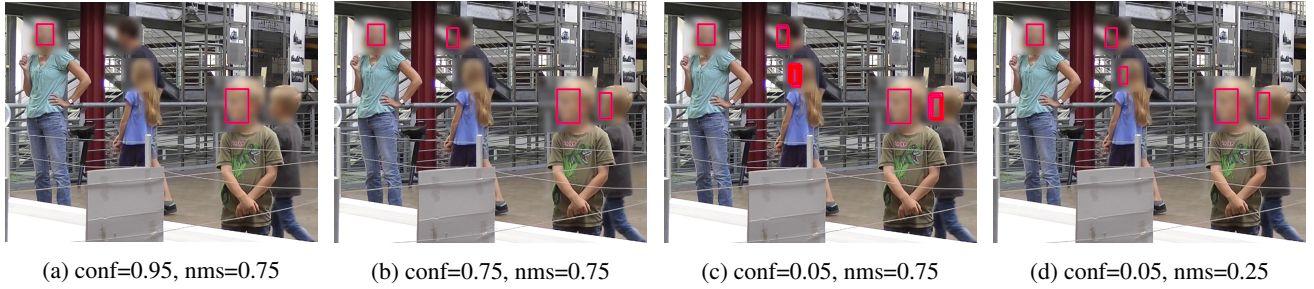


Figure 3: Effects of the two main face detection parameters. Lowering the confidence threshold (conf) increases the number of detected faces (compare (a, b, and c)). Multiple detections of the same face can be reduced using non-maximum suppression (nms, compare (c) with (d)).

missed face detection (false negative) is much higher than a false-positive detection, the threshold parameter should be set as low as possible.

2. Non-maximum suppression (NMS): Deep learning-based detectors may generate multiple, slightly different bounding boxes of varying confidence for the same object. NMS eliminates redundant bounding boxes and tries to select the optimal target boundary box (Gong et al., 2021).

The effect of these two main parameters is visualized in fig. 3, where faces in a still frame from our corpus were detected using the Dual-Shot Face Detector. Lowering the confidence threshold increased the number of detected faces, with very low thresholds producing multiple detections of the same face. This effect can be reduced using non-maximum suppression (though this is counterproductive for anonymization tasks; see next section 4.).

## 4. Detector Evaluation

Face detection methods typically generate bounding boxes (but methods exist that also provide the exact face outline). The accuracy of a face detector can be evaluated by comparing the detected bounding boxes with ground-truth labelled bounding boxes of a dataset.

### 4.1. Test-Dataset

To find optimal parameters, a small dataset with hand-selected frames from videos of a large multimodal corpus (Pitsch, 2020) was assembled. The corpus contains videos of an authentic, real-world situation in a museum, where people participated in a study on human-robot-interaction. The participants could freely walk and turn around to inspect exhibits, talk and interact with each other and the robot. Therefore, the videos are more difficult in terms of anonymization than, for example, videos recorded in a laboratory study with well-lit subjects recorded in a frontal view. The dataset shows faces and heads of persons in difficult-to-detect situations (e.g., partially occluded, overlapping, unusual poses, motion blurred, or under bad lighting conditions). It consists of 32 images containing 170 heads. 36% of faces are partially occluded, and 19% are facing away from the camera (only visible obliquely or fully from behind). Fig. 3 shows a sample frame from a video of the corpus. In this example, the challenge is that some faces are barely visible from behind.

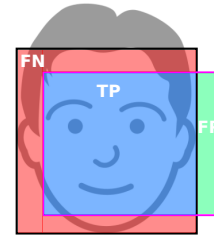


Figure 4: Face or head detection generates bounding boxes. Intersecting the annotated "ground truth" bounding box (black) with a detected, slightly misaligned bounding box (magenta) provides the number of true positive (TP, blue area), false positive (FP, green area) and false negative pixels (FN, red area). Face icon: CC 3.0, thenounproject.com

### 4.2. Evaluation & Parameter Optimization

The qualities of a feature detector are often visualized using a precision-recall curve. Intersecting a detected bounding box with the corresponding ground-truth bounding box provides the number of true-positive (TP) pixels, the number of false-positive (FP) pixels and the number of false-negative (FN) pixels; see fig. 4. Based on these values, two important metrics can be calculated: 1. The precision and 2. The recall of a detector:

$$\text{precision} = TP / (TP + FP)$$

$$\text{recall} = TP / (TP + FN)$$

A high precision value (close to 1) shows that a detector can find some target objects (e.g., faces) while minimizing false positive detections (i.e., detecting a face where no real face is located). A detector with a high recall value can find almost all objects (minimizing false negative detections, i.e., missing a truly existing face in an image), but often does so at the cost of reduced precision. For the purpose of anonymization, a high recall value is clearly preferred because the cost of a non-anonymized face is much higher than the cost of a false-positive detection.

The precision and recall of a detector are influenced by the parameters mentioned in section 3.3. Systematically changing, e.g., the confidence threshold from one to zero, yields the characteristic precision-recall curves for the different detectors.

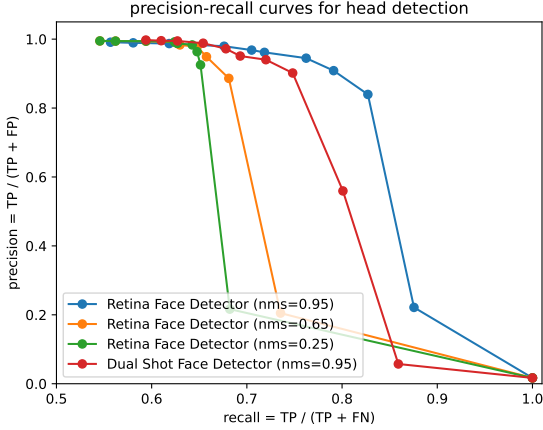


Figure 5: Precision-recall curves for the Dual-Shot Face Detector and the Retina Face Detector. Smaller confidence thresholds (decreasing from left to right; see table 1) result in better recall at the cost of decreasing precision. A larger threshold for non-maximum suppression (nms) improves both recall and precision in this dataset.

#### 4.2.1. Method

Each detector was tested on all images of our challenging dataset with three different values for non-maximum suppression (0.25, 0.65, and 0.95) and with twelve confidence thresholds ranging from zero to one using the values shown in table 1 (see first column). The generated bounding boxes were then used to set the values in a binary array, having the same size as the respective image, to true if that "binary pixel" is located inside a bounding box. This effectively generates a mask representing the union of all bounding boxes. A second mask was generated using the annotated ground-truth bounding boxes. Next, the number of TP-, FP-, and FN-pixels was calculated from both masks using logical operators, masking, and summation. The final step involves the calculation of precision and recall numbers, as detailed at the beginning of this section.

#### 4.2.2. Evaluation Results

Fig. 5 shows these curves for both the RFD and DSFD while also testing different values for the non-maximum suppression (nms) parameter.

The recall of both detectors is comparatively low because we observed that the bounding boxes generated by the evaluated face detectors often do not include the ears or, on rare occasions, parts of the chin or nose of a detected face. For the best possible anonymization, those facial parts should not be neglected.

One approach for increasing recall (at the cost of precision) is to expand the detected bounding boxes using two separate scaling factors for width and height expansion. Fig. 6 shows that bounding box expansion increases the recall while the precision stays at an acceptable level for anonymization purposes.

Further, the precision-recall curve for posture-based head-region detection is shown. Here, precision is even lower due to larger bounding boxes that cover the full head; see

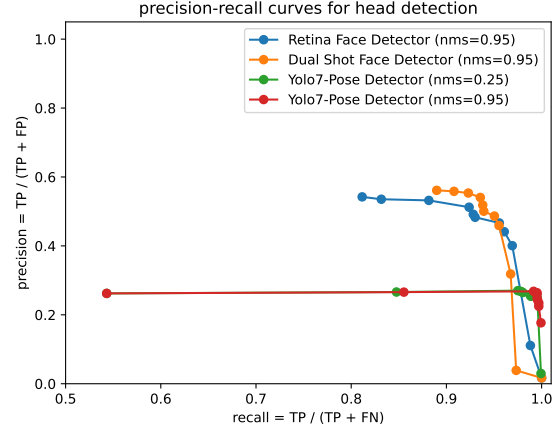


Figure 6: Precision-recall curves, this time with horizontally and vertically expanded bounding boxes for both face-detectors. Recall is substantially improved. The green and red traces show precision vs. recall for estimated head regions using the yolo7 pose detector.

	Detector					
	RFD		DSFD		Y7PD	
thr.	p	r	p	r	p	r
0	0.02	1	0.02	1	0.18	0.999
0.01	0.11	0.988	0.04	0.973	0.22	0.997
0.02	0.4	0.969	0.32	0.967	0.23	0.997
0.05	0.44	0.961	0.46	0.955	0.24	0.996
0.1	0.47	0.956	0.49	0.95	0.25	0.995
0.2	0.48	0.93	0.5	0.939	0.26	0.995
0.3	0.49	0.928	0.52	0.938	0.26	0.995
0.5	0.51	0.924	0.54	0.936	0.27	0.992
0.8	0.53	0.882	0.55	0.923	0.27	0.855
0.9	0.54	0.832	0.56	0.908	0.26	0.543
0.95	0.54	0.812	0.56	0.89	nan	0
1	nan	0	nan	0	nan	0

Table 1: Precision (p) - recall (r) values for the RetinaFace Detector (RFD), the Dual-Shot Face Detector (DSFD) and Yolo7 Pose Detection (Y7PD) with confidence threshold values (thr.) ranging from zero to one and a fixed nms of 0.95. The values correspond to the precision-recall curves shown in fig. 6.

the green boxes in fig. 1. But the recall values approach one for small confidence thresholds; hence, almost all faces and head regions are now detected.

## 5. Discussion & Outlook

For our anonymization purposes, full coverage of the faces of persons is very important. Other applications, where the bounding boxes should closely match the real face or head area, may require different threshold and nms parameters. A larger number of false-positive detections (e.g., due to expanded bounding boxes and low threshold values) was accepted for the task of anonymizing our specific corpus. We opted to combine bounding boxes from the retina face



detector and estimated head bounding boxes using yolo7 pose (confidence threshold = 0.1 and nms = 0.95 for both methods). Smaller threshold values than 0.1 yield a better recall, but occasionally resulted in very big bounding boxes covering large portions of a frame. Removing those spurious bounding boxes would have incurred an additional manual cleaning effort.

To further improve face detection reliability, it is planned to use a larger ensemble of face detectors together with weighted non-maximum suppression across detected bounding boxes from all detectors. On top, a predictive tracking algorithm like, e.g., a Kalman filter (Welch et al., 1995) could improve the tracking accuracy, especially for moving cameras.

Anonymization could be improved using more complex filters that may use less blurring to better preserve social cues but disturb face recognition by, e.g., applying a defined amount of random face morphing (Ferrara et al., 2022). Such filters should be carefully evaluated regarding their effects on privacy, both with respect to human and machine-based face identification capabilities.

Given the current technical capabilities, the question arises whether only a participant’s face or the entire body should be anonymized and whether the relevant parts should be made unidentifiable using classic methods (e.g., blurring) or whether they should be replaced with a deep-fake. This latter option is suggested by recent state-of-the-art frameworks like Deep Privacy 2 (Hukkelås and Lindseth, 2022). At first sight, deep-fakes maximize privacy. But such an approach reduces, removes, or - most problematic for analytic purposes - modifies social cues. For example, the deep-fake approach used in the Deep Privacy 2 framework can lead to different gaze directions, modified hand and body postures, different gender, age, and nationality. Hence, it is difficult to achieve a good balance between an acceptable level of anonymization and the potential loss of social cues that are important for interaction research.

Besides anonymization purposes, gathered body-posture trajectories, combined with 3D facial landmark localization (Khabaralak and Koriashkina, 2021) and gaze direction estimation (Ablavatski et al., 2020), will yield valuable data for multimodal conversation analytic methods and visualization.

## 6. Acknowledgements

This project was financed by the Volkswagenstiftung (grant number 90886, PI: Karola Pitsch).

## 7. Bibliographical References

- Ablavatski, A., Vakunov, A., Grishchenko, I., Raveendran, K., and Zhdanovich, M. (2020). Real-time pupil tracking from monocular video for digital puppetry. *arXiv preprint arXiv:2006.11341*.
- Daugman, J. (2006). Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, 94(11):1927–1935.
- Dawkins, M., Sherrill, L., Fieldhouse, K., Hoogs, A., Richards, B., Zhang, D., Prasad, L., Williams, K., Luffenburger, N., and Wang, G. (2017). An open-source platform for underwater image and video analytics. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 898–906. IEEE.
- Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.
- Ferrara, M., Franco, A., Maltoni, D., and Busch, C. (2022). Morphing attack potential. In *2022 International Workshop on Biometrics and Forensics (IWBIF)*, pages 1–6. IEEE.
- Gong, M., Wang, D., Zhao, X., Guo, H., Luo, D., and Song, M. (2021). A review of non-maximum suppression algorithms for deep learning target detection. In *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, volume 11763, pages 821–828. SPIE.
- Hukkelås, H. and Lindseth, F. (2022). Deepprivacy2: Towards realistic full-body anonymization. *arXiv preprint arXiv:2211.09454*.
- Khabaralak, K. and Koriashkina, L. (2021). Fast facial landmark detection and applications: A survey. *arXiv preprint arXiv:2101.10808*.
- Krause, A. F., Ferger, A., and Pitsch, K. (2023). Detecting and tracking persons in video recordings of authentic social interaction: Analysis and anonymization. accepted at CAMVA 2023.
- Kretzer, S. (2013). Arbeitspapier zur konzeptentwicklung der anonymisierungs-/pseudonymisierung in qualiservice.
- Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., and Huang, F. (2019). Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069.
- Omer, Y., Sapir, R., Hatuka, Y., and Yovel, G. (2019). What is a face? critical features for face detection. *Perception*, 48(5):437–446.
- Payal, P. and Goyani, M. M. (2020). A comprehensive study on face recognition: methods and challenges. *The Imaging Science Journal*, 68(2):114–127.
- Pitsch, K. (2020). Answering a robot’s questions: Participation dynamics of adult-child-groups in encounters with a museum guide robot. *Reseaux*, 220221(2):113–150.
- Roth, W.-M., Von Unger, H., et al. (2018). Current perspectives on research ethics in qualitative research. In *Forum qualitative sozialforschung/forum: Qualitative social research*, volume 19, pages 1–12. DEU.
- Rubinstein, I. S. and Hartzog, W. (2016). Anonymization and risk. *Wash. L. Rev.*, 91:703.
- Srivastava, B. M. L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., and Yamagishi, J. (2022). Privacy and utility of x-vector based speaker anonymization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2383–2395.
- Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolo7: Trainable bag-of-freebies sets new state-of-

the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Welch, G., Bishop, G., et al. (1995). An introduction to the kalman filter.

# **“Also ehrlich” – From adjectival use to interactive discourse marker**

**Lothar Lemnitzer<sup>1</sup>, Antonia Hamdi<sup>2</sup>**

<sup>1</sup>Berlin-Brandenburgische Akademie der Wissenschaften, <sup>2</sup>Universität Duisburg-Essen

E-mail: lemnitzer@bbaw.de, antonia.hamdi@stud.uni-duisburg-essen.de

## **Abstract**

In this paper we will take a closer look at the German word “ehrlich”. Traditionally, it is seen and described as an adjective. However, this word, as we will demonstrate with corpus data, has widened its domain of usage and is being used frequently, and in combination with other words, as an interactive unit (“interaktive Einheit”). As such, it is typically used in spoken discourse and in written dialogues, while having lost central aspects of its core meaning. Our findings are based on a German reference corpus and a corpus of Wikipedia discussion pages.

**Keywords:** interactive unit, German language, discourse analysis

## 1. Introduction

The German word „ehrlich” traditionally signifies a trait of human character („honest“, „sincere”) as well as a characteristic of human activity and its result („fair” as in „she acted fairly towards me” or „a fair deal”).

This type of usage is also registered in the dictionaries of contemporary German (Duden, DWDS, Wahrig etc.).

In the last decades, however, the word has come into a wider use while losing some aspects of its core meaning. It is nowadays frequently used as an interactive unit, typically in genres of spoken language and computer-mediated-communication. It co-occurs with a small set of (modal) adverbs. One of these collocations, „ehrlich gesagt” (en: „frankly speaking”), has already been subject of linguistic investigations. For example, Stoltenburg (2008) described the use of this phrase as a means to establish politeness in discourse. It allows speakers to distance themselves from former utterances and possible consequences of them (2008:275f.). Wolfgang Imo describes the phrase as an element of the comment adverb class (2012:70) and so does the Duden Grammar (2005:594).

In this paper, we will broaden the perspective and describe some other collocations with “ehrlich” in the framework of interactional linguistics (cf. Imo and Lanwer, 2019) and investigate their interactional function(s).

In section 2 we will formulate our research questions. Chapter 3 is devoted to a description of the theoretical framework on which our interpretation of the data is based. In section 4, the core of this paper, we will describe our database (corpora of various kinds) and our quantitative as well as qualitative data analysis. The examples are discussed and generalized in section 5. We finish this paper with conclusions and plans for further investigation.

## 2. The research question

In our daily use of German as native speakers and our use of interactive social media we stumbled upon a frequent use of the word „ehrlich” in contexts which do not support its usual meaning(s) as a qualitative adjective (en: „honest, sincere, fair”). We decided to take a closer look at it.

From this first intuitive observation two questions arose that we decided to investigate further with the use of several corpora a) what is the specific function of the word „ehrlich” when used in the non-traditional way? b) Which are the typical context of the word that “trigger” this particular function?

## 3. The theoretical framework

We will base our quantitative as well as qualitative analysis on the framework of „interactional linguistics”.

Interactional Linguistics is seen nowadays as an established sub-discipline of theoretical as well as applied linguistics. It originates in the work of Elizabeth Couper Kuhlén and Margret Selting (2000, 2001) and gained ground particularly in the linguistics community in Europe (cf.

Lindström 2009).

Interactional Linguistics investigates language in use quantitatively as well as qualitatively.

Recent investigations have shown that in the course of interaction, recurrent patterns emerge that become, over the time, more and more stable and lexicalized – grammatical constructions and idiomatic expressions are typical linguistic means to realise these communicative functions<sup>1</sup>. Such patterns are typically not syntactically integrated, they are an optional „add on” to the proposition(s) and act on a meta-pragmatic level<sup>2</sup>.

While the meta-pragmatic function of „ehrlich gesagt” has already been subject of linguistic investigations (see section 1), the functions of other patterns with „ehrlich” as their lexical core have still to be analysed. In contrast to recent studies that are based on (samples of) spoken language, we will be using corpora of written text and computer-mediated communication. Below, we will show why, in our opinion, this is an appropriate database.

Following Imo and Lanwer (2019) we will present a detailed qualitative analysis based on a small sample of patterns from these corpora. This a prerequisite for understanding the linguistic structures in dialogue in general and the place and function of the phrases which we will analyse in such linguistic structures.

## 4. Data Analysis

### 4.1 Quantitative Analysis

Based on the large corpora of the “Digitales Wörterbuch der deutschen Sprache” ([www.dwds.de](http://www.dwds.de), cf. Geyken et al., 2017) we looked at co-occurrences of „ehrlich” with other adverbs. As we have already pointed out, we will not include the phrase „ehrlich gesagt” into our analysis. As a result of exploring these co-occurrence data, we decided to narrow down our analysis on three pattern: „aber ehrlich”, „also ehrlich”, „mal ehrlich”. The reason for this decision is that these phrases frequently occur in a syntactically non-integrated position that is typical for interactive units (cf. Sieberg, 2016).

The distribution (relative frequency, calculated as parts per million) of the three phrases in the DWDS corpora is outlined in table 1. We chose the „Referenz- und Zeitungskorpus” (Z/R) and the „Webkorpus XL” (Web) from the DWDS. The latter corpus is around 10 times larger than the former corpus.<sup>3</sup>

Table 1: Relative frequency (ppm) in the DWDS corpora

Pattern	Z/R (ppm)	Web (ppm)
aber ehrlich	0,35	1,43
also ehrlich	0,05	0,38
mal ehrlich	0,51	4,51

In a second step, we broadened the data base to include the corpus of Wikipedia discussion pages. We checked various CMC corpora for occurrences of these patterns. However, in prototypical CMC corpora such as chat logs we could not find any or rather few examples. In the discussion section we explain why, in our opinion this is the case and

<sup>1</sup> Günthner (2009: 403) uses the term „sedimentierte Muster”.

<sup>2</sup> Torres Cajo 2017: 225.

<sup>3</sup> For corpus sizes, cf. <https://www.dwds.de/d/korpora/public> and

<https://www.dwds.de/d/korpora/webxl>. You have to be a registered user if you want to reproduce the results.

why the choice of corpus does matter.

Wikipedia discussion pages turned out to be an appropriate data source for our investigations and can, according to Beißwenger 2016 and Herzberg 2022, be seen as a (special) kind of CMC corpus.

The corpus is available via the corpus collection of the Leibniz-Institut für Deutsche Sprache via COSMAS IIweb. We drew our sample from the so-called wdd19 subcorpus<sup>4</sup>. Size of the corpus at the sampling date is 711.935 texts with 415.929.118 Tokens<sup>5</sup>.

The size of this second data sample is presented in table 2.

Table 2: Relative frequency (ppm) in the wdd19 corpus

Pattern	Wdd19 (ppm)
mal ehrlich	3,72
also ehrlich	0,66
aber ehrlich	1,80

After having collected, i.e. exported the corpus data into tables, we had to do some clean-up tasks. The patterns under investigation appear frequently as free combinations, i.e. the word „ehrlich“ is used in a literal sense (these are false positives) – such examples had to be removed from the data. As a consequence, we narrowed down our data set on such patterns where the word occurred in a non-integrated position<sup>6</sup>. In the reduced data set the share of false positives is much lower.

In our data, the patterns occur in sentence initial positions, with a comma or colon as a connecting element. Less frequently, the phrases are inserted into the sentence or, rarely, in sentence final position. They might also occur as independent, yet incomplete sentences.

We counted the positional distribution in a sample of the Wikipedia corpus, see table 3.

Table 3: Distribution of the patterns according to sentence position

Pattern	Sentence initial	Inserted	Sentence final	Isolated
Mal ehrlich	18/23	1/23	0/23	4/23
Also ehrlich	17/23	1/23	2/23	3/23
Aber ehrlich	3/5	2/5	0/5	0/5

However, due to the small size of the sample, these figures should be taken with a grain of salt. The tendency towards sentence initial position is nonetheless obvious.<sup>7</sup>

The ultimate goal was to select (and present in this paper) prototypical examples of the use of these phrases as interactional units as a basis for the qualitative analysis.

## 4.2 Qualitative Analysis

In what follows, we will present, for each pattern, two examples from the data and a resp. interpretation of these

examples. These examples are prototypical for the use of these phrases.

Instead of translating the examples, we only rephrase them. The core phrases of these examples, the reason why we present them, are language specific and their resp. sense or function will be lost in the translation process.

### 4.2.1. aber ehrlich

*Du betonst zu recht die Teamarbeit hier; ja, jeder kann und darf diesen Artikel verbessern. **Aber ehrlich:** Wenn dein Tonfall hier symptomatisch für deine Arbeits- und Argumentationsweise ist, prophezeie ich dir keine lange Zukunft hier.* (WDD19/D0100.30875 Diskussion:Dürkopp Typ P 16.)

(Participant initially agrees with the discussion partner but continues with criticising his / her style of discussing).

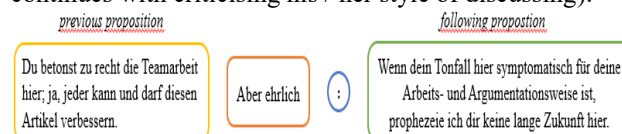


Fig. 1 : *Aber ehrlich*, example 1 above

*Auf die Frage, ob es ihn nicht störe, dass dieser Golfstaat die LGBT+-Symbole aus den Stadien und Straßen verbannt hat, sagte er: „Man muss anerkennen, dass Katar diese WM sehr gut organisiert hat. (...) Aber es stimmt, es gibt (hier) noch vieles zu regeln, es gibt viele Länder, wo noch vieles zu regeln ist. **Aber ehrlich,** seien wir jetzt erst mal glücklich.“* (WM-Euphorie in Frankreich: Liebe zu zwei Teams. TAZ Verlags- und Vertriebs GmbH, 2022-12-15)

(Moroccan soccer player marks the sceptical remark of an interviewer as being irrelevant in the situation of having won a game and refuses to answer it.)

### 4.2.2. also ehrlich

*Bleibt die Aussage, der Artikel sei oberflächlich und mit nicht verarbeiteter Primär- und Sekundärliteratur vollgestopft. **Also ehrlich:** Wenn man sich nicht die Mühe macht, Probleme wirklich herauszuarbeiten und zunächst auf der Artikel-Disk zur Diskussion zu stellen, dann erwarte ich wenigstens, dass man sich mit der Artikel-Historie auseinandersetzt und herauszufinden versucht, wer die Autoren waren, wie deren Vorgehensweise zu beurteilen ist und was das über die Qualität des Artikels aussagt.* (WDD19/A0055.03746 Diskussion:Atlantis/Archiv/012.)

(Participant bluntly refuses the proposal of the discussion partner and answers with his / her own proposal what should have been done instead)

*Viele Menschen glauben, dass sie bereits mit wenig Selbstvertrauen geboren wurden. **Also ehrlich,** was für ein Blödsinn! Niemand kommt mit einem Mangel an Selbstvertrauen auf die Welt. (Ich zeig dir, wo der Hammer hängt – Die 12 besten Wege zu mehr Selbstvertrauen.* Hafawo, 2015-07-05)

sentence or clause and are followed by a clause or sentence delimiter (comma, full stop, semi-colon...).

<sup>7</sup> Also see Imo (2007: 63), „die Satzperipherie [ist] generell der ideale Ort für metapragmatische Marker ...“.

<sup>4</sup> wdd19 = Wikipedia-Diskussionen zu Artikeln bis 2019.

<sup>5</sup> <https://cosmas2.ids-mannheim.de/cosmas2-web/faces/investigation/queryString.xhtml> (after registration).

<sup>6</sup> In short, “syntactically non-integrated positioned” has been defined by us as: the two words appear at the beginning of a

(An opinion that is claimed to be shared by many people is clearly rejected by the author (*was für ein Blödsinn* [en „what a rubbish”]) and countered by a contradictory opinion of the author.)

#### 4.2.3. mal ehrlich

*Die Enttäuschung kommt oft aus deutschen Generalstabskreisen, die vergeblich auf eine Unterstützung ihrer Putschpläne 1938 durch England hofften. Doch **mal ehrlich**: welches Land hat je offiziell Putschisten gegen eine legitime Regierung unterstützt? Unreal, aber genug als Ausrede für mangelnde Handlungsbereitschaft. Könnte man meinen.* (WDD19/A0059.74030 Diskussion:Appeasement-Politik/Archiv)

(Participant answers an opinion with a rhetorical question and thus challenges its relevance)

*Die Kosten für die Rufnummernmitnahme wurden zwar vom Gesetzgeber auf 6,82 Euro gedeckelt, bei fraenk wird aber von vornherein keine Gebühr erhoben. Das ist fair. ... Kein MMS-Versand möglich. **Mal ehrlich**, wer braucht MMS? (fraenk im Test: Tarif im Telekom-Netz mit mehr Datenvolumen. GIGA, 2023-04-05)*

(Participant counters the opinion of somebody else with a rhetorical question that challenges the relevance of this opinion).

There are of course less prototypical examples with the phrase (in the following the phrase „mal ehrlich”) as an interactive unit in communicative function. In the following examples, the phrase is addressed to a statement of the speaker himself (or herself). The primary function is to put more emphasis to this proposition (in the sense „believe me, this is true”):

*Es ist nichts los in Rostock. Mal ehrlich, das Leben pulsiert nicht gerade in den Straßen.* [Nothing happens in (city of) Rostock. Life on the streets is all but vibrating]

*Taylor sichert den Eckball und begibt sich in den Strafraum, um den Ball selbst einnicken zu können. Mal ehrlich, neun von zehn Stürmern hätten das so gemacht.* [Taylor saves the corner ball in order to head it in himself. Nine out of ten strikers would have done so.]<sup>8</sup>

## 5. Discussion of the examples

With the use of all the aforementioned, prototypical interactive units speakers establish a link between a previously mentioned proposition (of the partner in the dialogue or discussion) and their own response to that proposition. Therefore, these interactive units are of the responsive type. We deliberately included in the examples in section 4.2 as much context as needed to understand the specific role and function of the interactive unit. Second,

the speaker sets the tone or frames his / her own response (fig 1). In non-prototypical uses, one of these functions can be missing, as we have shown above.

With the use of „mal ehrlich”, the speaker’s proposition typically has the form of a rhetorical question that (indirectly) challenges the proposition that is addressed (as useless, superfluous etc). On the one hand, this phrase is closest to the core meaning of „ehrllich”, as it could be rephrased with *be honest, face the facts*. On the other hand, it belongs to the group of constructions that assume a commentary function<sup>9</sup>. It typically occurs in a syntactically non-integrated position a fact that strengthens our analysis that the phrase is used to organise the discourse by linking two propositions. Semantically, the phrase signals that the previous proposal etc. is challenged, whereas „mal ehrlich” expresses the mildest form of challenge.

That the proposition takes the form of a rhetorical question indicates that it is addressed not only to the other participant of the dialogue, but to a broader audience. The speaker indirectly asks for approval of his / her position by the audience

„Aber ehrlich” opens a statement of the participant that is more confronting and less polite than „mal ehrlich”. „Aber ehrlich” not only serves as a link to a previous proposition of the discussion partner (see fig. 1), but also introduces a kind of face saving interaction. Thus, the possibly face threatening statement<sup>10</sup> that follows the interactive unit is mitigated by it in a way that a respectful tone of communication is maintained. Consequently, we can assign the function of establishing or maintaining an atmosphere of respect as the core function of this phrase.

With „also ehrlich” a previous opinion, proposal etc. is rejected and replaced by the opinion, proposal etc. that the participant claims to be the appropriate one. The latter is the core of the proposition that follows the initial phrase. „Also ehrlich” is the most blunt and direct reaction to the challenged proposal, there is an undertone of indignation connected with that impression. With the word „also” as part of the phrase, the intention of the discussion partner to maintain the turn and follow with an own proposal is enforced. The core meaning of the word „ehrllich” is nearly completely lost in this phrase.

It is common to these formulaic expressions that they „modalize” a proposition. The semantically emptied core word „ehrllich” is reduced to the function of a marker that designates the attitude of the speaker (in short: being sincere). They realise a meta-pragmatic framing of the proposition that corresponds well with the syntactically non-integrated position – typically the sentence initial position. They are directed either to the previously mentioned or the following proposition or to both and assume a pivotal function.

„Ehrllich gesagt” addresses potentially face-threatening utterances of the speaker and offers the dialogue partner the chance to dismiss his or her position and thus save him-/herself from the threat. This function could not be found in the data for either of the interactive units that we have

<sup>8</sup> Both examples are taken from the DWDS Webkorpus.

<sup>9</sup> Imo (2012: 80ff) uses the terms “Projektoronstruktion” or „Kommentarphrase”. We are, however, not aware of an English

equivalent for these terms.

<sup>10</sup> Cf. Brown/Levinson, 1987.

presented here. All these expressions strengthen the speaker's own proposition and can be seen as a more or less strong confrontation with the (given) proposition of the dialogue partner. The validity and coherence with a discursive world that has been established by the interacting dialogue partners is at stake here. Furthermore, these expressions are typically realised as a part of a „familiar“ style register (in German: *nähesprachlich*). Even the majority of examples from the DWDS corpora can be seen as oral register in a written medium.

A final remark on the choice of corpora for this investigation is at place here. We were surprised by the fact that the corpora of the DWDS, which mainly include written texts, provided rich material, i.e. many diverse examples of those interactive units that we have been focussing on, while central or prototypical CMC corpora did not. In light of the fact that these interactive units also have a pivot role between two propositions, we understood this issue better. Corpora of dynamic media such as chats do not lend themselves for such investigations. We have learned that there is an interaction between the research query, the phenomena to be investigated and the data.

## 6. Conclusions and further work

With „*ehrlich*“, we have examined but one lexical unit that is frequently and recurrently being used, in combination with other (modal) adverbs, in dialogic functions. The core meaning of the word is more or less opaque in these kinds of usages. Dictionaries of contemporary German should therefore register these kinds of uses of „*ehrlich*“ alongside a description of its proper meanings. Dictionaries of contemporary spoken German such as LeGeDe<sup>11</sup> would be even a better place for an account of such phrases. Furthermore, these and similar phrases can be used in the teaching of German as first and second language. They can serve to illustrate the function of such linguistic entities (discourse markers) as means to maintain an atmosphere of politeness and respect even in confronting situations (face saving). This is particularly important in digitally mediated communication where the participants do not see or even know each other. We can imagine to present and analyze such formulae as „*mal ehrlich*“ in connection with types of argumentation, in particular such with indirect or normative arguments<sup>12</sup>. Teachers can also introduce them as a part of a toolkit of „modalizers“<sup>13</sup> in the context of teaching / learning strategies of argumentation. They can raise awareness for the special, non-propositional function of these elements that is in many cases overlooked or even misinterpreted by students. In particular, the examples from the Wikipedia corpus provide illustrative material for this teaching goal.<sup>14</sup> Such teaching models are in the spirit of a didactical move in Germany towards the integration of (corpora of) authentic language into the classroom.<sup>15</sup> As a result of such analyses, students are able, after having

investigated authentic examples with these phrases, to understand the strategic reference to politeness as a concept that underlies or frames this type of discourse, and, more generally, as a reference to a socially shared and accepted value that underlies this kind of discourse.

An important issue for further work is: does this kind of exemplary analysis scale up? In other words: can we find further bigrams, trigrams etc. of words that assume the same functional roles, i.e. as interactive (responsive) units?

The detection of such interactive units on a broader scale is not possible with more quantitative data analysis for the following reasons. Typically, both words (in the case of binary units) are highly frequent. Therefore, statistics used for co-occurrence analysis is not reliable. The recall will be very low since the absolute frequency of occurrence of both parts of a the potential lexical unit plays a major role in most statistics such as MI or LogDice. Accordingly, the „Wortprofil“ of the DWDS lists only „*mal ehrlich*“ as significant co-occurrence for the word „*ehrlich*.“ On the other hand, such statistics produce too many so-called false positives: significant co-occurrences which are irrelevant for our purposes.

A promising alternative is pattern based analysis. The idea is to break down the concept of syntactically isolated into search queries on the corpora using their query languages. In DDC for the corpora of the DWDS, we can formulate the search query `"$p=ADV WITH $.=0 $p=ADJ* $p={'$', ' '$. '}"`<sup>16</sup> to be interpreted as „retrieve patterns of adverb in sentence-initial position, followed by an adjective followed by a clause or sentence delimiter (comma, semi-colon, full stop etc.)“. We will surely get many false positives, but sorting the data by their frequency of occurrence will help to find the interesting patterns. The DDC/DWDS query `COUNT (" $p=ADV WITH $.=0 $p=ADJ* $p={'$', ' '$. '}" ) #BY[$w, $w+1] #DESC COUNT` will sort the patterns by their frequency of occurrence. For the Kernkorpus of the DWDS the search engine retrieves a list, with „*nun gut*“, „*so weit*“, „*sehr gut*“ on top of it.<sup>17</sup>

We also have to face a loss in recall because we will not get the „ADV ADV“ patterns, or the „ADV INTJ“ patterns. However, we can find them if we apply a larger set of queries on the corpora.

These kinds of queries are purely explorative and only gives us an idea of which patterns are worth a closer look with an extended data search, data analysis etc. Nevertheless, the above question can be answered positively – yes, it scales up.

Corpus queries of this kind and complexity are probably not easy to reproduce on smaller corpus collections and their resp. search engines. We therefore recommend to perform the first, exploratory step on the DWDS corpora

<sup>11</sup> <https://www.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch/>

<sup>12</sup> Cf. Schurf/Wagener 2016, 303.

<sup>13</sup> Cf. <https://de.frwiki.wiki/wiki/Modalisateur>.

<sup>14</sup> Cf. Hug 2017 as a example for such a teaching unit.

<sup>15</sup> In German: „Bildungsstandards“. To learn more about the current discussion on the nationl levein Germany cf. <https://www.kmk.org/themen/qualitaetssicherung-in-schulen/>

[bildungsstandards.html](https://www.kmk.org/themen/qualitaetssicherung-in-schulen/).

<sup>16</sup> For further details please consult the DDC documentation page at <https://www.dwds.de/d/korpussuche> or the DWDS/DDC query tutorial at <https://www.dwds.de/b/category/tutorials/> (both texts are in German).

<sup>17</sup> Note that the first word, due to its sentence-initial position, is always written with a capital letter, but this is not relevant in our context.



and to follow up with more detailed data collection requests afterwards. We cannot yet give a decisive answer for the Korap search engine at the Leibniz-Institut für Deutsche Sprache in Mannheim. Some of the corpora are part-of-speech-annotated, which is a necessary pre-requisite, and some are not.<sup>18</sup>

For the examples in this paper, we would like to broaden the scope of our investigation in two regards: a) sifting through corpus examples, we encountered more complex, but still recurrent co-occurrences such as: „*aber mal ehrlich*“, „*also mal ehrlich*“. Differences between the henceforth described phrases and those more complex ones (and other combinations) should be further investigated, b) there are similar words in German that are on the way of assuming particular functions as discourse markers, thereby losing their core meaning, e.g. *gut*<sup>19</sup>. We want to include those words in our further investigation.

## 7. References

- Arens, Katja (2023): Strukturieren und Evaluieren im Gespräch. Lexikalische Diskurspartikeln als Ressourcen der Gesprächsorganisation. Heidelberg: Winter.
- Beißwenger, M. (2016). Praktiken in der internetbasierten Kommunikation. In Deppermann, Arnulf/Feilke, Helmut/Linke, Angelika (Eds.), *Sprachliche und kommunikative Praktiken*. Berlin, Boston: De Gruyter, pp. 279-311. [online: <https://www.degruyter.com/document/doi/10.1515/9783110451542-012/html>].
- Brown, P. and Levinson, S.C. (1987): *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.
- Couper-Kuhlen, E. (2021). Language over Time. Some old and new uses of OKAY in American English. *Interactional Linguistics* 1 (1) pp. 33-63 [online: <https://researchportal.helsinki.fi/en/publications/language-over-time-some-old-and-new-uses-of-okay-in-american-engl>].
- Couper-Kuhlen, E. and Selting, M. (2001). *Studies in Interactional Linguistics*. Amsterdam: John Benjamins.
- Dittmar, N. (2002). Lakmüstest für funktionale Beschreibungen am Beispiel von auch (Fokuspartikel, FP), eigentlich (Modalpartikel, MP) und also (Diskursmarker, DM). In: Fabricius-Hansen, C., Leirbukt, O. and Letnes, O. (Eds.): *Modus, Modalverben, Modalpartikel*. Trier: Wissenschaftlicher Verlag, pp. 142-177.
- Geyken, A.; Barbaresi, A.; Didakowski, J.; Jurish, B.; Wiegand, F. and Lemnitzer, L. (2017). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). *Zeitschrift für germanistische Linguistik*, 45 (2), pp. 327-344.
- Günthner, Susanne (2009): Konstruktionen in der kommunikativen Praxis. Zur Notwendigkeit einer interaktionalen Anreicherung konstruktionsgrammatischer Ansätze. In: *Zeitschrift für germanistische Linguistik* 37(3). S. 402-426.
- Hug, Michael (2017): „Es sollte vielleicht ...“ – Modalisieren beim argumentativen Schreiben. *Praxis Deutsch* 262. S. 50-59.
- Imo, Wolfgang (2007): *Construction Grammar und Gesprochene-Sprache-Forschung: Konstruktionen mit zehn matrixsatzfähigen Verben im gesprochenen Deutsch*. Tübingen: Max Niemeyer Verlag.
- Imo, W. (2012). Wortart Diskursmarker?. In Rothstein, B. (Ed.), *Nicht-flektierende Wortarten*. Berlin, Boston: De Gruyter, pp. 48-88.
- Imo, W. and Lanwer, J.-P. (2019). *Interaktionale Linguistik. Eine Einführung*. Springer Verlag.
- Lemnitzer, Lothar und Nils Diewald (2022): Abfrage und Analyse von Korpusbelegen. In: *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium*, ed. by Michael Beißwenger, Lothar Lemnitzer, Carolin Müller-Spitzer, utb:Fink, Stuttgart, pp. 374-390.
- Lindström, J. (2009). Interactional Linguistics. In D'hondt, S.; Östman, J.A. and Verschueren, J. (Eds.), *The pragmatics of interaction*. pp. 96-103.
- Schurf, B. and Wagener, A. (2016): *Texte, Themen und Strukturen. Deutschbuch für die Oberstufe*. Nordrhein-Westfalen. Berlin: Cornelsen.
- Selting, M. and Couper-Kuhlen, E. (2000). Argumente für die Entwicklung einer 'interaktionalen Linguistik'. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 1, pp. 76-95.
- Sieberg, B. (2016). Reaktive. Vorschlag für eine Erweiterung der Kategorie Responsive. In Handwerker, B.; Bäuerle, R. and Sieberg, B. (Eds.), *Gesprochene Fremdsprache Deutsch* (= Perspektiven Deutsch als Fremdsprache, Band 32), Baltmannsweiler: Schneider Verlag Hohengehren, pp. 101-117.
- Stoltenburg, B. (2009). Was wir sagen, wenn wir es „ehrlich“ sagen... Äußerungskommentierende Formeln bei Stellungnahmen am Beispiel von „ehrlich gesagt“. In *Grammatik im Gespräch*. De Gruyter, pp. 249-280.
- Storrer, A. and Herzberg, L. (2022). Alles okay! Korpusgestützte Untersuchungen zum Internationalismus OKAY. In Beißwenger, M.; Lemnitzer, L. and Müller-Spitzer, C. (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium*, utb:Fink, Stuttgart, pp. 37-59.
- Torres Cajo, Sarah (2017): "das is SO lächerlich; ohne SCHEISS jetzt ma" – Zur affektiven Äußerungsmodalisierung durch *ohne Scheiß*-Konstruktionen im gesprochenen Deutsch. In: *Sprachreport Gesprächsforschung* 18. S. 223-240.

<sup>18</sup> For further investigation into this issue we recommend <https://korap.ids-mannheim.de/> and the search engine synopsis in Lemnitzer / Diewald (2022).

<sup>19</sup> Katja Arens dedicates a part of her dissertation thesis to this particle, cf. Arens 2023.

# **The recontextualization of expert knowledge: intertextual patterns in digital science dissemination**

**Rosa Lorés**

Research Institute of Employment, Digital Society and Sustainability (IEDIS)  
Department of English and German Studies  
Universidad de Zaragoza (Spain)  
E-mail: [rlores@unizar.es](mailto:rlores@unizar.es)

## **Abstract**

Technology is having an unprecedented impact on the communication of specialized knowledge, which takes advantage of the use of digital modes and media to reach multiple, diversified audiences. To serve this purpose, expert knowledge is subjected to various processes of recontextualization. I here explore the intertextual patterns used in digital scientific dissemination to recontextualize expert knowledge in order to reach less specialized audiences, as well as the role played by digital affordances to shape the intertextual patterns identified. For such purposes, I focus on a corpus of 30 online scientific feature articles, where, among other features, the use of quotation practices is explored as well as possible emerging intertextual patterns. Results reveal the existence of patterns of intertextuality in the feature articles analysed, in which digital affordances (i.e. hyperlinks) combine with other “offline” intertextual resources for different recontextualization purposes and in various ways, depending on the level of expertise and specialization aimed at in the text.

**Keywords:** intertextuality, digital scientific communication, recontextualization, feature articles

## 1. Introduction

More than ever in the history of humankind, our contemporary technological, globalized, world is fostering the transfer of information across individuals and communities. Focusing on the communication of science, specialized or expert knowledge is now making use of digital modes and media to reach multiple, diversified audiences, under the form of a wide array of practices. To serve the communicative purposes that these digital practices have, specialized knowledge is subjected to various discursual processes of transformation, in an attempt to adapt to other communicative contexts, other audiences, and other purposes, different from those in which and for whom this expert knowledge originated. A major transformation process taking place in digital scientific communication is recontextualization, which refers to the processes by which information is appropriated and manipulated for different contexts (Bernstein, 1996; Linell, 1998; Calsamiglia & Van Dijk, 2004; Bondi et al., 2015; Johansson, 2019). As Giménez et al. (2020: 296) claim, these processes usually involve “repurposing the intended meanings of the original text to meet the new focus and real or perceived expectations of a new audience and thus the need to examine multiple instances of re-contextualization”.

Therefore, recontextualization does not only imply the reformulation, rephrasing or even simplification of textual material. Recontextualization goes beyond a mere textual exercise and involves processes of reinterpretation and reshaping (Luzón, 2013; Li, 2015; Giménez et al., 2020; Engberg, 2021), sometimes of resemiotization, frequently going beyond the verbal mode.

A major instance of recontextualization in the dissemination of science is intertextuality, defined by Bazerman (2004: 86) as “the explicit and implicit relations that a text or utterance has to prior, contemporary and potential future texts”. Practices of intertextuality basically allow bringing other voices into the discourse (Kristeva, 1986; Fairclough, 1992), thus contributing to the recontextualization of specialized knowledge for other purposes and audiences. According to Farrelli (2020: 2) “[a]t the level of discourse, social practices have patterns of intertextuality whereby some text-types, sources of text or even specific texts are typically referred to whilst others are not...”. What is important, though, is not which texts are referred to, but how they are used and what they are used for, that is, how intertextuality contributes to social action (Bazerman, 2004; Luzón, 2023).

My contention here is that, in digital scientific dissemination as in any other given social practice, these typicalities exist, and specific intertextual patterns can be identified as contributors to the recontextualization of expert knowledge for multiple audiences.

Thus, in the present study the following research questions are posed:

1. Which intertextual patterns are used as discursual resources in digital scientific dissemination practices to recontextualize expert knowledge?

2. How do the identified intertextual patterns contribute to the recontextualization of specialized knowledge for digital scientific dissemination?

I seek to investigate these research questions in one of the more widespread genres for the digital dissemination of scientific knowledge, the online feature article.

Scientific feature articles are usually written by journalists and aim to provide background information on a newsworthy topic, drawing on several expert or specialized sources. They differ from breaking news articles in that instead of reporting news about a particular situation, they share a general perspective on a subject. They attempt to be engaging and usually offer a personal perspective.

One of the main challenges in the exploration of intertextuality in digital discourse is how to operationalize it. Several attempts which have inspired the present proposal have been made to tackle this type of analysis. Thus, in order to explore the form and function of intertextuality within the perspective of CDA, Farrelli (2020) proposes four key concepts: inter-text, network of intertexts (groups of texts that connect to each other through intertextuality), networks of social practices and typicality (what patterns of intertextuality are typical for a social practice). Myers (2003) raises complex questions which have to do with genre issues on the understanding that scientific discourse involves a range of genres and practices and that popularizations are an important part of this range; Giménez et al. (2020) incorporate the concept of “re-entextualization” (Blommaert 2005) to describe how texts are decontextualized, refocused and reorganized, after a number of transformations. To analyze the processes of entextualization and re-entextualization of scientific knowledge, Giménez et al. (2020) propose to focus on two levels of analysis: textual elements (grammatical features, lexis, etc.) and rhetorical functions (attribution of authority, meaning making, visual purpose, etc.)

Enlightened by these scholarly contributions, the present study focuses on the following aspects:

1. the use of direct and indirect textual quotations
2. the lexical elements used to introduce these quotations (i.e. use of proper or general names)
3. the type of textual networks created, that is, the type of texts to which feature articles are related by means of intertextual practices
4. the intertextual patterns which emerge as a result of the intertextual markers and relations identified.

The exploration of these four aspects and the insights gathered from the results obtained may contribute to the understanding of intertextuality as part of the recontextualization practices which characterize the discourse of digital science dissemination.

## 2. Methodology and corpus

For the purposes of the present study I collected a corpus of 30 feature articles from two science popularization digital sources, written by a variety of journalists: *Smithsonian Magazine* (<https://www.smithsonianmag.com/>) and *Popular Science*

(<https://www.popsci.com/>), all of them published during the last two years (since February 2021 till March 2023), dealing with the topic of health, both physical and mental. Whereas the corpus is apparently limited in size, it is large enough to carry out an intense exploration of intertextual markers, networks and patterns, as intended. As Bazerman (2004: 91) says, intertextual analysis is quite intensive. Intensive analysis should avoid extensive corpora, at least till patterns are identified and can then be explored in larger collections of texts.

The methodological procedure applied was the following: first, the browser extension GoFullPage was used to download all the web texts in pdf format, including multimodal elements. Then, the verbal content of the web texts was transformed into word texts for further manual analysis when needed. Three texts from each digital publication were manually analyzed by way of a pilot study. As a result, a taxonomy of features to operationalize intertextuality corresponding to the four aspects mentioned above was identified. Next, the NVivo Pro software tool for qualitative and quantitative analysis was used for the exploration of the whole corpus. This time, the extension NCapture was used to upload all the files in the software tool, which was fed with the categories previously identified. Then, the whole corpus was analyzed by attending to these categories. As a result, the initial proposal was revised and a more fine-grained operating taxonomy of intertextual features emerged.

This taxonomy is, therefore, both a methodological toolkit and the outcome of the exploration of the digital scientific feature article as a recontextualizing practice of expert knowledge.

### 3. Results and discussion

The analysis of the corpus of online feature articles compiled for the purposes of the present study resulted in the identification of three types of intertextual references:

- 1) **Text quotations**, which might be either i) direct quotations of text of various lengths or of single concepts (the source being experts and researchers in the topic of the article), and ii) indirect quotations, in which the source can be introduced by using a proper name (usually the name of the researcher(s)), by using general names (i.e. *researchers/experts/scientists...*), or by means of abstract rhetors (Hyland, 1996), that is, nouns referring to research activity (i.e. *the study/experiment/research*).
- 2) **Digital quotations**, which here refer to texts incorporated into the article by means of the digital affordance of hyperlinking. Hyperlinking is used to create a variety of intertextual networks which connects the feature article with both monomodal (only verbal) and multimodal texts, including texts already existent in the “offline world” and, also, “digitally-indigenous” texts. As observed, the texts hyperlinked range from more specialized sources (research articles and research article abstracts) to less expert (breaking news and posts in the same magazine or in other popularizing sources). They also include reference material found on the

Internet (i.e. Wikipedia), information found in websites of public and private institutions, and even media artifacts, such as podcasts, links to TV series, and YouTube videos.

Digital affordances allow more sophisticated ways to create intertextual networks than those traditionally found in offline texts, with cases of “embedded quotation”, where other texts are referred to simultaneously using a textual quotation (indirect) and a digital quotation (hyperlink) to the full source from which the quotation is taken.

no cure. In a small pilot study, Kirkland, Tchkonina and collaborators administered D+Q to 14 people with the condition, three times a week for three weeks. They [reported notable improvement](#) in the ability of participants to stand up from a chair and to walk for six minutes. But the  
[*Smithsonian* 8]

- 3) **Multimodal quotations**, which are embedded elements that are copy-pasted from other sources and recontextualized in online feature articles. Examples of these digital quotations are pictures, visuals illustrating the information given in the text, graphs and figures bringing real data into the popularized text, videos showcasing the research under focus, and mass media products including extracts from a TV series.

The quantitative analysis of the corpus under study yielded some interesting results.

Starting with **textual quotations**, items were identified for both direct and indirect quotations. Direct quotation is made by attributing text to the authors of the study or studies reported or to other experts on the topic, identified by their proper name (and position).

Quotations were text stretches of various lengths, ranging from one single concept to a collection of sentences.

Textual quotations	Direct quotation		
	Text attributed to experts	Text attributed to non-experts	Total
<i>Popular Science</i>	111 /6.15	10 /0.55	121 / 6.69
<i>Smithsonian</i>	233 /7.65	3 / 0.1	236 / 7.74
Total	344 /7.09	13 /0.27	357 /7.35

Table 1. Raw numbers and frequency of direct quotations per 1000 words.

As observed, direct quotation from expert sources meant 96.36% (344 out of 357) of all the instances of direct quotation recorded. Slight differences are found between the two publications with respect to text attributed to other sources (non-experts) but the frequencies are so small that these differences are by no means significant.

Indirect quotation was introduced both by general nouns referring to the sources (i.e. *experts, researchers, informants...*) or by mentioning experts or non-experts by name. The use of abstract rhetors (i.e. *study, experiment, research*) was also part of this intertextual practice.

Textual quotations	Indirect quotation			
	Text attributed to experts	Text attributed to non experts	Text attributed to abstract rhetors	Total
<i>Popular Science</i>	56 /3.1	19/1.05	29 / 1.6	104 /5.76
<i>Smithsonian</i>	167 /5.48	29 /0.95	28 /0.92	224 /7.35
<b>Total</b>	223 / 4.59	48/0.99	57 /1.18	328 /6.76

Table 2. Raw numbers and frequency of indirect quotations per 1000 words.

The quantitative data show that, as expected, the appeal to experts as source is higher than to other sources. Thus referring to researchers either through general nouns or by proper name means almost 68% (223 out of 328) of all the indirect quotations found in the corpus, whereas referring to non-experts sources means only 14.6 % (57 out of 328) of the cases, with abstract rhetors having a stronger presence (17.4%). Differences were also found regarding the lexical elements used to introduce indirect quotations, with experts identified by their proper name being the major lexical pattern used, with 178 cases recorded in the corpus, meaning 54.3% of all the indirect quotations recorded.

No significant differences are found between both publications perhaps apart from the fact that *Popular Science* only referred to non-experts by name whereas in *Smithsonian* indirect quotations from non-experts were introduced by general nouns (i.e. people) or by proper name to similar extents.

In all, contrasting the use of direct and indirect textual quotations, we observe that there is a balance between both types in terms of frequency, and very similar data are gathered:

Textual quotations	Direct quotations	Indirect quotations
<i>Popular Science</i>	121 / 6.69	104 /5.76
<i>Smithsonian</i>	236 / 7.74	224 /7.35
<b>Total</b>	357 /7.35	328 /6.76

Table 3. Raw numbers and contrastive frequency per 1000 words of direct and indirect quotations.

With regard to the position in the text in which general and proper nouns were found, tendencies were observed for proper nouns to appear at the beginning of the text, where the study is introduced and details are given of the expert voice to which the scientific knowledge is attributed. As expected, appealing to this expert voice from the very beginning confers the journalist with the *auctoritas* requested to confer credibility to their popularizing texts. No differences, though, were recorded in relation to any strategic positioning of textual direct and indirect quotations, which were found to combine along the text without any predictable pattern.

**Digital quotations** were another major type of intertextual practice identified in digital feature articles. Characterized by the technical affordance of hyperlinking, instead of

incorporating text as such, as textual quotations do, digital quotations encourage readers to explore other sites (texts, webs, blogs, media) in search of complementary scientific knowledge. Digital quotations were found to establish intertextual networks with research articles, research article abstracts, reports and information from institutional websites, popularizing texts from the same and other sites, or breaking news in digital newspapers. The tables below show data for the different types of digital quotations identified. Firstly, Table 4 reflects data for what Puschmann (2015) calls “primary output”, that is, the formal publication of scientific findings, which are hyperlinks to research articles and abstracts as well as links to academic book announcements:

Digital quotations	Hyperlinks to primary output			
	Research articles	Research article abstracts	Academic book announcements	Total
<i>Popular Science</i>	41/2.27	10/0.55	1/0.06	52 /2.88
<i>Smithsonian</i>	73/2.4	45/1.48	2/0.07	120 /3.94
<b>Total</b>	114/2.35	55/1.13	3/0.06	172/3.54

Table 4. Raw numbers and frequency of digital quotations of “primary output” sources per 1000 words.

Table 5 below contains data for “secondary output”, which here refers to scientific knowledge disseminated online through various channels (some of them already existing in the offline world, such as popularizations, and others, which are digitally-indigenous, like the blog).

Digital quotations	Hyperlinks to secondary output			
	Information in public and private websites	Personal blogs	Articles in the same magazine	Articles in different magazines
<i>Popular Science</i>	61/3.38	16/0.86	48 /2.66	20/1.11
<i>Smithsonian</i>	62/2.03	8/0.26	5/0.16	24/0.79
<b>Total</b>	123/2.53	24/0.5	53/1.09	44/0.9
Digital quotations	Reference materials (Wikipedia)	Media	Total	
<i>Popular Science</i>	14/0.76	0/0	159 /8.8	
<i>Smithsonian</i>	5/0.16	5/0.16	109 /3.58	
<b>Total</b>	10/0.39	5/0.1	268 /5.52	

Table 5. Raw numbers and frequency of digital quotations of “secondary output” sources per 1000 words.

As observed, hyperlinks to “primary output” are well represented (3.54), establishing intertextual networks with research articles and abstracts of the articles where the scientific knowledge reported was first published. However,

in terms of quantitative data, digital quotations linking to secondary output (popularizations, reports, media coverage) is higher (5.52), especially in the case of one of the sites (*Popular Science*), mainly due to the hyperlinks with public and private websites from where further information is derived and to links with feature articles in the same magazine, which might raise issues connected with self-promotion and visibility.

Hyperlinking to media (audiovisual) products is also found but only in one of the magazines (*Smithsonian*) and mainly in one single article (4 out of the total 5 cases recorded), showing that these affordances are not exploited to the extent they are in other digital texts like social media (i.e. Twitter, see Adami 2014; Zappavigna 2022; Luzón 2023):

The underlying theme of *The Bleeding Edge*, the 2018 documentary about the medical device industry, was that “innovative” doesn’t necessarily mean better care. Another 2018 film, *Upgrade*, warned audiences about “helpful” scientists offering state-of-the-art biotech devices.

[*Smithsonian* 12]

### c) Multimodal quotations

As indicated above, multimodal quotations refer to multimodal elements which are directly copy-pasted into the feature article. Table 6 below shows the quantitative data found for this type of quotation.

Multimodal quotations			
	Videos	Pictures	Media (TV)
<i>Popular Science</i>	1/0.56	5/0.28	0/0
<i>Smithsonian</i>	2/0.07	4/0.13	1/0.03
<b>Total</b>	3/0.06	9/0.19	1/0.02
	Graphs/ Figures	Visuals	Total
<i>Popular Science</i>	8/0.44	1/0.56	15 /0.83
<i>Smithsonian</i>	5/0.16	10/0.33	22/0.72
<b>Total</b>	13/0.27	11/0.27	37/0.76

Table 6. Raw numbers and frequency of types of multimodal quotations per 1000 words.

Although the frequencies per 1000 words are very small, some tendencies can be observed in the use of multimodal quotations. Data show a preference in online feature articles for pictures, graphs and visuals which, in all, amount to a total of 89.2% of non-verbal elements in the texts (33 out of 37). That is, these feature articles seem to favour the more static modes, also present in offline texts, whereas audiovisual modes (i.e. video or media) represent only 10.8% of the non-verbal elements. [It should also be noted that all the articles analyzed contained a picture at the top. These pictures were not added to the counting, as it was understood that they are a conventional element, part of the layout of the text, and their counting would have distorted the analysis of results.]

Although both publications share a preference for static nonverbal elements, as happens in other types of quotations and intertextual relations created, they show differences with respect to the nonverbal element included. Thus, *Popular Science* favours pictures and graphs/figures

whereas *Smithsonian* favours graphs/figures and other visuals (drawings or representations of objects, body parts, etc.).

A final quantitative analysis was carried out comparing the three types of quotation:

	Textual quotation	Digital quotation	Multimodal quotation
<i>Popular Science</i>	225 /12.46	225 / 12.46	15 /0.83
<i>Smithsonian</i>	460 /15.09	241 /7.9	22 /0.72
<b>Total</b>	685 / 14.12	466 /9.6	37 /0.76

Table 7. Raw numbers and frequency of the three types of quotations per 1000 words.

Data show that online feature articles primarily make use of offline quotation resources to establish intertextual networks, by means of textual (direct and indirect) quotations. Having said that, it is observed that they take advantage of the digital platform on which they are published and resort to technical affordances, incorporating digital quotations by means of hyperlinking to quite a substantial degree. Here, relevant differences are found between the two publications, with *Popular Science* making use of textual and digital quotations to exactly the same degree, and *Smithsonian* clearly opting for textual over digital quotations. These differences can probably be explained in terms of in-house conventions, although a larger corpus might cast a clearer light on this question. Multimodal quotations have a lower presence, as compared to textual and digital quotations, and as compared to the presence they have in other digital environments for science dissemination (i.e. social media).

In response to the first research question, the quantitative data reveal the existence of patterns of intertextuality in the corpus under study. These patterns show that digital scientific dissemination and, more specifically, the feature article published online on popularizing sites, is characterized by the use of three types of quotations: textual, digital and multimodal, which, as seen above, are present to various degrees in the corpus, with the latter being (still?) not very frequent.

The primary intertextual practices identified, textual and digital quotations, seem to serve different discursive purposes, concerning the level of expertise and specialization they bring into the text. Thus, textual (direct and indirect) quotations, which are intertextual resources also present in offline environments, tend to bring almost exclusively the voice of the expert scientist. They do so by creating intertextual networks which link the feature article to the primary scientific source (i.e. the voice of the expert as reflected in the research article or study published). In contrast, digital quotations, which exploit the technical affordances of the online platform (hyperlinking) open the scope for a wider variety of levels of (non) expertise, by linking not only with the research article or the research article abstract from which the information derives, but mainly with other popularizations or less specialized texts, occasionally introducing elements of popular culture such as clips from TV series and films. Multimodal quotations



in online feature articles are present to a much lower degree showing a stronger preference for “static” modes (i.e. pictures, graphs, figures and visuals) rather than for “animated” modes (i.e. audiovisuals).

Responding to the second research question, it can be claimed that, whereas the conventional, offline, intertextual resources mainly bring the voice of the authority into the text (thus serving the purpose of conferring authority to the voice of the journalist writing the feature article), digital quotations highly contribute to the recontextualization of the specialized knowledge for dissemination purposes by widening the range of voices incorporated into the text (i.e. institutions, public and private organizations, other specialized journalists, the public in general) thus offering a range of “multiple readings” with which the diversified audiences accessing the feature article may engage.

#### 4. Conclusion

In all, it can be claimed that intertextuality, as a recontextualizing tool, finds in digital communication a fertile ecosystem to expand and evolve, where both conventional (offline) intertextual resources and digitally native practices are exploited for the purposes of communicating expert knowledge to and engage with, probably not so specialized readers. To do so, the text needs to be conferred with the *auctoritas* of the expert voices from whom knowledge originates. The journalist acquires a “credible voice” by quoting the primary texts either textually or by means of hyperlinking. Moreover, the versatility of digital intertextual practices allows journalists to use them in order to contribute to another main function of popularizing texts: engaging with non-specialized audiences. This is done by various means which include, for instance, embedding digital quotations within text attributed to expert sources, or by fostering the “approachability” of the authoritative voice by hyperlinking with their personal blogs or pages. Digital affordances are also exploited to incorporate other layers of popularization which the offline mode is more resistant to or simply unable to integrate, such as hyperlinks to popular articles and multimodal texts (videos, etc.). Embedding multimodal quotations also fosters the engaging potential of online feature articles.

Compared to the digital intertextual practices found in social media (see Adami 2014; Puschmann 2015; Zapavigna 2022; Luzón 2023; Sancho 2023), online feature articles may still be on the “analogue” side of the spectrum of digital scientific popularization. The question is to see, as the living organism that it is, how is the genre of the online feature article going to evolve and whether, attending to the demands and expectations of an increasingly “digitalized audience” will correspondingly widen the range of intertextual resources incorporated.

#### 5. References

Adami, E. (2014). Retweeting, reposting, repinning; reshaping identities online: Towards a social semiotic multimodal analysis of digital remediation. *LEA - Lingue e Letterature d'Oriente e d'Occidente*, 3(3), 223–243.

Bazerman, C. (2004). Intertextuality : How texts rely on other texts. In C. Bazerman, & P. Prior (Eds.), *What Writing Does and How it Does it: An Introduction to Analyzing Texts and Textual Practices*. New Jersey, USA: Lawrence Erlbaum Associates, pp. 83-96.

Bernstein, B. (1996). *Pedagogy, Symbolic Control and Identity. Theory, Research, Critique*. London, UK: Taylor and Francis.

Blommaert, J. (2005). *Discourse: A Critical Introduction*. Cambridge: Cambridge University Press.

Bondi, M., Cacchiani, S., Mazzi, D. (2015). *Discourse in and through the Media: Recontextualizing and Reconceptualizing Expert Discourse*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Calsamiglia, H., Van Dijk, T. (2004). Popularization discourse and knowledge about the genome. *Discourse & Society*, 15(4), pp. 369--389. <https://doi.org/10.1177/0957926504043705>

Engberg, J. (2021). Prologue. In G. Pontrandolfo, & S. Piccioni (Eds.), *Comunicación especializada y divulgación en la red. Aproximaciones basadas en corpus*. London, UK: Routledge, pp. vii--x.

Fairclough, N. (1992). *Discourse and Social Change*. Cambridge, UK: Polity Press.

Farrelli, M. (2020). Rethinking intertextuality in CDA. *Critical Discourse Studies*, 17(4), pp. 359--376. <https://doi.org/10.1080/17405904.2019.1609538>

Giménez, J., Baldwin, M., Breen, P., Green, J., Roque Gutiérrez, E., Paterson, R., Pearson, J., Percy, M., Specht, D., Waddell, G. (2020). Reproduced, reinterpreted, lost: Trajectories of scientific knowledge across contexts. *Text&Talk* 40(3), pp. 293--324. <https://doi.org/10.1515/text-2020-2059>

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics* 17(4), pp. 433--454. <https://doi.org/10.1093/applin/17.4.433>

Johansson, M. (2019). Digital and written quotations in a news text: The hybrid genre of political opinion review. In P. Bou-Franch, & P. Garcés-Conejos Blitvich (Eds.), *Analyzing Digital Discourse. New Insights and Future Directions*. Switzerland: Springer, pp. 133--162.

Kristeva, J. (1986). Word, dialogue and novel. In Moi, T. (Ed). *The Kristeva Reader*. London, UK: Blackwell, pp. 35--61.

Li, Y. (2015). ‘Standing on the shoulders of giants’: Recontextualization in writing from sources. *Science and Engineering Ethics* 21, pp. 1297--1314. <https://doi.org/10.1007/s11948-014-9590-4>

Linell, P. (1998). Discourse across boundaries: On recontextualizations and the blending of voices in professional discourse. *Text* 24(3), pp. 143--157. <https://doi.org/10.1515/text.1.1998.18.2.143>

Luzón, M.J. (2013). Public communication of science in blogs: Recontextualizing scientific discourse for a diversified audience. *Written Communication* 30(4), 428 – 457. <https://doi.org/10.1177/0741088313493610>

Luzón, M.J. (2023). Forms and functions of intertextuality in academic tweets composed by research groups. *Journal of English for Academic Purposes* 64, 101254



- Myers, G. (2003). Discourse Studies of scientific popularization: questioning the boundaries. *Discourse Studies* 5(2), pp. 265--279. <https://doi.org/10.1177/1461445603005002006>
- Puschmann, C. (2015). A digital mob in the ivory tower? Context collapse in scholarly communication online. In M. Bondi, S. Cacchiani, & D. Mazzi (Eds.), *Discourse in and through the Media: Recontextualizing and Reconceptualizing Expert Discourse* Cambridge Scholars Publishing. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, pp. 22–45.
- Sancho-Ortiz, A. (under review). Analysing social media in EFL teaching and learning: Twitter for science dissemination. *Profile: Issues in Teachers' Professional Development*, 25(2).
- Zappavigna, M. (2022). Social media quotation practices and ambient affiliation: Weaponising ironic quotation for humorous ridicule in political discourse. *Journal of Pragmatics*, 191, 98–112.

# Studying the distribution of reply relations in Wikipedia talk pages

## Abstract

This paper presents an extended annotation and analysis of interpretative reply relations focusing on a comparison of reply relation types and targets between conflictual pages and neutral pages of German Wikipedia (WP) talk pages. We briefly present the different categories identified for interpretative reply relations to analyze the relationship between WP postings as well as linguistic cues for each category. We investigate referencing strategies of WP authors in discussion page postings, illustrated by means of reply relation types and targets taking into account the degree of disagreement displayed on a WP talk page. We provide richly annotated data that can be used for further analyses such as the identification of interactional relations on higher levels, or for training tasks in machine learning algorithms.

**Keywords:** Wikipedia talk pages, reply relations, referencing strategies

## 1. Introduction

This paper presents an extended annotation and analysis of reply relation types and targets in Wikipedia (WP) talk pages focusing on the investigation of reply relation (RR) types as well as target locations in Wikipedia talk pages containing different levels of disagreement.

Reply relations are a special kind of interactional relations that hold between postings, i.e. user contributions on a talk page. When Wikipedia authors communicate with each other on a talk page, a set of reply relations between postings on a talk page arises by the fact that the content of (one or more) previous posting(s) is directly addressed. Because computer-mediated communication (CMC) interactions vary based on genre and topic, reply relations are not limited to question-answer patterns. Reply relations involve any response or reaction that occurs when two authors communicate with one other.

We think that the annotation of reply relations is emerging as a promising method to reconstruct interaction structures on article discussion pages. By identifying reply relations, Wikipedia's usual convention of indentation can be substantiated or corrected if necessary, resulting in a more accurate representation of the underlying discussion structure.

### Cleanup

Does this complete the article cleanup? **Smack** (or others): you decide. -- [hike395](#) 02:16, 31 May 2005 (UTC)

It looks pretty good. [...] --[Smack](#) ([talk](#)) 19:24, 31 May 2005 (UTC)

Example 1<sup>1</sup>: Interpretative reply relation, type: addressing, linguistic cue - username “Smack”.

Indentations are a formal means to express reply relations in WP talk pages. The term *interpretative reply relations* refers to all reply action that is not realized formally but signaled by other structural or linguistic means.

In Example 1, the reply action is not realized formally by technical reply or indentation, but signaled by different linguistic means, e.g. via addressing, as in this case the use of *Smack* by author [hike395](#). We refer to such indicators as *linguistic cues*. The term *reply target* denotes a previous posting which is being referred to by the current posting. In Example 1, author Smack refers to [hike395](#)' posting which means that [hike395](#)'s posting is the reply target of Smack's answer. Besides different

forms of addressing, such as in Example 1, Q-A structures, or quotes can be considered as cues for interpretative reply relation types. For the presentation, we will summarize our taxonomy of interpretative reply relation types which abstracts overgroups of linguistic cues in talk page postings.

## 2. Wikipedia talk pages

We aim to provide enriched CMC data by annotating reply relations between postings on Wikipedia talk pages. There are different strategies when trying to reconstruct reply relations, such as focusing on the microstructure (i.e. internal structure) of postings by annotating speech acts, as in Ferschke et al. 2012. We take a closer look at the mesostructure, i.e., the relation *between* postings, and build upon Lungen/Herzberg (2019). By annotating reply relations between postings, we make explicit that a posting most of the time represents a reply to a previous posting and also to which posting exactly. In Wikipedia talk pages, the primary order structure is termed a *thread* (cf. Beißwenger et al. 2012). A thread contains a variable number of postings which the authors group thematically under different headings, such as “Cleanup” in Example 1. Wikipedia authors are requested to indent their contributions on the wiki page to build thread structures as known from other discussion forums. As a result, the amount of indentation is a property of the posting rather than something imposed by the server (cf. Beißwenger et al. 2012).

## 3. Linguistic Annotation

Identifying and annotating aspects such as the addressing cue in Example 1 is a first step to reconstructing the reply sequences in Wikipedia discussions. For a complete analysis, one would have to take into account the articles, the revision histories (of articles and talk pages), and the linked pages as well. The approach is a step towards the representation of interaction structures in CMC corpora, which will also allow for quantitative studies, similar to speech act annotations in speech corpora.

### 3.1 Research Questions

The annotation process addressed several goals. After demonstrating subtypes of reply relations and the reply strategies that occur within the extensive background of a Wikipedia talk page, we wanted to focus on the distribution of these reply types and strategies taking into account the degree of disagreement displayed on a WP talk page.

<sup>1</sup> [https://en.wikipedia.org/wiki/Talk:Hiking/Archive\\_1](https://en.wikipedia.org/wiki/Talk:Hiking/Archive_1).

Therefore, the research questions are as follows:

RQ1: Do conflictual pages and neutral pages differ in the distribution of reply targets?

RQ1a: Where in relation to the replying posting is the reply target posting located in conflictual vs. in neutral pages?

RQ1b: How frequently does the reply target annotated actually match the one indicated by the indentation, i.e. the “parent” posting exactly one indentation level higher? And if the annotated reply target is not the parent posting, is it then the immediately preceding posting? Do the respective figures in the conflictual pages differ significantly from the neutral pages?

RQ1c: How frequently is the reply target to be found in a different thread altogether in conflictual vs. in neutral pages?

RQ2: Do conflictual pages and neutral pages differ in the distribution of reply type categories?

RQ2a: Which reply relation type can be identified for each subcorpus?

RQ2b: Which reply relation type occurs most often in each subcorpus?

### 3.2 Data: Two Subcorpora of Wikipedia talk pages

Kittur et al. (2009) have shown in early Wikipedia research that articles of certain categories entail a high potential for disagreement and conflict (cf. Kittur et al. 2009). Categories in Wikipedia serve to group article pages according to certain characteristics. In addition to articles on religion and politics, article pages of the philosophy category and on personalities in particular contain an increased potential for conflict (cf. Kittur et al. 2009, p. 1512; Hara et al. 2010).

The Wikipedia Demo Corpus German Talk Pages Subcorpus<sup>2</sup> (WDC) provides the database of conflicting talk pages. Table 1 displays the WP pages (left column) of the WDC as well as the talk pages of the neutral corpus (right column). The neutral corpus consists of eight WP talk pages of less conflicting categories, such as technology, cities, animals, and represents the comparative, neutral data basis.

WDC; list of annotated conflict-prone WP pages	Neutral corpus; list of annotated “neutral” WP pages
Flüchtlingskrise in Europa ab 2015 ( <i>Refugee crisis in Europe from 2015</i> ) Chiropraktik ( <i>Chiropractics</i> ) Wladimir Wladimirowitsch Putin ( <i>Vladimir Putin</i> )	Berlin ( <i>Berlin</i> ) Streifenhörnchen ( <i>Chipmunk</i> ) Großer Panda ( <i>Giant panda</i> ) Fernglas ( <i>Binoculars</i> ) Stadtbahn Bonn

<sup>2</sup> The *Wikipedia Demo Corpus* was developed within the framework of a multilingual research project and is currently available via the Corpus platform *KorAP*: <https://korap.ids-mannheim.de/instance/wikidemo>.

Terroranschläge am 11. September 2001 ( <i>Terrorist attacks on 11 September 2001</i> ) Psychoanalyse ( <i>Psychoanalysis</i> ) Gentechnisch veränderter Organismus ( <i>Genetically modified organism</i> ) Feminismus ( <i>Feminism</i> ) The Legend of Zelda ( <i>The Legend of Zelda</i> )	( <i>Bonn city rail</i> ) Schwarzweißfotografie ( <i>Black and white photography</i> ) Grammatik ( <i>Grammar</i> ) Wandern ( <i>Hiking</i> )
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 1: Data basis: German Wikipedia talk pages, English translations added.

### 3.3 Annotation process and guidelines

The categories for annotating interpretative reply relations are based on suggested categories mentioned in Lungen/Herzberg (2019). These suggestions were transformed into a set of nine categories and annotated in three annotation rounds by two encoders<sup>3</sup> for the WDC data set, and in one annotation round for the neutral corpus data set.

For the annotations, simplified I5 versions of the pages devoid of all inline annotations (e.g. *italics*) were prepared, and the annotators used the simplified I5 XML files in the *Oxygen XML Editor* as well as annotation guidelines which explained the attributes and elements accordingly. Three attributes were annotated during the process: @relationTarget, @relationType, and @cueTarget, the latter in combination with the element <cue>.

To finish the annotation process, we adjudicated the annotations of both subcorpora (relation targets, types and cues) to create *master annotations* which constitute a gold standard dataset<sup>4</sup>. We took over the roles of adjudicators, as it is essential “to have adjudicators who were involved in creating the annotation guidelines, as they will have the best understanding of the purpose of the annotation” (Pustejovsky/Stubbs 2013, 134).

## 4. Results

The results section provides answers to the research questions using the WDC and neutral corpus master files. We extended the results of the WDC annotation process by comparisons to neutral WP sites in order to find out whether the observations made by Kittur et al. (2009) apply to referencing strategies as well.

<sup>3</sup> There were different encoders involved in the annotation processes: two encoders annotated the WDC data over a total of three consecutive annotation rounds. As the categories were solidified in the initial annotation process and the annotation guidelines existed in final form, two different encoders got by with fewer rounds of annotation when annotating the neutral corpus. Overall, the process also took less time, which can also be attributed to the different sizes of the subcorpora, with the WDC containing 572,968 tokens and the neutral corpus 21,131 tokens, as well as posting and thread sizes, cf. Table 2.

<sup>4</sup> Implementing this additional step is beneficial when planning to train and test machine learning (ML) algorithms.

We assumed that the more disagreement is displayed in the author's exchanges on a WP talk page, the more complex the referencing between the postings will get, creating long and branched discussion threads.

Before reporting on the results of the research questions, Table 2 presents some descriptive statistics about the sizes of pages, threads and posts in the two subcorpora.

	<b>Conflictual pages</b>	<b>Neutral pages</b>
avg #threads by page*	27.24	16.62
avg #posts by page*	278.27	47.50
avg #posts by thread*	10.20	2.86
avg #tokens by page	17,362.67	2,641.38
avg #tokens by post	67.80	54.53

Table 2: Sizes of pages, threads and posts in the two WP subcorpora. The asterisk \* symbolizes a significant size difference between the subcorpora<sup>5</sup>.

The two subcorpora differ significantly in the size of threads per page, posts per page as well as posts per thread. On the conflictual WDC talk pages there are more threads that contain a larger amount of postings which are longer as well in comparison to the talk pages of the neutral corpus. This confirms our assumption that the greater the amount of displayed disagreement in the author's exchanges on a WP talk page, the longer and more complex discussion threads are emerging.

<b>Reply relation target</b>	<b>% in conflictual pages</b>	<b>% in neutral pages</b>
Target is in the same thread	99.66	99.12
Target is in a different thread	0.34	0.88
Target is parent posting	66.30	76.79

<sup>5</sup> We used Pearson's chi-square statistic  $\chi^2$  to calculate differences in reply relation type distribution between the subcorpora, cf. <https://www.socscistatistics.com/tests/chisquare2/default2.aspx>. The p-values range between  $< .00001$  (avg #threads by page and avg #posts by thread), and .047415 (avg #posts by page). The results are significant at  $p < .05$ .

Target is parent and parent is preceding posting	93.79	96.51
Posting has more than one target	4.83	5.59

Table 3: Distribution of reply type targets in the two WP subcorpora.

Table 3 presents the distribution difference in percent of reply type targets in the two WP subcorpora. When investigating RR targets, we wanted to identify where in relation to the replying posting the reply target posting is located in conflictual in comparison to neutral pages. For both subcorpora, the clear majority of targets is located within the same thread: 99.66 % of annotated targets in the WDC and 99.12 % in the neutral corpus respectively (RQ1a). In 66.30 % of the annotated WDC data, the reply target annotated actually matched the one indicated by the indentation, i.e. the "parent" posting exactly one indentation level higher (RQ1b). In the neutral corpus, the amount is slightly higher with a total of 76.79 % that matched the "parent" posting exactly one indentation level higher. That means that the indentation had been used correctly (i.e. according to the Wikipedia guidelines). We then asked how frequently the reply target is, if not the parent posting, the immediately preceding posting. In 93.79 % of annotated targets in the WDC, the targets corresponded to the immediately preceding posting and were the parent posting at the same time. This result is almost identical in the neutral corpus, with 96.51 % annotated targets that corresponded to the immediately preceding posting and were the parent posting at the same time. Lastly, we analyzed whether the frequency of the reply target to be found in a different thread altogether differs between the subcorpora (RQ1c). Again, both subcorpora show a similar distribution. In around 5 % of postings, reply relations are identified to refer to more than just one other posting, i.e. where several interpretative reply relations that were identified within one posting show replies to more than one previous posting. The reply relations from postings like this can currently not be correctly identified by relying on the indentation only.

The respective figures in the conflictual pages do not differ significantly from the neutral pages. To conclude, the level of disagreement does not lead to a more branched and expanded discussion thread as the distribution of the annotated reply relation target locations does not differ between the subcorpora.

By contrast, the analyses of the RR types distribution revealed differences between the conflictual pages and neutral pages (RQ2). The annotators distinguished between eight reply relation types<sup>6</sup> (RQ2a), cf. the column "Reply relation type" in Table 4, while it was

<sup>6</sup> Additionally to the presented eight RR types in Table 4, the category "title-relation" was annotated as well. We do not include it here and in other presented results in this paper because it had been annotated largely automatically.

possible to identify more than just one type per posting.<sup>7</sup> The relation types arise from abstracting over the nature and forms of their linguistic indicators in the postings, cf. Example 1 in which the linguistic cue *Smack* used by author hike395 allows for interpreting the reply relation type “addressing”.

Reply relation type	% in conflictual pages	% in neutral pages
2ndPerson*	28.18	6.09
implied*	23.97	8.12
anaphor*	12.61	21.83
response-token*	12.10	25.89
quoting	9.02	7.11
addressing	8.90	6.60
QA-relation*	5.16	24.37
no relation annotated	0.07	0.00
<u>Sum</u>	<u>100</u>	<u>100</u>

Table 4<sup>8</sup>: Distribution of reply type categories in the two WP subcorpora, sorted by frequencies highest to low of the WDC. The asterisk \* symbolizes a significant difference in the RR type distribution between the subcorpora<sup>9</sup>.

As results of annotating RR types in the WDC conflict-prone WP talk pages, the two relation types occurring most often are “2ndPerson” and “implied” for both encoders<sup>10</sup> (RQ2b), cf. Table 4. These two types

<sup>7</sup> e.g., relationTarget="p1 p2" relationType="2ndPerson QA-relation" cueTarget="c2 c3" would encode that the posting includes two different types of reply relations, a “2ndPerson” as well as a “QA-relation”.

<sup>8</sup> We calculated the frequencies relatively in % to take into account the different subcorpora sizes.

<sup>9</sup> We used Pearson’s chi-square statistic  $\chi^2$  to calculate differences in reply relation type distribution between the subcorpora, cf. <https://www.socscistatistics.com/tests/chisquare2/default2.aspx>. The p-values range between  $< .00001$  (2ndPerson, response-token, QA-relation), .000015 (implied) and .001303 (anaphor). The results are significant at  $p < .05$ .

<sup>10</sup> We calculated an inter-rater agreement between the encoders for eight categories using Cohen’s  $\kappa$ . We counted all pairings of relation types at postings with one identical relation target, according to the following rules: label an empty relation type as the type ‘NO\_REL’, if there are one or more identical pairs of relation types, count the first one, and if there is no identical pair of relation types, count the first non-identical pair.

$\kappa$  for the conflictual pages, i.e. over the sum of all WDC postings, was 0.63;  $\kappa$  for the neutral pages was 0.63 as well. An agreement level between 0.61–0.80 classifies as a substantial agreement (Landis/Koch 1977, 165). This result shows that RR annotations in both subcorpora can be covered substantially well

count for over 50% of all reply types assigned by them. Putting this into perspective in terms of the relation type “implied”, we can see that for almost a quarter of all relations between postings in the WDC corpus, no specific textual cues could be identified, such as a greeting, “Hi Anna”, or a direct request to action, for example in questions like “Can you change...?”.

In the neutral subcorpus the two relation types occurring most often are “response-token” and “QA-relation” for both encoders (RQ2b). Comparable to the WDC RR category type results, also two RR types count for over 50% of all reply types assigned. However, in the neutral corpus, the RR type “anaphor” was identified with almost similar frequency, turning the dual lead into a trio.

Interestingly, the aforementioned three RR types “response-token”, “QA-relation” and “anaphor” can be identified significantly less often in the WDC. The relations between postings in the WDC arise rather implicitly by interpreting the contents of all participants’ postings involved and understanding their connection whereas on the neutral pages, referencing strategies between postings are signaled explicitly.

To conclude, the level of disagreement on a Wikipedia talk page impacts the distribution of RR types, but not RR targets. We could identify that on shorter and content-wise, more neutral talk pages, the distribution of RR types, namely the following five out of eight RR categories: “response-token”, “QA-relation”, “anaphor”, “2ndPerson” and “implied” differs significantly with regard to the conflictuality degree of a talk page. The category “implied” that contains relations established implicitly by the general content of the postings and the readers ability to infer and understand that a reply relation holds without reading a specific linguistic cue, finds proportionately large usage on the conflict-prone WDC talk pages.

Identifying interpretative reply relations helps account for postings whose references cannot be reconstructed via the indentation itself. By annotating @relationType and @relationTarget to identify the posting which another posting refers to, we know the extensiveness of discussion threads and the multipurposeness of one singular posting in which the author can address numerous issues simultaneously. Moreover, the developed reply relation type categories can be applied to a variety of talk pages, regardless of their potential for disagreement.

## 5. References

- Beißwenger, M./Ermakova, M./Geyken, A./Lemnitzer, L./Storrer, A. (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (Issue 3). <https://doi.org/10.4000/jtei.476>.
- Ferschke, O./Gurevych, I./Chebotar, Y. (2012): Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. Präsentiert auf: EACL 2012,

with the developed RR type categories, regardless of the level of disagreement on a talk page.



- Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics. S. 777–786. <https://www.aclweb.org/anthology/E12-1079>.
- Hara, Noriko/Shachaf, Pnina/Hew, Khe Foon (2010): Cross-Cultural Analysis of the Wikipedia Community. In: Journal of the American Society for Information Science and Technology 61(10), S. 2097–2108. <https://doi.org/10.1002/asi.21373>.
- Kittur, Aniket/Chi, Ed H./Suh, Bongwon (2009): What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure. Präsentiert auf: CHI '09: CHI Conference on Human Factors in Computing Systems, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Boston MA USA: ACM. S. 1509–1512. <https://doi.org/10.1145/1518701.1518930>.
- Landis, J. R./Koch, G. G. (1977): The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174. <https://doi.org/10.2307/2529310>.
- Lüngen, H./Herzberg, L. (2019): Types and annotation of reply relations in computer-mediated communication. In: European Journal of Applied Linguistics 7(2), S. 305–332. <https://doi.org/10.1515/eujal-2019-0006>.
- Pustejovsky, J./Stubbs, Amber (2013): Natural language annotation for machine learning. Sebastopol, CA: O'Reilly Media.
- WikiDemo Corpus in *KorAP* (*Corpus analysis platform*). <https://korap.ids-mannheim.de/instance/wikidemo>
- Wikipedia. <https://www.wikipedia.de/>

# MMWAH! Compiling a Corpus of Multilingual / Multimodal WhatsApp Discussions by Swedish-speaking Young Adults in Finland

**Martti Mäkinen**

Hanken School of Economics  
E-mail: [martti.makinen@hanken.fi](mailto:martti.makinen@hanken.fi)

## Abstract

This WiP paper will report the compilation of the corpus of *Multilingual / Multimodal WhatsApp discussions at Hanken* (MMWAH). The target data donors are Swedish-speaking young adults in Finland. This demographic group is of particular interest in the study of bi- and multilingualism and the coexistence of Swedish, English and Finnish in CMC. Furthermore, the corpus will lend itself to research on multimodal / polysemiotic practices in instant messaging and the maintenance of social networks and identity-building through the linguistic and other semiotic resources at the speakers' disposal.

**Keywords:** CMC, instant messaging, multilingualism, multimodality, linguistic practices in CMC, corpus compilation, Finnish-Swedish

## 1. Introduction

Digital interaction in social media has become one of the most important forms of informal written communication. Swedish-speaking young adults in Finland use several languages in parallel in their instant messaging (47 % of them use English or Finnish in addition to Swedish; Stenberg-Sirén, 2020), and that makes platforms such as WhatsApp important seats of language contact and linguistic innovations in Finland. The current project sets out to chart the language repertoires of the above-mentioned young adults, focusing on the co-existence of the (at least) three languages in WA discussions and how the participants in discussions create and maintain social networks through the use of the languages and other semiotic resources.

To study the phenomena described, a corpus of WhatsApp discussions, *Multilingual and Multimodal WhatsApp discussions at Hanken* (MMWAH, Hanken = Hanken School of Economics), is being compiled in the project. This Work-in-progress paper will report the process of the corpus compilation, with the criteria and choices made in collecting the data, touching upon the practical, legal, and marketing issues of the work.

## 2. About the project

Project Multilingual Instant Messaging: Focus on WhatsApp in Finnish-Swedish Digital Communication is a three-year project funded by the Swedish Cultural Foundation in Finland begun in August 2022. It is an associated project to Dynamics of Digitally-mediated

Language (DDL) that has brought researchers of Swedish, Finnish and English together to look into multilingual and polysemiotic practices in CMC in Finland.

The current project employs a small research team, PI and two research assistants, of which one is a student of language technology and the other a software engineer. The main function of the project is to compile a corpus for studying the interplay of English and the domestic languages in Finland, but also for the research purposes of DDL.

## 3. Criteria and motivation for data collection

The target group of the data collection is Swedish-speaking young adults, 18-30 years of age, in Finland. The reasons to study this group in particular are many. As native speakers of the official minority language in Finland (5.2 % of Finnish citizens are speakers of Swedish (Statistics Finland, 2022)), many of them have either active or at least passive knowledge of Finnish (the domestic languages of Finland, Finnish and Swedish, are both compulsory subjects at school, irrespective of one's mother tongue). Among the target group, English is ubiquitous, and hence the members of the group communicate more or less comfortably in three languages, one of which is not an Indo-European language. Of course, there are individuals in the group whose third language beyond Swedish and English is not Finnish; nevertheless, the project sees this group as an interesting manifestation of multilingual practices, and the data collected will provide possibilities for multiple lines of research.

WhatsApp has been chosen as the instant messaging platform of interest, due to its wide use in Finland. According to AudienceProject (2020), WA is the most used IM platform in Finland, with 87 % of the population having some experience in its use. There would be other platforms that provide IM functionality and that would be, perhaps, more preferred by young adults (e.g. Discord, Snapchat etc.). However, WhatsApp is geared towards the ease of textual communication (as opposed to the original idea of platforms that aim for the ease of video distribution), and that is the motivating factor for the current project.

The corpus will be, to some extent, similar to the ones compiled in projects MoCoDa2 (Beißwenger, Imo, Fladrich & Ziegler, 2019) WhatsUp, Switzerland? (Überwasser, 2021; Überwasser & Stark, 2017) and the Dutch WhatsApp corpus (Verheijen & Stoop, 2016). Nevertheless, after the beginning of these projects, GDPR has changed the game of collecting personal communication to some extent (cf. section 4).

### 3.1 Relevant research questions

The corpus will lend itself for several kinds of studies. The most immediate research questions for the current project will be about code switches between Swedish / Finnish and English, and also the triggering events of such switches (cf.



Peterson, Biri & Vaattovaara, 2022). Other questions of interest are the co-existence of the three expected languages, the dynamic interplay between the domestic languages, and identity-building of the user groups through the available linguistic and other semiotic repertoires (cf. Peterson, Hiltunen & Vaattovaara, 2022).

Such questions on mediated interaction practices will combine ethnographic and quantitative research methods. Themes such as linguistic routines, multi-lingual, multi-modal and stylistic repertoires, and how participants of discussions develop and adhere to interactive norms can be scrutinized with the help of the MMWAH corpus (cf. Leppänen *et al.*, 2009; Yus, 2014; Seufert *et al.*, 2016).

#### 4. GDPR and other complications on corpus compilation

Since the implementation of the General Data Protection Regulation (GDPR) of the European Union in 2018, also the collection of language data must comply with the new regulations. It protects the privacy of participants in research so that the collected data originating with natural persons will not allow direct or indirect identification of the persons in question (GDPR, Regulation 2016/679). In practice, storage limitations set by GDPR demand that collected language data is stored on servers within the EU throughout the project's lifespan and after. The lawfulness of data processing requires the informed consent of the data subjects themselves, for which reason an informed consent is collected from all research participants.

Additionally, according to GDPR (Article 5), data should be minimised, meaning that collection of superfluous data should be avoided. Thus, all processed and stored data must have an obvious purpose for the research. The processes and processors of data have been described in the project data management plan, also available to data donors (cf. MMWAH).

Finally, according to WhatsApp rules, the discussion data must be collected from the platform users as donations, in a similar manner as in the afore-mentioned projects dealing with WA data. The marketing of the project will be shortly described in section 5.1. The multimodal aspect of the project means that we are also asking for any media shared in discussions. That unfortunately decreases the number of messages in a discussion to a maximum of 10,000 messages. In our experience, the number will be even lower than that, depending on the max. allowed size of email attachments. Nevertheless, we have decided to use the email as the mode of donations, as we wish to make the process as approachable to the potential donors as possible.

#### 5. Compilation apparatus

As said, WA discussion data can only be collected through donations. The system (still under development and testing) directs the donations to a dedicated project mail inbox. A pre-processor script scrapes the exported discussion and

media, saves it on the project server and sends a link to the donor email address that lets them to edit the discussion, mostly to delete sensitive messages and media.

With the same message we send a link to a web survey tool that collects demographic information about the donor (year of birth, gender, linguistic background = languages known, education) and also contact information to the other participants in the discussion. The last item is something that has been discussed a lot. It has transpired that not all HEIs (or their lawyers) interpret GDPR alike, and to be certain that the letter of GDPR is obeyed, we will ask for an informed consent from all the participants of discussions.

An automated script will delete the originally donated material when a) the donor has completed the edition phase, or when b) the time window for editing has passed (cf. MoCoDa2, Beißwenger *et al.*, 2019). This will ensure that no direct nor strong indirect identifiers are entered into the corpus and that sensitive, unused data is removed from our servers.

#### 5.1 Marketing MMWAH project

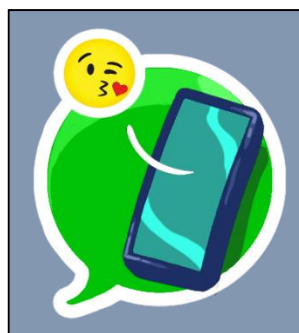


Figure 1: MMWAH project logo

As the data must be collected as donations from the target group, the project has planned a marketing campaign for obtaining the data. The plan includes F2F project pitches to relevant audiences (student bodies, sports teams, amateur theatres, SIGs), online video material (both advertisements and practical tutorials), airtime with national and local Swedish-speaking radio channels, and of course some attention to the visual image and logo of the project.

At the time of writing this paper, the F2F pitches have been begun; the online and on-the-air elements of the campaign will wait until the automated data collection apparatus is fully implemented. This is projected to take place in the second year of the project.

#### 6. Corpus scope and flavours

The projected size of MMWAH is 300,000–500,000 words which will make it a comparable resource of Finnish-Swedish in comparison with other corpora in the same language, and almost unparalleled among unmoderated language corpora of Finnish-Swedish. Of course, the size depends directly on the interest of potential donors, thus also the figure is subject to change.

WA discussions for MMWAH will be tagged for languages used, POS, lemmas, and demographic parametres, avoiding the disclosure of indirect identifiers of the data donors. The container file type will be XML, and the basic encoding

will be carried out according to TEI-CMC (Beißwenger & Lungen, 2020; Chanier *et al.* 2016).

Also other “flavours” of the corpus will be published, especially for the purpose of depositing the corpus in various corpus repositories. E.g., potential future repositories would be “language banks” in Finland and in Sweden (Kielipankki and Språkbanken respectively), and for that the corpus should be compatible with KORP, their corpus search engine, with POS-tagged vertical files.

## 7. Corpus afterlife in OA repositories

The publication of MMWAH will take place through two main channels, Kielipankki (Kielipankki) in Finland and Språkbanken (Språkbanken) in Sweden, in Open Access format and under a Creative Commons license. Also, the project will follow FAIR principles, to ensure the accessibility of the collected data (cf. Frey *et al.*, 2019).

## 8. References

- AudienceProject. (2020). AudienceProject Insights 2020. App and Social Media Usage. Available: [https://www.audienceproject.com/wp-content/uploads/AudienceProject\\_Study\\_App\\_and\\_Social\\_Media\\_Usage\\_2020\\_pdf.pdf?x62193](https://www.audienceproject.com/wp-content/uploads/AudienceProject_Study_App_and_Social_Media_Usage_2020_pdf.pdf?x62193).
- Beißwenger, M., Imo, W., Fladrich, M. and Ziegler, E. (2019). <https://www.mocoda2.de>: a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions. *European Journal of Applied Linguistics*, 7(2), pp. 333–344.
- Beißwenger, M., Lungen, H., (2020). Cmc-core: A basic schema for encoding CMC corpora in TEI. *Corpus* [online], 20. Available: <http://journals.openedition.org/corpus/4553>.
- Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham C. R., Hriba, L., Longhi, J. and Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, 29(2), pp.1–30.
- DDL = Dynamics of Digitally-mediated Language. [online] Available: <https://sites.utu.fi/digitallymediatedlanguage/>, accessed 04/2023.
- Frey, JC., König, A., Stemle E. W. (2019). How FAIR are CMC Corpora? In Julien Longhi, Claudia Marinica (Eds.) *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)* Cergy-Pontoise, France: The Institute of Digital Humanities of Cergy-Pontoise University, pp. 26–31.
- GDPR = Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at: <https://gdpr-info.eu/>, accessed 04/2023.
- Kielipankki = *Language Bank of Finland* [online]. <https://www.kielipankki.fi/language-bank/>, accessed 04/2023.
- Leppänen, S., Pitkänen-Huhta, A., Piirainen-Marsh, A., Nikula, T., Peuronen, S. (2009). Young People's Translocal New Media Uses: A Multiperspective Analysis of Language Choice and Heteroglossia, *Journal of Computer-Mediated Communication*, 14(4), pp. 1080–1107.
- MMWAH = Multilingual Instant Messaging: Focus on WhatsApp in Finnish-Swedish Digital Communication, project homepage. [online] Available at: <https://www.hanken.fi/sv/institutioner-center/hankens-center-sprak-och-affarskommunikation/forskning/snabbmeddelanden-pa>, accessed 04/2023.
- Peterson, E., Hiltunen, T., Vaattovaara, J. 2022. A place for plis in Finnish: A discourse-pragmatic variation account of position. – Elizabeth Peterson, Turo Hiltunen & Joseph Kern (eds.), *Discourse-Pragmatic Variation and Change: Theory, Innovations, Contact*, pp. 272–292. Cambridge University Press. DOI: 10.1017/9781108864183.015.
- Peterson, E., Biri, Y., Vaattovaara, J. 2022. Grammatical and social structures of English-sourced swear words in Finnish discourse. – Martín-Solano, R. & San Segundo, R. (eds.), *Corpus linguistics and Anglicisms*, pp. 49–70. Peter Lang Publishing. DOI: 10.3726/b19222.
- Seufert, M., Hoßfeld, T., Schwind, A., Burger V. and Tran-Gia, P. (2016). Group-based communication in WhatsApp. 2016 IFIP Networking Conference (IFIP Networking) and Workshops, Vienna, Austria, pp. 536–541.
- Språkbanken = *The National Language Bank* [online]. Available at: <https://sprakbanken.se/sprakbankeninenglish.html>, accessed 04/2023.
- Statistics Finland, 2022. Finland's national statistical institute. [online] Available at: [https://www.stat.fi/index\\_en.html](https://www.stat.fi/index_en.html), accessed 04/2023.
- Stenberg-Sirén, J. (2020). *Svenska, finska, engelska – komplement eller alternativ? En inblick i ungas språkanvändning*. Helsinki: Magma. Available: [http://magma.fi/wp-content/uploads/2020/01/magma1\\_2020\\_webb-1.pdf](http://magma.fi/wp-content/uploads/2020/01/magma1_2020_webb-1.pdf).
- Ueberwasser, S. (2021). Look Back Without Anger: Recapitulation of the Corpus *What's up, Switzerland?*. In Iris Hendrickx, Lieke Verheijen, and Lidwien van de Wijngaert (eds.) *Proceedings of the 8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)*. Nijmegen, NL: Radboud University, pp. 98–100.
- Ueberwasser, S. und Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5), pp. 105–126.
- Verheijen, L., Stoop, W. (2016). Collecting Facebook Posts and WhatsApp Chats. In P. Sojka, A. Horák, I. Kopeček & L. Pala, (Eds.) *Text, Speech, and Dialogue. TSD 2016*. vol 9924, pp. 249–258.

Yus, F. (2014). Contextual constraints and non-propositional effects in WhatsApp communication. *Journal of Pragmatics*, 114, pp. 66—86.

# CoDEC-M: The multi-lingual manosphere subcorpus of the Corpus of Digital Extremism and Conspiracies

Rachel McCullough, Daniel Drylie, Mindi Barta, Daniel Smith

Bolante.NET, Old Dominion University, Oregon State University, Bolante.NET

E-mail: rechalmccullough@protonmail.com, ddrylie@odu.edu, bartam@oregonstate.edu, danielvsmithpsyd@gmail.com

## Abstract

The Corpus of Digital Extremism and Conspiracies (CoDEC) is an open-source, open-access corpus made up of several subcorpora documenting different online spaces where extremists and conspiracy theorists gather. CoDEC-M is a subcorpus that addresses a growing interest in the manosphere and a gap in knowledge on the language used in non-English speaking parts of the manosphere. By comparing the most frequent keywords and bigrams in the Russian and English sections of CoDEC-M, we may better understand language contact and transfer between the English-speaking manosphere and its non-English speaking equivalents, as well as shed light on shared themes and linguistic innovations in non-English speaking communities that self-identify as part of the manosphere.

**Keywords:** corpus linguistics, multilingual corpora, language and gender, computer mediated discourse, Russian language

## 1. Introduction

Media coverage of different groups belonging to the manosphere, particularly incels and red pillers, has accompanied the widespread social isolation and rise in right wing rhetoric of the last decade. The language of these communities has also received more mainstream attention. While academic research has been conducted on the manosphere, both linguistic research on non-English manosphere populations and cross-cultural research on these groups is fairly limited. This pilot study using CoDEC-M is designed to remedy this by comparing the top 20 keywords and bigrams in both the English and Russian sections of CoDEC-M by relative frequency compared to reference corpora in each target language.

Academics have documented incel communities and their speech, but previous corpus analysis of CMC data focusing on the manosphere has targeted primarily English data (e.g. Thomas, 2022; Bogetic, 2022). While a semantic analysis of Russian incels was conducted in 2021 (Voroshilova & Pesterev), the authors were unable to find many texts from self-identified incels (p. 186), and their results deviated from the explicit misogyny and violent rhetoric that studies of English-speaking incels would lead one to expect, suggesting a need for further investigation.

## 2. Methods

CoDEC-M consists of data in two languages taken from two sources. For this study, we compared datasets from two languages: 4.4 million tokens in Russian and 2.1 million tokens in English. The disparity in size of these two sections is due to space limitations in the text analysis software Sketch Engine (SkE). The English section of CoDEC-M was scraped using Selenium on the entirety of incels.is, and the Russian section was scraped using Beautiful Soup with manually retrieved links for 472 threads from 2ch.hk's ongoing /incel/ thread on the /sex/ board. Once compiled, the corpora were preprocessed and run through the Wordlist and Keywords tools in SkE to determine the top keywords and bigrams via relative frequency (RF) when compared to the English and Russian TenTen corpora.

## 3. Results & Implications

The top 20 keywords and bigrams in both sections of CoDEC-M are presented in Table 1 on the next page. Russian data is presented in both Russian and English.

Profanity is characteristic of the discourse of both groups (see Table 1). The top three keywords by relative frequency parallel, with the Rank 1 word by RF referring to women (“foid”; “тян”), Rank 2 being the word for incel (“incel”; “инцел”), and Rank 3 referring to non-incels (“normie”; “чед”). Notably, the Russian keywords contain borrowings, suggesting language transfer from English.

Discussion of sex and dating is frequent in both communities, but the English language data contains bigrams referring to race (Table 1), while more Russian bigrams reference physical characteristics—presumably those of incels themselves (e.g., пониже рост, *shorter height*). The English data also contains self-referential terms like “incel forum” and “incel community” which are conspicuously absent from the Russian data, possibly due to the lack of a dedicated forum for Russian incels.

This comparison illustrates that within self-identified incel communities there is a shared focus on gender roles and physical appearance. While there is a partially shared lexis, further research is needed to understand the specialized Russian vocabulary with no equivalent in the English data, like тян (*chan*) and скиф (*skuf*).

## 4. References

- Bogetic, K. (2022). Race and the language of incels: Figurative neologisms in an emerging English cryptolect. *English Today*, pp. 1-11.
- Kapiec. (2021). In *Slovar molodozhnogo slenga*. Retrieved 3 Aug. 2023, from <https://slang.su/id/23910>.
- Russian incelosphere. (2023). In *Incels Wiki*. Retrieved 1 Aug., 2023, from <https://archive.ph/bsqg3>.
- Thomas, M. (2022). A quantitative analysis of the language used by violent and non-violent incels. [Master's thesis, University of North Carolina at Chapel Hill]. *Carolina Digital Repository*.
- Voroshilova, A. & Pesterev, D.O. (2021). Russian Incels Web Community: Thematic and Semantic Analysis. *Communication Strategies in Digital Society Seminar (ComSDS)*, pp. 185-190.

#	CoDEC-M (EN)	CoDEC-M (RU)	CoDEC-M (EN)	CoDEC-M (RU)
1	foid	тян <sup>1</sup> ( <i>chan</i> ; a young woman)	white foid	невольное воздержание ( <i>involuntary abstinence</i> )
2	incel	инцел <sup>2</sup> ( <i>incel*</i> )	incel forum	осознанное воздержание ( <i>conscious abstinence</i> )
3	normie	чед <sup>3</sup> ( <i>chad*</i> )	foid worship	пониже рост ( <i>shorter height</i> )
4	jfl	ебать <sup>4</sup> ( <i>fuck</i> )	white woman	сын шлюхи ( <i>son of a whore</i> )
5	cuck <sup>5</sup>	спок ( <i>good night</i> )	black pill	желание секса ( <i>desire for sex</i> )
6	blackpill	хуй <sup>6</sup> ( <i>dick</i> )	average look <sup>7</sup>	высочайший рост ( <i>highest growth</i> )
7	subhuman	пиздец ( <i>fucked up</i> )	average height	тёмная триада ( <i>dark triad</i> )
8	nigger	блять ( <i>fuck</i> )	virtue signal	линия роста ( <i>growth line</i> )
9	fakecel	бетабакс ( <i>betabucks*</i> )	low inhib	ростом волос ( <i>hair growth</i> )
10	mog	всратый ( <i>shitted-in</i> )	white knight	фаза знакомства ( <i>dating phase</i> )
11	inceldom	пизда ( <i>pussy, cunt</i> )	white guy	наибольший хуй ( <i>biggest dick</i> )
12	manlet	нормис ( <i>normies*</i> )	short man	процентом жира ( <i>fat percentage</i> )
13	tbh	ирл ( <i>irl*</i> )	white girl	размер хуя ( <i>dick size</i> )
14	chad	опухший ( <i>swollen, puffy</i> ; referring to a swollen face due to alcohol consumption)	high inhib	черта лица ( <i>facial features</i> )
15	truecel	двачую ( <i>I agree; me, too</i> )	dating app	линия волос ( <i>hairline</i> )
16	simp	скуф ( <i>skuf</i> ; an unappealing man, age 30-55) ( <i>Incels Wiki, 2023</i> )	asian woman	основная теория ( <i>basic theory</i> )
17	faggot	кунов ( <i>kun</i> ; a young man)	clown world	красивейший парень ( <i>handsome guy</i> )
18	brocel	лвл ( <i>lvl*</i> , level)	low tier	зона глаз ( <i>eye area</i> )
19	trucel	подкатывать ( <i>pull up</i> ; approach a girl to get to know her) ( <i>Slovar molodozhnogo slenga, 2021</i> )	tier normie <sup>8</sup>	красивейший человек ( <i>handsome man</i> )
20	oneitis	анон ( <i>anon*</i> )	incel community	глазе жертвы ( <i>prey eyes</i> )

Table 1: Top 20 Keywords and Bigrams in the English and Russian sections of CoDEC-M. Bigrams are regularized as part of pre-processing, and duplicate word stems have been excluded from the top 20 keywords. Asterisks in the translations indicate that the translated word is a borrowing from English.

<sup>1</sup> Duplicate word stems: тянки (RF of 407.89), тянок (RF of 303.70), тянка (RF of 170.63)

<sup>2</sup> Duplicate word stems: инцелы (RF of 430.16), инцелов (RF of 289.64), инцела (RF of 162.51)

<sup>3</sup> Duplicate word stems: чеды (RF of 157.89), чедов (RF of 157.32), чеда (RF of 187.90); Spelling variations: “чэд” (RF of 155.81).

<sup>4</sup> Duplicate word stems: ебанный (RF of 160.53) ебалю (RF of 274.44), ебанный (RF of 160.53).

<sup>5</sup> Duplicate word stems: “cucked” (RF of 238.5).

<sup>6</sup> Duplicate word stems: “нахуй” (RF of 585.03), “похуй” (“fuck off,” RF of 264.35), “хуйня” (RF of 309.65), нихуя (RF of 212.53), дохуя (RF of 117.50)

<sup>7</sup> Because lemmatization was performed as a preprocessing step, the bigram “average-looking” is rendered “average look.”

<sup>8</sup> Hyphens constitute word breaks in SkE, so the bigram “tier normie” surfaces with a higher RF than phrases like “low-tier normie”.

# Little Big Data: Karelian Twitter Corpus

Ilia Moshnikov<sup>1\*</sup>, Eugenia Rykova<sup>1,2\*</sup>

<sup>1</sup>University of Eastern Finland

<sup>2</sup>Technical University of Applied Science TH Wildau

E-mail: [ilia.moshnikov@uef.fi](mailto:ilia.moshnikov@uef.fi), [eugeniia.rykova@th-wildau.de](mailto:eugeniia.rykova@th-wildau.de)

\*The two authors have contributed equally to the present paper.

## Abstract

This paper investigates Karelian language visibility on Twitter and describes the first corresponding data collection using language-related keywords and hashtags. In total, 2626 entries written fully or partially in Livvi, South and Viena Karelian were scraped with Postman API. The visibility of Karelian on Twitter has been considerably increasing in the past few years, Livvi-Karelian being the most prominent dialect. The data were analysed linguistically (manually and with language detection software) and thematically. Although language-related topics are the most popular, there is a substantial number of entries in eight further topics. Applicability of the collected data for linguistic and sociological research, and further data collection considerations are discussed.

**Keywords:** Karelian, minority language recognition, Twitter

## 1. Introduction

Twitter is a leading microblog platform, which can serve for data collection on various research questions (Grillenberger, 2021). Languages other than English have been receiving more attention in the last decade. However, the studies including minority languages are still scarce (see Cunliffe, 2019; Valijärvi and Khan, 2023). Thus, Karelian language is not even separately discussed in Twitter linguistic repertoire of Finland (Hiippala et al., 2020). Lack of corresponding automatic language processing tools hinders the process, too. In this paper, an approach to investigate Karelian language visibility on Twitter and collect the corresponding data is described, and considerations for further data collection are discussed. The following questions are addressed: How to collect data in Karelian from Twitter? How present is Karelian on Twitter throughout the years? What dialects of Karelian are the most visible on Twitter? What are the topics of tweets published in Karelian?

## 2. Research background

### 2.1 The Karelian language and its usage online

Karelian is a minority, critically endangered Finnic language mainly spoken in Russia and in Finland. Currently, the total number of Karelian speakers is roughly about 20,000 people (Sarhimaa, 2017; Federal State Statistics Service, 2021). Linguistically, the Karelian language is divided into two main dialects: Olonets (or Livvi) Karelian, and Proper Karelian. The latter consists of Viena (North) Karelian and South Karelian (Koivisto, 2018). Several written standards of Karelian make the use of the language online diverse.

The first signs of using Karelian online are from the late 1990s. The first websites in Karelian were launched in the early 2000s. From the 2010s, the use of Karelian on social media started significantly growing. Salonen (2017)

studied language use in internet services and software, and the visibility of the language on social and digital media. Moshnikov (2022a, 2022b) studied the use of Karelian as a language of websites from the virtual linguistic landscape and language ideologies theory as well as the use of the language online by Karelian speakers. According to the research, while Facebook is the most popular social media for consuming and creating content in Karelian, the use of Karelian on other social media platforms, including Twitter and Instagram, has increased (Moshnikov, 2022a). As a new domain, the language use on social media demonstrates ongoing trends and changes in language itself, and also reflects certain sociocultural processes. Language use in different domains and its responsiveness to new domains and media are the keystones in language survival and vitality (Drude and Intangible Cultural Heritage Unit's Ad Hoc Expert Group, 2003).

### 2.2 Twitter

Twitter is a social media micro-blogging platform, where users can publish short messages, or tweets, of a maximum of 280 characters (140 characters until November 2017) and receive feedback from other users (Fausto and Aventurier, 2016). As social media interaction in general, Twitter is a multilingual source of data, corresponding to Big Data's definition: it has volume, velocity, and variety (Kitchin, 2013). Unlike Facebook, Twitter has long allowed to collect data via Twitter API. In February 2023, Twitter announced elimination of the free API access, which would make further data collections more difficult (Willingham, 2023). However, it is not clear yet, which changes exactly academics will have to face as the corresponding information has not been updated since March 30, 2023 (Twitter Developers, 2023).

Local communities adapt social media platforms, including Twitter, for their purposes and interests using specific hashtags. Speakers of a particular language create their own hashtag systems, which makes it easier to find tweets based on a concrete topic, place, or language (Cocq, 2015;

Outakoski et al., 2018; McMonagle, 2019). Communities of speakers also create networks to support and encourage language use and learning. In small communities, the role of an individual active user could be crucial.

### 3. Research data and methods

#### 3.1 Data scraping

Postman API software (2022) was selected to collect data from Twitter using Academic Research access due to convenient way of modifying the search parameters and stating the necessary information sections to be retrieved (cf. Rykova et al., 2023). Unlike other language-specific Twitter data collections (e.g., AbdelHamid, 2022; Rykova et al., 2023), Karelian data cannot be collected via specifying the language in the query as Karelian is neither among those whose identification is supported by Twitter API, nor built-in in other software libraries for Twitter data collection. Post-hoc language identification of Karelian (cf. Ljubešić et al., 2014; Nguyen et al., 2015) is also difficult due to scarcity of corresponding resources and dialect variability. Thus, the applicability of HeLI-OTS 1.4 language identifier (Jauhiainen et al., 2022) to Twitter entries is first researched in the current paper.

First, a full-archive search was performed with the help of the following keywords and hashtags: *karjalan kieli*, *karjalan kielet*, *karjalan kielen*, *karjalan kielien*, *karjalan kieltä*, *karjalan kieliiä*, *karjalakse*, *karjalaksi*, *#karjalakse*, *#karjalaksi*, *#karjalankieli* (case and special characters can be ignored). It was assumed that the users would use these hashtags to highlight the use of the language as the speakers of other minority languages do (Cocq, 2015; McMonagle et al., 2019), and keep Karelian apart from the Finnish language. Hashtag *#karjala* was not included in the search because it is often used in irrelevant posts about beer or geographically related questions, discussed in Finland. The forms of the nominative, genitive, and partitive have been chosen according to their frequency and usage in the closely related Finnish language (Hakulinen et al., 2004, §1228).

Since Karelian cannot be detected via Twitter API, but is recognized as Finnish, the search query contained the parameter of Finnish as the language of the entry. Additionally, tweets had to be organic, not an advertisement. The data were retrieved starting from 2007. After the initial search, additional searches for the parent tweets of the retrieved comments (multiple tweets query) and user information (user lookup query) were performed. Thus, the data included entries, information on public metrics, location (if given), and the author. For the comments, the entries that allowed to trace the conversation back (references) to the parent tweet were included.

#### 3.2 Data reduction

As of March 14, 2023, the collected data consisted of 15,428 entries. Removing retweets allowed to reduce the number of entries to 8463. Removing duplicates – tweets with the same content from the same or different users, which were not marked as retweet, and could be copying

the same links or self-repetitions – resulted in the final number of 8224 entries.

#### 3.3 Data labelling

The text of entries was subject to automatic language detection with the help of HeLI-OTS 1.4 (Jauhiainen et al., 2022). This language identifier includes two dialects of Karelian: Livvi-Karelian (*olo*) and Ludic Karelian (*lud*), although nowadays Ludic is generally considered an independent language (Pahomov, 2017). Besides identified language itself, information on the algorithm confidence score and the second probable language was also saved.

The language was also labelled manually by the first author of the study, who is a native Livvi-Karelian speaker. Manual labels included more specific information on Karelian dialects: in column ‘language’ generally marked as *olo* or *krl*, and the latter further specified in a separate column ‘dialect’ (South or Viena Karelian). If an entry contained several sentences written in up to five different languages, the languages were listed in order of appearance. Non-text entries, the ones with languages mixed within a sentence, or separate sentences written in more than five languages were labelled as “other”.

Manual labelling also included assigning topics to the entries written fully or partially in one of the Karelian dialects. The selection of topic was data-driven: relevant groups were identified and refined during the labelling process.

### 4. Results

#### 4.1 General results

In total, there are 2626 entries in the final dataset that are fully or partially written in one of the Karelian dialects – 31.9% of the cleaned-up data. Figure 1 shows the distribution of these entries by years and dialects. Year 2023 is not included as it is not complete yet. If an entry contains more than one dialect, it is counted for each of them. Thus, the total number of entries in the graph is higher than the actual number of entries in the database. In general, 88% of the one-language entries were in Livvi-Karelian, 8% in South Karelian, and 4% in Viena Karelian.

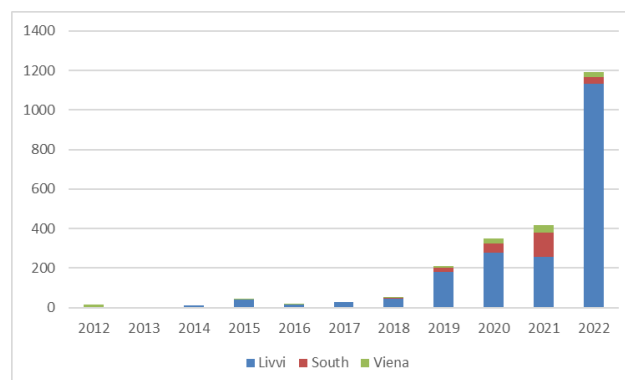


Figure 1: Entries in Karelian per year



## 4.2 Language detection

The comparison of automatically identified and manually labelled languages can be seen in Figure 2. It must be noted that this matrix is not a classic confusion matrix as its true and predicted labels are asymmetric. Manual labelling allows including more than one language (e.g., *krl* + *eng/fin* means (not Livvi) Karelian followed by either English or Finnish, 3+ (*olo*) means three and more languages, including Livvi-Karelian), while automatic detection outputs only one. Besides that, automatic detection has (erroneously) output languages that are not present in the original data.

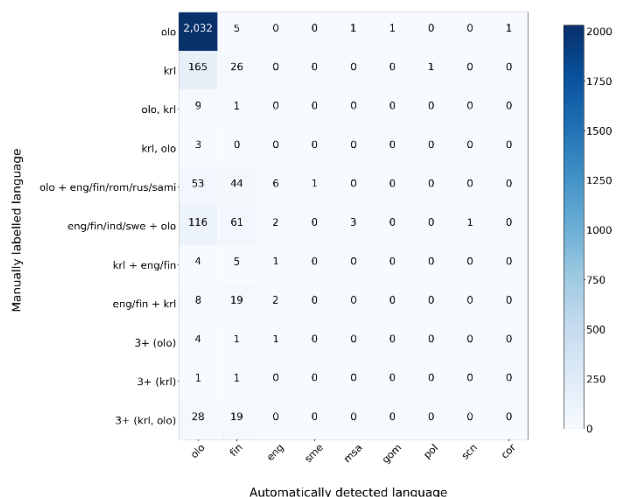


Figure 2: Confusion matrix of manually and automatically identified languages

ISO 639-3 language codes used in Figure 2 are (in order of appearance): *olo* – Livvi-Karelian, *krl* – South or Viena Karelian, *eng* – English, *fin* – Finnish, *rom* – Romani, *rus* – Russian, *ind* – Indonesian, *swe* – Swedish, *sme* – Northern Sami, *msa* – Malay (macrolanguage), *gom* – Goan Konkani, *pol* – Polish, *scn* – Sicilian, *cor* – Cornish. The language marked as Indonesian (*ind*) has such label based on the location of the entries author. However, this variety of Malay is not included separately in the languages detected by HeLI-OTS 1.4, which makes Malay macrolanguage (*msa*) the closest possible label for the absent *ind*. Sami languages are manually marked as a group (*sami*), without further distinction.

If an entry is written in South Karelian, its language is detected as Livvi-Karelian in 90% of the cases and as Finnish – in 9.3% of the cases. From entries written in Viena Karelian, detected language is Livvi-Karelian in 72.1% of the cases, and Finnish otherwise (27.9%).

Mean confidence score of the algorithm in cases, when Livvi-Karelian (*olo*) is recognized as such, is 1.58 (SD (standard deviation) = 0.43). Mean confidence score of the algorithm in cases, when other dialects (*krl*) are recognized as Livvi-Karelian (*olo*), is 0.94 (SD = 0.6), while when they are recognized as Finnish, the mean confidence score is 0.3 (SD = 0.24). In 80% of such cases, Livvi-Karelian is the second language in the language probabilities list.

In cases with two and three or more languages, when the

recognized language is one of the entry languages (or *olo* for *krl*), mean confidence score is 0.54 (SD = 0.44 and SD = 0.42, respectively). In cases, when a language absent in the original data (*gol*, *pol*, *scn*, *cor*) is detected, mean confidence score is 0.25 (SD = 0.24).

## 4.3 Topics

Ten topics identified during the manual labelling procedure are presented in Table 1, including the corresponding number of entries and an example. It must be noted that 98% of the topic ‘religion’ comprises the extracts from the Bible or other (Christian) religious texts, posted by the same user (Jyrki Kuusirati), starting in February 2021. Most of the entries in the topic ‘media’ are links to news sources, accompanied by the title (and sometimes subtitle) of the corresponding news article.

## 5. Discussion

The method of using language-related keywords and hashtags has proven to be successful to collect Twitter entries in Karelian. From the data available from Twitter until March 2023, 2626 original entries in three Karelian dialects were collected. The predominant dialect is Livvi-Karelian, which is in line with other research (Moshnikov, 2022a). Despite the use of such specific hashtags, language-related topics were not the only ones identified in the data. The research data show that certain events and holidays increase the activity of users. For example, the launch of Yle News in Karelian (Yle Uudizet karjalakse) in 2015 or the establishment of the Association of Young Karelians in Finland (Karjalazet Nuoret Suomes, KNŠ) in November 2019 clearly increased activity in Karelian on Twitter. Some spikes in the use of Karelian on Twitter can also be observed on specific dates, for example, Karelian Language Day (November 27<sup>th</sup>).

Automatic detection of language with the help of HeLI-OTS 1.4 (Jauhiainen et al., 2022) allows to identify Livvi-Karelian with 99.6% sensitivity. Two other Karelian dialects, namely South Karelian and Viena Karelian, are identified as Livvi-Karelian with 90% and 72.1% sensitivity, respectively, although the mean confidence of the algorithm becomes lower. Entries with separate phrases or sentences in different languages are usually identified with one of the languages present in the entry. However, it is difficult to predict, which phrase/sentence would influence the identification. In the future, such entries should be further split. Mean confidence score is relatively low in cases when identified language is unrelated to the actual language of the text. Thus, the proposed language detection algorithm can be used for Twitter (or other social media) data scraping in Karelian. Mean confidence scores and second possible language could decrease the possibility of missing relevant data and provide information of the possible dialect.

Certain authors (both individuals and organisations), regularly writing in Karelian on Twitter, have been identified. Further data collections could focus on these particular authors and their interactions with other Twitter users (cf. Ljubešić et al., 2014; Nguyen et al., 2015).

Topic	Description	N of entries	Example
Religion	Tweets related to the religious holidays or Bible.	1455+28	<i>Hyviä äijänpäivän pruzazniekkua! Kristoz voskres! Hristos nouzi kuollielois! #äijänpäivy #äijypäivy #karjalakse</i>  'Happy Easter holidays! Christ has resurrected! Christ has risen from the dead!'
Personal	Tweets identified as an opinion or experience.	327	<i>Tänäpäinä lähen otuskah, ga loma vuottau dačalla! Hyviä heinäkuudu #karjalakse #tiedäjättijetäh</i>  'I'm going on holiday today, but the iron scrap is waiting for me at the cottage! Have a good July'
Vocabulary	Tweets related to the learning of the language from the perspective of the vocabulary (e.g., translations and presenting variants from the different dialects of Karelian).	313	<i>Tänäpäi aijankohtaine sanaine karjalakse on huračču = vasenkätinen. Huraččuloin päiviä pietäh 13. elokuudu jo vuvves 1976. #sanainekarjalakse</i>  'A relevant word in Karelian today is 'huračču' - left-handed. Left Handers Day has been celebrated on 13 August since 1976.'
Language status and policy	Tweets related to the language status, policy, and revitalization process in a broad understanding.	217	<i>Karjalan kieli on oma kieli, ei suomen kielen murreh.</i>  'Karelian is a proper language, not a dialect of Finnish.'
Media	Tweets of news or other mass-media sources.	106	<i>Yle Uudizet karjalakse: Päivännouzu-Suomen yliopisto tahtou jatkaa karjalan kielen elvytändiä da kehitändiä</i>  'Yle News in Karelian: The University of Eastern Finland would like to continue its work on the revitalisation and development of the Karelian language.'
Research	Research related topics.	54	<i>Hyvä karjalan kielen maltai! Vastua kyzelyh karjalan kielen käyttöh näh! #karjalakse #karjalankieli</i>  'Dear Karelian speaker! Answer the questionnaire on the use of the Karelian language!'
Language learning	Education related topics including university studies and other language courses.	44	<i>Zavodimma egläin Karjalan Liiton karjalan kursan. Mie opastan varzinkarjalua / suvikaarjalua. Opastujat ollah kaikin puolin Suomie, 20 rištikanzuo. Keski-igä ozapuulieh 27,5 vuotta.</i>  'Yesterday, together with the Karelian Union, we started a Karelian language course. I am teaching Karelian Proper/South Karelian. The students are from all over Finland, 20 people. Average age of the participants is 27.5 years old.'
Culture	Culture related topics.	41	<i>Elbyygö karjalan kieli? Tulgua terveh Lieksan 11. kul'ttuuraseminuarah piätinččänä 4. muarienkuuda. Väl'l'ä piäzy!</i>  'Will the Karelian language be revived? Welcome to the 11th Lieksa Cultural Seminar on Friday, 4 March. Free entry!'
Politics	Tweets related to elections or political parties.	12	<i>Minule mugon! Iänestä minuu Jovensuun kunduvalličuksis 2021!</i>  'For me it's like this. Vote for me in the Joensuu Municipal Election 2021!'
Other	Other topics not related to other groups mentioned here.	29	<i>Hyviä puolistusvoimien flagupruazniekkua! #karjalan #kieli #puolistusvoimat #flagu #pruazniekku #Suomi</i>  'Happy Flag Day of the Finnish Defence Forces!'

Table 1: Topics identified in the collected data.

During the manual labelling of the entries, 10 topics were identified in the data. However, the topic ‘religion’, comprising the largest number of entries, mainly consists of direct citations of the (Christian) religious texts. While it is relevant to the general visibility of Karelian online, these collected texts cannot be considered as representing personal voices on social media. The same holds true for the majority of entries in the topic ‘media’ because they only copy the titles and subtitles of the news articles and provide corresponding links. The topic ‘vocabulary’ predominantly contains word or phrase lists translated into one or more of Karelian dialects. While such posts are also relevant for visibility and could be used as language learning resource, their applicability to other research fields is questionable. That reduces our dataset to 752 Twitter entries, written fully or partially in one or more Karelian dialects, that could be used for deeper linguistic and sociological analysis. The data can be analysed in the context of language or dialect contacts, lexical and morphological variation, and from the perspective of the translation studies and discourse analysis. Tweets and discussions related to the status of the Karelian language are interesting from the perspective of language revitalization and policy. The modern use of Karelian online has also an important symbolic meaning for the Karelian-speaking community. The collected corpus becomes even more important in the current context of changes in Twitter API access.

## 6. Conclusion

To the best of authors’ knowledge, this paper describes the first corpus of Twitter entries in Karelian. Using language-related keywords and hashtags, 2626 original entries corresponding to 10 different topics were collected. 29% of the material is found useful for further linguistic and sociological analysis. Future data collection is discussed.

## 7. References

- AbdelHamid, M., Jafar, A. and Rahal, Y. (2022). Levantine hate speech detection in twitter. *Social Network Analysis and Mining*, 12, 121.
- Cocq, C. (2015). Indigenous voices on the web: Folksonomies and endangered languages. *Journal of American Folklore*, 128(509), pp. 273--285.
- Cunliffe, D. (2019). Minority languages and social media. In G. Hogan-Brun & B. O’Rourke (Eds.), *The Palgrave Handbook of Minority Languages and Communities*. London: Palgrave Macmillan, pp. 451--480.
- Drude, S. and Intangible Cultural Heritage Unit’s Ad Hoc Expert Group. (2003). Language vitality and endangerment. (11 July, 2023).
- Fausto, S. and Aventurier, P. (2016). Scientific literature on Twitter as a subject research: Findings based on bibliometric analysis. In C. Levallois, M. Marchand, T. Mata, & A. Panisson (Eds.), *Twitter for Research Handbook 2015--2016*, pp. 1--14.
- Federal State Statistics Service. (2021). Vserossiiskaja perepis’ naselenija 2020 [Russian Census 2020]. [https://rosstat.gov.ru/vpn\\_popul](https://rosstat.gov.ru/vpn_popul). (19 April, 2023).
- Grillenberger, A. (2021). Twitterdaten analysieren mithilfe der blockbasierten Programmiersprache SNAP! [Analyse Twitter data using the block-based programming language SNAP!]. *LOG IN*, 41, pp. 54--60.
- Hakulinen, A., Vilkkumäki M., Korhonen R., Koivisto V., Heinonen T.R. and Alho I. (2004). *Iso suomen kielioppi [Descriptive Grammar of Finnish]*. Online version. Helsinki: Finnish Literature Society. <http://scripta.kotus.fi/visk>. (10 July, 2023).
- Hiippala, T., Väisänen, T., Toivonen, T. and Järvi, O. (2020). Mapping the languages of Twitter in Finland: Richness and diversity in space and time. *Neuphilologische Mitteilungen*, 121(1), pp. 12--44.
- Jauhainen, T., Jauhainen, H. and Lindén, K. (2022). HeLI-OTS, Off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3912--3922. Marseille, France. European Language Resources Association.
- Kitchin R. (2013). Big data and human geography: Opportunities, challenges, and risks. *Dialogues in Human Geography*, 3(3), pp. 262--267.
- Koivisto, V. (2018). Border Karelian dialects – a diffuse variety of Karelian. In M. Palander, H. Riionheimo, & V. Koivisto (Eds.), *On the border of language and dialect*. Studia Fennica Linguistica 21. Helsinki: Finnish Literature Society, pp. 56--84.
- Ljubešić, N., Fišer, D. and Erjavec, T. (2014). TweetCat: a tool for building Twitter corpora of smaller languages. *Proceedings of LREC 2014*, pp. 2279--2283.
- McMonagle, S., Cunliffe, D., Jongbloed-Faber, L. and Jarvis, P. (2019). What can hashtags tell us about minority languages on Twitter? A comparison of #cymraeg, #frysk, and #gaeilge. *Journal of Multilingual and Multicultural Development*, 40(1), pp. 32--49.
- Moshnikov, I. (2022a). The use of the Karelian language online: Current trends and challenges. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(2), pp. 275--305.
- Moshnikov, I. (2022b). The use of the Karelian language online: websites in Karelian. In T. Seppälä, S. Lesonen, P. Iikkanen, & S. D’hondt (Eds.), *Kieli, muutokset, yhteiskunta - Language, change, society*. AFinLA Yearbook 2022, pp. 192--216.
- Nguyen, D., Trieschnigg, D. and Cornips L. (2015). Audience and the use of minority languages on Twitter. *Proceedings of the ninth international AAAI conference on web and social media*, 9 (1), pp. 666--669.
- Outakoski, H., Cocq, C. and Steggo, P. (2018). Strengthening indigenous languages in the digital age: social media-supported learning in Sápmi. *Media International Australia*, 169(1), pp. 21--31.
- Pahomov, M. (2017). *Lyydiläiskysymys: Kansa vai heimo, kieli vai murre? [The Ludian Question: Nation or tribe, language or dialect?]*. Helsinki: Helsingin yliopisto & Lyydiläinen Seura.
- Postman. (2023). Postman API Tool. (14 March, 2023).
- Rykova, E., Stieben, C., Dostovalova, O. and Wicker, H. (2023). Connected Driving in German-Speaking Social Media. *Social Sciences*, 12(1): 46.

- Salonen, T. (2017). Karelian – a digital language? In C. Soria, I. Russo, & V. Quochi (Eds.), *Reports on digital language diversity in Europe*. (21 April, 2023).
- Sarhimaa, A. (2017). *Vaietut ja vaiennetut. Karjalankieliset karjalaiset Suomessa [Silent and being forced to be silent: Karelian-speaking Karelians in Finland]*. Tietolipas 256. Helsinki: SKS.
- Twitter Developers. (2023, March 30). Announcing new access tiers for the Twitter API. *Twitter*. <https://twittercommunity.com/t/announcing-new-access-tiers-for-the-twitter-api/188728> (11 July, 2023).
- Valijärvi, R.L., and Kahn, L. (2023). The Role of new media in minority- and endangered-language communities. In E. Derhemi and C. Moseley (Eds.), *Endangered Languages in the 21st Century*. Abingdon, Oxon, England: Routledge, pp. 139--157.
- Willingham, A.J. (2023, February 3). Why Twitter users are upset about the platform's latest change. *CNN*. <https://edition.cnn.com/2023/02/03/tech/twitter-api-what-is-pricing-change-ccc/index.html>. (11 July, 2023).

# Multimodal Intertextual Practices in Video Film Reviews

Anastasiia Piroh

Universität Duisburg-Essen

E-mail: anastasiia.piroh@stud.uni-due.de

## Abstract

A paradigmatic shift in the functions and formats of film reviews has altered the ways that intertextuality, or the relations of the text to other texts, is rendered in various film review formats. As one of the fastest growing multimodal social networking sites, *YouTube* offers a nuanced inventory for expressions of intertextuality, which extend beyond exclusively textual practices. This study is part of a larger project which investigates the semiotic resources of evaluation in online video film reviews. Based on a detailed case study of a sample from the Corpus of Online Video Film Reviews (CoVFR), this paper aims to explore the diversity and complexity of multimodal intertextual practices in video film reviews.

**Keywords:** video film review, intertextuality, semiotic resources, interdiscursivity, multimodality

## 1. Introduction

Almost every spoken or written utterance or text is a product of the texts that came before. Intertextuality and interdiscursivity bear special significance in the context of digital genres, particularly in online reviews whose primary function is the evaluation of products and services (Vásquez, 2014). This paper aims to analyse intertextuality and interdiscursivity from a multimodal perspective, focusing on the semiotic resources that construe the meaning of online video film reviews. Section 2 defines the genre and main characteristics of online film reviews. Section 3 elucidates the theoretical foundations of intertextuality and interdiscursivity. Section 4 introduces the data and methods utilised in the study. In Section 5 I provide a concise summary of the findings of this case study. Finally, section 6 presents some concluding remarks.

## 2. Video film reviews as a genre

The genre of film reviews has a long history, starting at the dawn of cinematography. Traditionally associated with high-end journalism and professional film criticism, film reviews underwent substantial changes when they expanded to online media. The online movie review is typically written by a non-professional user, with the intention of sharing information and evaluation with an audience of peers (Taboada, 2011). A paradigmatic shift can therefore be observed, from one-to-many lecture in traditional media to a many-to-many open dialogue and exchange between a reviewer and their audience, as well as members of the audience. Robert Koehler (as cited by Battaglia, 2010) points out that the Web is helping make film criticism more accessible, but also more difficult to define. The increased accessibility of online film reviews broaches the question of functional shifts within the genre. According to McWhirter (2016), prior to the networked digital media age, the institution of film criticism operated on a functional continuum between aesthetically-based judgement and a substantive social function. Digitalisation, and consequent democratisation, of the reviews foregrounded evaluation as their primary purpose. Vásquez (2014) considers reviews hybrid texts that may also include description and narration. The interrelated change of formal patterns and functions, as well as the rising popularity of film reviews among a

larger audience emerged with the first video film reviews, at that time broadcasted on television. The American TV-show *At the Movies*, hosted in the 1970s by Gene Siskel and Roger Ebert, brought exposure to the genre, placing the appeal and personality of a film critic at the forefront (Battaglia, 2010). The goal to appeal to a broader audience, e.g. by using witticisms, alliteration and catchphrases, democratised the genre, a trend which we can also observe in modern online video film reviews.

In contrast to the traditional review aggregator websites, such as *Rotten Tomatoes*, *Metacritic* and *IMDB*, *YouTube* grants its content creators full creative freedom and a broad amount of resources to address specific topics and issues. As opposed to written CMC texts, *YouTube* reviews can draw on multiple semiotic resources, such as moving images, spoken word, captions, photos, clickable icons and links (Benson, 2015). Video reviews on this site, despite their accessibility and primary goal of entertainment, have proven to be a complex and nuanced subgenre of online film reviews. In contrast to their written counterparts, video film reviews can rely on the aforementioned semiotic resources to construe the meaning of the text, as will be elaborated on in the following sections of the paper.

## 3. Intertextuality and interdiscursivity as integral characteristics of video film reviews

While intertextuality is conceptualised as ‘making reference to other texts’ (Vásquez, 2015) and a ‘text internal phenomenon’ (Bhatia, 2010), Bloor & Bloor (2007) relate interdiscursivity to ‘genre-mixing’, or the hybridisation of the one genre of text-type with the other’. Both intertextuality and interdiscursivity play an essential role in the meaning-making process. Meaning does not have to be confined to a single text but is derived from an interdiscursive, intertextual web of social and historical practices (Benwell & Stokoe, 2006). Therefore intertextuality can be conceptualised as a text-creating practice, as well as a discourse property of texts (Vásquez, 2015), while interdiscursivity refers to the mixing of various genres, discourses, and cultures (Bhatia, 2010). Although interdiscursivity has not been as widely researched as intertextuality, Vásquez (2015) and Lam (2013) confirm that genre interaction, which often includes borrowing and sharing semiotic resources, has proven to have pervaded in the context of digital

communication.

Intertextuality and interdiscursivity are evident in the majority of online discourses. In the case of film reviews, at least two discourses, the reviewer's discourse and film discourse, interact, regardless of their format. While written film reviews utilise a more traditional range of textual practices, e.g. paraphrasing and recontextualisation, in order to relay the intertextual references, video film reviews are in a position to leverage visual and aural modes to fulfil the same goal. For instance, instead of only narrating interviews with third parties that are relevant to the reviewed film, a video film reviewer can edit in a corresponding clip, or alternatively supplement it with a voice-over, therefore re-conceptualising intertextuality in this review format.. The following sections illustrate multimodal intertextual and interdiscursive practices employed by film reviewers on *YouTube*.

#### 4. Data and methods

The data sampled for this case study comes from a larger project which aims to explore the interplay of semiotic resources of evaluation in video film reviews, taking a micro-diachronic perspective. Due to the complexity of multimodal data, I have annotated a portion of a specialised corpus of OVFR including the 42 most viewed video film reviews from various *YouTube* channels that have been active since 2012. The content creators targeted in this case study are considered independent, i.e. not hired or commissioned to produce their film reviews.

My analysis focuses primarily on the video material and the process of inductive identification of diverse types of intertextuality and interdiscursivity in the reviews (coded with the assistance of the video annotation software, MAXQDA 2022). While I will also provide some preliminary quantitative results, the approach of this study is essentially exploratory and qualitative. Due to space limitations, the common patterns of intertextuality and interdiscursivity will be illustrated by examples from a selected OVFR *Half in the Bag: Ghostbusters: Afterlife (SPOILERS)*<sup>1</sup> which contained the highest frequency of intertextual references (83 instances). The video is 58 minutes in length and had approximately 2,5 mil. views at the time of data collection (March 2023).

The Youtube video this study draws on is publicly available online. Yet, notions of publicness and privateness are not only tied to availability, but also to expectations of internet users as to how their data are normally used. Therefore studies involving social media data of this kind also need to consider ethical questions. According to the guidelines established by the Association of Internet Researchers, there is an important distinction between 'participants ... best understood as "subjects"', and 'authors whose texts/artefacts are intended as public' (cited by Vásquez, 2014). The content chosen for analysis falls under the second category of sources. Furthermore, I take into consideration that the public nature of the reviews uploaded to YouTube also refers to the authors'

intention to share their views and opinions with the audience. In order to comply with ethical research protocols, only the first names of video participants will be mentioned in this paper for the sake of clarity. Seeing that my study focuses primarily on the video content and properties of the reviews, as opposed to the personal information about their authors, my work conforms to the general guidelines and does not require any special permissions.

#### 5. Results

In her monograph, Vásquez (2014) concludes that in contrast to previous studies (see Pollach, 2006), online reviews do not remain isolated from other texts, but demonstrate richly varied forms of intertextuality.

My case study provides further evidence of this, as the selected video film reviews showcase a wide array of multimodal intertextual practices. Table 1 contains the most frequently used forms of intertextuality in the annotated portion of the corpus. I identified: cultural references (that include references to popular and Internet culture phenomena), references to third parties (often mentioning the personalities directly related to the production of the reviewed film and reactions to other film critics) and references to the reviewer's own body of work.

Form of intertextuality	Number of instances	Percentage
Cultural references	270	77.8%
References to third parties	49	13.1%
Self-references	21	6.0%
Interdiscursivity	11	2.9%

Table 1: Frequencies of intertextual practices in OVFR

Drawing on the frameworks developed by Ungood Hughes & Riley (2012) and Janney (2012), I have summarised the most prominent semiotic resources of intertextuality observed in the analysed corpus in Table 2.

Modes	Submodes	Examples
Verbal and paralinguistic	Polylogue, vocal and prosodic elements	Discussion between the participants, intonation
Kinesic	Gestures and facial expressions	Shrugging, pointing towards the interlocutors
Editing	Edited-in clips, visual effects	Clips of movie scenes and interviews, filters
Graphic design	Picture and text	Overlaid photos and screenshots, captions
Sound construction	Sound effects, musical scoring	Sound effects (cash register, reverb), score ( <i>Ghostbusters</i> theme,

<sup>1</sup> See <https://youtu.be/5vzSPROcXIU> (accessed on March 6th, 2023).

		piano), voice-over
Staging	Sets and props	Black backdrop, props relevant to the reviewed film

Table 2: Semiotic resources of intertextuality

Due to the multitude of relevant materials and space limitations, I necessarily have to take a selective approach and will therefore provide a detailed description of examples representative of particular intertextual practices, clustered into the three major types of references established in the data.

### 5.1 Cultural references

References to well-known cultural or popular cultural events constitute the most frequent form of intertextuality in the data set, as nearly 78% per cent of the OVFRs included this type of intertextual practices. Most often, this form of intertextuality encompasses references to other films as means to compare the merits of film-making, however references to various forms of entertainment (e.g. comic books, video games, musicals, etc.) have also occurred throughout the corpus.

The selected exemplary review of *Ghostbusters: Afterlife* (2021, dir. Jason Reitman) showcases that aside from embedding the discourse of the reviewed film, into the discourse of their OVFR, the authors of *RedLetterMedia* allude to or directly mention a number of cultural phenomena in order to draw comparisons or corroborate their opinions.

Figure 1 illustrates a complex instance of intertextual layering in this video film review. The reviewed film is introduced with a help of a montage of scenes from the film itself and a voice-over, narrating the basic information about it. However, the tone and pacing of the introduction change when the reviewer refers to the film as: “the latest entry in our continuing collective pop culture equivalent of *The Chris Farley Show*”, thus opening the overarching theme of the film’s similarities with the original *Ghostbusters* (1984, dir. Ivan Reitman) and the general nostalgia exploitation caused by the resurgence of the 1980s popular culture in the media. Mimicking Chris Farley’s satirical interview style, the reviewer continues to list the most recognisable scenes from the original film:

(1) Remember Twinkies? Remember the ‘Stay Puff’ marshmallow man? Remember the Nestle’s crunch bar that Bill Murray handed Harold Ramis in the throwaway improvised gag? ... Remember literally the entire third act of the first movie? I love the 80s!

While the reviewer asks these questions, a scene from the 1984 *Ghostbusters* film is played, which is interjected by an edited-in meme clip of a pointing Leonardo DiCaprio from the 2019 film *Once Upon a Time in Hollywood* (dir. Quentin Tarantino).



Figure 1: Film scenes interjected by a reference to meme culture

In the online context, this reaction image and/or clip is often employed to communicate excited recognition. The reviewers rely on it to deliver the point that the 2021 film relies on the sentimental value of the original *Ghostbusters* as its major selling point.

Commercialisation of the franchise is the strongest point made in this review, its iterations appear throughout the video’s runtime. Other cultural references include the mentions of well-known American brands (e.g. Walmart, Baskin Robbins, Funko Pop etc.), which were also present in the film. The reviewers draw attention to the fact that paid advertisement takes away from the film’s potential value, accompanying each mention of the brands with a cash register sound effect.

Remarkably, brand references and the review’s main theme of commoditization are expressed not only through the mode of editing, but also through staging.



Figure 2: Dan Aykroyd’s Crystal Skull Vodka and a Ghostbusters figure used as props

Figure 2 depicts a skull-shaped bottle of vodka by a brand founded by the 1984 *Ghostbusters* actor, Dan Aykroyd and a *Ghostbusters* figure as props on the set. While the latter can be interpreted as another allusion to the reviewers’ conviction that the film’s aspiration to achieve commercial success overshadows its artistic value, the former reference is quite obscure and requires familiarity with popular culture. Additionally, the vodka bottle prop has already made its appearance in previous reviews of *RedLetterMedia*, which layers a reference to the reviewers’ work.

### 5.2 References to third parties

Sometimes, individuals refer to third parties, such as film directors, actors, film crew members, film audience, other reviewers, *YouTube* comments, etc. The sample film review inserts clips from Bill Murray’s interview about *Ghostbusters: Afterlife*, as well as images and videos of other cast members of the original *Ghostbusters* films, particularly Harold Ramis to whom the reviewed film is dedicated. There is also a clip from an interview with the



film's director. The quoted texts are rooted in a context, however, juxtaposed to the reviewers' predominantly negative evaluation, we can observe the stark contrast between the director's intended vision and the unfavourable response to the film.

The authors of the *Ghostbusters: Afterlife* review also refer to some texts written by movie-goers who shared their positive evaluation of the film. Such references occur in the form of recontextualised (possibly also simulated) quotations (2), as well as in the form of overlaid screenshots, e.g. from Twitter (Fig. 3).

(2) I bet if you look up, like, reactions to this movie... People are saying, "I got so emotional at the end, I-I cried when I saw the PKE metre". I can guarantee, yeah, that's the reaction. I don't understand how people can think like that.

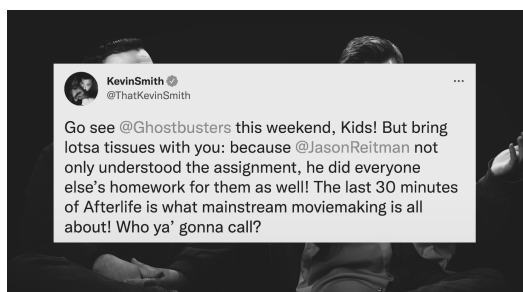


Figure 3: Reference to other review text

In example 2, *RedLetterMedia* reviewers express their direct disagreement with the positive reviews of the film (*I don't understand how people can think like that.*). Figure 3, however, exhibits a more implicit form of disagreement, expressed through the modes of web design and sound construction. The sequence consists of the screenshots of other online reviews, with an emotional piano score accompanying the screenshots' dissolve, while one of the reviewers provides a facetiously dramatic narration of the texts. While the reviewers make no evaluative claims about the third-party statements presented, the audience is capable of discerning *RedLetterMedia*'s veiled opposition to the quoted texts.

### 5.3 Self-references

Despite a lower number of occurrences in the annotated corpus, self-references represent a notable form of intertextuality. Commonly, references to one's own body of work take place when film franchises are the subject of multiple OVFRs, hence serving as the way for the reviewer to restate their previous evaluations.

For instance, the sample OVFR opens with the footage of the channel's previous video, *Rich and Jay Talk About Ghostbusters: Afterlife*<sup>2</sup>, in which the reviewers converse about the film's trailer and make predictions for the upcoming *Ghostbusters* sequel. The clip from this video is inserted into the review, and has the caption 'December

2019' at the bottom of the frame. In order to further indicate that the clip is situated in the past, a diffused glow filter is applied to the footage, with reverb audio effects accompanying visual cues. Later in the review, one of the channel members, Jay, reiterates his initial prediction for *Ghostbusters: Afterlife* that had already been brought up in their 2019 video:

(3) I think...we talked about the trailer. It's like, that's not fan service, like, that's the tech. Like that's just what these movies are.

Example 3 shows that the reviewers substantiate their evaluative claims of the film's technical aspects by reiterating their initial reaction to the film's trailer and the information that was inferred from its contents.

Aside from their evaluative purpose, self-references can serve as means to entertain and engage with the audience. For instance, in order to address the fact that this video is filmed on a set different from the one where *Half in the Bag* reviews usually take place, one of the reviewers comments:

(4) We're in the black void. Last time the three of us were here together was after watching *Star Wars: Episode 10*.

Furthermore, he makes an additional intertextual reference to the prior review the channel had posted.

The following excerpt illustrates the references to the fictional characters and storylines created by the channel's authors for the sake of framing their film review series *Half in the Bag*.

(5) But what about the ongoing Mr. Plinkett storyline? Will you ever get the antibodies from Mr. Plinkett's blood?

More specifically, example 5 illustrates a reference to Mr. Plinkett, a character that originated on *RedLetterMedia* and made recurring appearances in the channel's extensive body of work. Frequently employed, recurring intertextual elements, such as references to the channel's most recognisable characters, demonstrate the intertextual link between the sample review and a larger discourse system that connects it with the review series, and by extension, with the channel's overall lore. Video film reviews can thus transform and develop the structure and purpose of film reviews in the traditional sense of the genre, by not only utilising a wider array of technical tools, but also by constructing a larger, overarching discourse system.

### 5.4 Interdiscursive practices in OVFR

OVFRs feature several instances of interdiscursivity, in which various genre conventions are borrowed and blended into the film review. The examples discussed here draw on the discourse of blockbuster films, as the reviewers criticise the weak narrative in *Ghostbusters: Afterlife* and suggest alternatives to the film's plot progression.

<sup>2</sup> See <https://youtu.be/5vzSPROcX1U> (accessed on March 6th, 2023).

(6) ...after *Ghostbusters 2* what do you do?.. You franchise it?.. and then it's Bill Murray on the phone, you know, "Hey listen, our Cleveland... chapter of the Ghostbusters is having problems with their-, with their equipment, you know. Can we get cheaper people to fix it?" ... he's doing his Bill Murray thing right? And then he cuts the budget to save money and that's when things go bad... and that's when the doom threat comes in...

(7) ... at the end they need their help to save them because all the dead miner ghosts are coming to storm their house, like townspeople with pitchforks... "Light them up, kids!" (Laughter) ... "Hey Phoebe? Put down the flask and pick up the proton pack!"

Excerpts 6 and 7 showcase how the reviewers incorporate the dialogue conventions of franchise blockbuster films (e.g. dialogue-based humour, exposition) (Schauer, 2007) in order to provide their assessment of the film's lacking aspects. Moreover, parody genre conventions are mixed into the review, as the participants impersonate the characters from the discussed films, which is indicated by their intonation and often lowered pitch.

## 6. Conclusions

This case study on multimodal intertextual and interdiscursive practices in online video film reviews has demonstrated a broad array of semiotic resources, including verbal elements, cinematography, sound construction, graphic design and staging, as shown in Table 2 and in the detailed description of the selected examples.

OVFRs are a genre which grants practically unlimited options in matters of structure and delivery of the material. Independent video film reviewers unrestrictedly regulate their evaluations as well as semiotic resources that convey them.

Video film reviewers engage with a variety of existing texts and discourses in constructing their reviews. Interaction, appropriating and sharing across genres discourse communities, disciplinary cultures has been made more accessible in the context of digital communication (Lam, 2013). This case study illustrates how multimodal intertextual references are used to corroborate the evaluations provided by the review authors, engage with the audience and agree or disagree with the evaluations of other parties. Among the most prominent intertextual practices are reviewers' multimodal references to cultural phenomena as a way to draw comparisons between the reviewed film and the films from the same franchise, as well as comment on the commercialisation in contemporary film industry.

The table below summarises the frequencies of intersection between the identified forms of intertextuality and semiotic resources that conveyed them.

	Clips	Images	Gestures	Intonation	Staging & sound design
Cultural references	116	91	97	68	35
Reference to third parties	16	12	31	22	15
Self-references	8	2	5	5	15

Table 3: Intersection between forms of intertextuality and semiotic resources

Edited in clips from films, advertisements and interviews, as well as graphic design mode (e.g. images of prominent personalities in the film-making industry) were employed in the analysed corpus as a visual tool that assists the viewers in discerning the reference and ensuring that its effect is communicated properly. Modes of sound construction and staging, as well as paralinguistic and kinesic modes, occurred in combination with the aforementioned resources in order to enhance and, in several instances, dramatise the evaluative statements made by the reviewers.

Certainly, the success of such practices, and their meaning-making, largely depends on the audience's awareness of specific forms of popular culture. Vásquez (2015) points out that "intertextual references may also help to forge a sense of virtual co-membership among reviewers and reviewers who participate in shared discourse systems ... and they also serve to create a connection between author and audience". Therefore, further studies of intertextuality in OVFRs may require a more thorough investigation of its community-building properties.

## 7. References

- Battaglia, J. (2010). Everyone's a Critic: Film Criticism Through History and Into the Digital Age. Senior Honours Thesis. The College at Brockport.
- Benson, P. (2015). YouTube as text: Spoken interaction analysis and digital discourse. In R. Jones, A. Chik, & C. Hafner (Eds.), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age*. London, UK: Routledge, pp. 81--96.
- Benwell, B. and Stokoe, E. (2006). *Discourse and Identity*. Edinburgh: Edinburgh University Press.
- Bhatia, V. (2010). Interdiscursivity in Professional Communication. *Discourse & Communication*, 4, pp. 32--50.
- Bloor, T., Bloor, M. (2007) *The Practice of Critical Discourse Analysis: An Introduction*. London: Hodder Arnold.
- Janney, R. (2012). Pragmatics and Cinematic Discourse. In *Lodz Papers in Pragmatics*, 8 (1), pp. 85--113
- Lam, P. W. Y. (2013). Interdiscursivity, hypertextuality, multimodality: a corpus-based multimodal move analysis of Internet group buying deals, *Journal of Pragmatics*, 51: pp. 13--39.

- McWhirter, A. (2016). *Film Criticism and Digital Cultures: Journalism, Social Media and the Democratisation of Opinion*. I.B. Tauris.
- Pollach, I. (2006) Electronic word of mouth: a genre analysis of product reviews on consumer opinion websites. In *Proceedings of the 39th Hawaii International Conference on System Sciences, IEEE Computer Society*.
- Schauer, B. (2007) Critics, Clones and Narrative in the Franchise Blockbuster. In *New Review of Film and Television Studies*, 5 (2), pp. 191--210.
- Taboada, M. (2011). Stages in an online review genre. In *Text & Talk - An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 31. pp. 247--269.
- Ungoed Hughes, A., & Riley, H. (2012). The multi-modal Matrix: Common Semiotic Principles in the Seven Modes of Narrative Film. In *Proceedings of the 10th World Congress of the International Association for Semiotic Studies*, pp. 2123--2132.
- Vásquez, C. (2014). *The Discourse of Online Consumer Reviews*. London: Bloomsbury.
- Vásquez, C. (2015). Forms of intertextuality and interdiscursivity in online reviews. In R. Jones, A. Chik, & C. Hafner (Eds.), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age*. London: Routledge, pp. 66--80.

# Scientific communication on social media: Analysing Twitter for knowledge recontextualisation

Ana E. Sancho-Ortiz

Research Institute of Employment, Digital Society and Sustainability (IEDIS)

Department of English and German Studies

Universidad de Zaragoza (Spain)

E-mail: [a.sancho@unizar.es](mailto:a.sancho@unizar.es)

## Abstract

In the last decades and as a result of the growing concern with ensuring the democratisation of science, Twitter has gained importance within the scientific community as a means for the transmission of specialised knowledge both within expert and non-expert audiences. Considering this, the present paper studies the presence of science communication and dissemination on Twitter within a non-strictly scientific context, taking the official accounts of Greenpeace and WWF as its object of analysis. For this purpose, a total of 100 tweets from these accounts were gathered and analysed through manual reading. Based on this analysis, it was found that the use of Twitter for scientific communication primarily responds to informative and engagement purposes, which are materialised through the convergence of the verbal and visual modes and the combination of diverse types of hyperlinks (e.g., hashtags, tags and outbound links). As a result, it was concluded that the scientific use of Twitter features and affordances primarily relates to the recontextualisation and not the generation of specialised knowledge, which has certain implications for the role played by the audience.

**Keywords:** hyperlinks, multimodality, science dissemination, Twitter

## 1. Introduction

Technological advancements have prompted the reconsideration of scientific research as a collaborative experience which concerns every member of society (Kurtulmus, 2021). Such reconsideration has resulted in a renewed interest within the scientific community in enhancing knowledge communication and dissemination (Bondi & Cacchiani, 2021). When applied to the transmission of information beyond specialised boundaries, dissemination becomes popularisation as it “involve[s] the transformation of knowledge into ‘everyday’ or ‘lay’ knowledge, as well as a recontextualization of scientific discourse” (Calsamiglia & Van Dijk, 2004, p. 370). This recontextualisation requires the presence of a mediator as the person (or group of people) in charge of adapting specialised knowledge to the level of expertise of the non-expert audience (Moirand, 2003). In the context of the media, this adaptation process is characterised by the constant interaction between two sets of dimensions, the communicative and the cognitive ones—which, according to Moirand (2003), respectively comprise “the enunciative standpoints [...] of the mediator, utterer and addressee and the representations of the discourse of other groups” (p. 177), and “the designations and reformulations of the states and objects of knowledge” (Moirand, 2003).

Within digital communication, the dynamics of scientific research and dissemination have been influenced by the development of communication platforms with innovative digital affordances, whose exploitation has led to an outburst of online—scientific and academic—genres (Luzón & Pérez-Llantada, 2019). These emerging genres are characterised by their great hybridity and interdiscursivity (Belcher, 2023), qualities enhanced by the strong reliance on hyperlinking found in computer-mediated communication (Belcher, 2023). Following this idea, in their exploration of the growing ecology of digital genres, Askehave and Nielsen (2005) proposed a bidimensional

model whereby these new genres are conceived as both the final text (i.e., a product) consumed by internet users and the newly opened navigation paths (i.e., a means) through which users access other digital texts. Based on this view, these scholars consider that hypertextuality and multi-mediality constitute the essence of digital texts as it is thanks to the combination of hypertextual layers and diverse media in one single text that bidimensionality is obtained.

As regards hypertextuality, a hypertext must be conceived as “a system of non-hierarchical text blocks where the textual elements (nodes) are connected by links” (Askehave & Nielsen, 2005, p. 126). In digital hypertexts, hypertextuality is made explicit and visual to the reader with the introduction of hyperlinks: a navigation feature which allows users to access diverse information sources (Jays et al., 2022). Apart from this informative function, hyperlinks also foster “the emergence of searchable talk, that is, online discourse where the primary function appears to be affiliation” (Zappavigna, 2011, p. 789) between users. The extent to which these embedded hyperlinked paths are followed seems to be partly determined by their visual saliency, an aspect which connects with the other key feature of digital genres according to Askehave and Nielsen (2005): multi-mediality. When commenting on the potential of web texts to combine “text, image, sound, and animation” (p. 125), these scholars conceive (these four) media as meaning-making systems whose combination forces users to carry out “modal shifts” (Askehave & Nielsen, 2005) to successfully understand the text. This idea of media as meaning-making systems has been explored in detail within multimodal research, an approach to the study of human communication which understands any communicative situation as the result of the interaction between semiotic modes (Norris, 2004).

Concerning the concept of semiotic mode—which would, to some extent, equate to Askehave and Nielsen’s conception “medium”, there is no one-size-fits-all definition for it. Still, most scholars conceive them as

meaning-making systems (Bednarek & Caple, 2017; Norris & Maier, 2014) or resources (Bezemer & Kress, 2008) which are “socially shaped and culturally given” (Kress, 2010, p. 79). Two generally accepted mode categories are the verbal mode (traditionally, ‘text’) and the visual mode (namely, images). Traditionally, and as a result of the dominance of linguistics in communication studies, the verbal mode has been established as the prevalent meaning-making system (Kress, 2000). However, scholars such as Kress & Van Leeuwen (2006) have highlighted the semiotic capacity of the verbal mode by equating it to language and grammar in terms of its structural complexity.

In spite of the lack of consensus within mode categories, what seems to be clear is that the specific combination of semiotic resources is equally dependent on a speaker’s individual interest and the sociocultural context in which the communicative situation takes place (Bezemer & Kress, 2016). The conceptualisation of communication and the modes interacting to achieve it as context-dependent phenomena have been extensively explored in language-related domains. An example of these is the field of Genre Studies within which the identification of a genre has been usually linked to the (set of) communicative purpose(s) with which texts are produced in recurrent communicative situations taking place in specific contexts (Askehave, 1999).

Going back to digital communication, the identification of recurrent patterns might be hindered by the aforementioned interdiscursivity and hybridity of digital genres (Belcher, 2023). In one online platform one may find never-ending instances of genre types, realised in the form of merged, hybridised text types which accomplish a variety of –in Askehave’s terms– ‘official’ and ‘hidden’ purposes. To these difficulties in analysing online communication, one might add the fact that online content is constantly being updated and reinvented. In an attempt to fight its immeasurability, many scholars have turned to the quantifying methods of corpus linguistics and studied reduced samples of online content as representatives of general practices (e.g., Siever et al, 2020; Yilmaz et al., 2023). One of the platforms which has received most attention given its informative and interactive layout is Twitter (Squires, 2016), on which one can find works on topics as diverse as politics (Maireder & Ausserhofer, 2014), education (e.g., Tang & Hew, 2017) or science (Insall, 2023). This microblogging platform, originally launched in 2006 as a space to post short, real-time messages known as tweets (Squires, 2016), was bought in October 2022 by the South-African entrepreneur Elon Musk. With this purchase, Musk announced in his own Twitter account that he and his new team would be redesigning the platform to transform it into “the everything app”, a site where users could carry out an infinite diversity of digital practices. As a consequence, the platform has already been subject to many functional and design changes, among which there stand out the limitation in the amount of posts that be read in a day and the renaming of the platform into “X” (X) (and of tweets into “posts”).

Considering all the ideas mentioned above, this paper is aimed to explore how science communication and knowledge recontextualisation work as digitally mediated practices, focusing more specifically on the microblogging platform Twitter (or X). The main intention of this paper is thus to consider how scientific findings are shaped into a digital environment which, far from being originally intended for scientific communication, was aimed at promoting laypeople interaction and conversation around trivial topics (Squires, 2016). In this sense, to understand this paper it must be considered that the data gathered for this study and the initial analyses carried out as part of it date back to the first semester of 2022, before the purchase and renaming of the platform from Twitter into X. As a consequence, it was decided that, to respect the name the platform had when the data analysed was originally collected and posted, the terms “Twitter” and “tweet(s)” will still be used in this paper in substitution to “X” and “post(s)”. Considering all this, the study concentrates on two organisational accounts, @WWF and @Greenpeace, analysing, first, the communicative purpose(s) for which this platform is used and, second, the various forms and uses of multimodal and hyperlinking practices on which these accounts rely to reach such purposes.

## 2. Methodology and corpus

The present paper consists of a corpus-based study carried out around a closed set of tweets with recontextualized scientific knowledge extracted from two international Twitter accounts: @WWF and @Greenpeace. As has been previously explained, the analysis here presented is the result of an exploratory study whose aim was to look into ways in which science communication takes place on Twitter. The ultimate aim is to contribute to the understanding of scientific communication practices in non-expert digital environments—as is the case of social networking sites.

As regards the choice of the accounts, these were selected following a thematic criterion driven by a personal interest in analysing a topic with high social and scientific relevance in the present-day context. Thus, it was established that all tweets had to address topics related to the environmental crisis, an issue which has gained considerable relevance within the scientific community, especially after the formulation of the 2030 Agenda framework for Sustainable Development.

Regarding the methodology followed to select the tweets included in the corpus, these were manually gathered via screenshots according to chronological and numerical criteria. Thus, a sample of 100 was extracted from the “Tweets and replies” section of each account, starting on February 16th 2022 and going backwards in time until the set number (50 per account) was reached. The use of specific downloading software tools (e.g., GoFullPage or NCapture) was discarded given the practicality of screenshots when it comes to selecting specific content (in this case, the tweet) and the need to discard contextual information (e.g., the suggested accounts provided by Twitter or replies to the tweet). This type of contextual

information would have been distracting for the main aim of the study, i.e., getting an overview of the practices carried out by the two accounts as instances of science communication on Twitter.

With regard to the analysis of the corpus, the tweets were analysed individually through manual reading attending to four criteria: linguistic realisations, multimodal elements, hyperlinking, and communicative purpose(s). For the linguistic realisations, the tweets were manually tagged considering orthotypographic aspects (namely punctuation marks and instances of non-standard capitalization), semantic categories (with a focus on nouns), verbal tenses and moods, and layout or move structure. As regards multimodal elements, tagging was carried out in terms of the verbal mode (including any instance of verbal text), the visual mode (including emoji and image files) and the spatial mode (considering the disposition of the previous elements in the tweet). For hyperlinking, the categories used for tagging were internal hyperlinks (namely quote tweets, hashtags and tagging) and outbound links (to other platforms). Retweets (a type of internal hyperlink) were left out in this analysis for two reasons: 1) my interest in the way science dissemination accounts exploit Twitter features to craft scientific content, and 2) the considerable difference in the frequency of retweets found between @WWF (with 2 retweets in its timespan) and @Greenpeace (with over 50 retweets). Lastly, communicative purposes were marked based on the features identified in the previous criteria and the dimensions proposed by Moirand (2003) as regards scientific communication: informative purposes (for the cognitive dimension) and engagement purposes (for the communicative dimension).

After the analysis of each individual tweet, the contrasted data led to a preliminary set of shared communication patterns which were identified, first, in each account, and then, in the two accounts together as examples of science communication on Twitter. The data and patterns here found were taken as an initial step in the attempt to cast some light on what the most salient features of science communication and knowledge recontextualisation could be in the context of Twitter communication.

### 3. Results and discussion

The analysis of the WWF and Greenpeace accounts led to the identification of two potential common patterns in this digital practice (science communication on Twitter): a) the identification of shared communicative purpose(s), and b) the convenient exploitation of multimodality (namely the verbal and visual modes) and hyperlinking to achieve such purpose(s).

Regarding communicative purposes, the analysis of the corpus reveals that there are two main co-existent purposes with which WWF and Greenpeace post their tweets: i) informative, seen in those tweets aimed at providing the audience with scientific data, and ii) engaging, found in those intended to involve the reader in the topics and issues addressed. The identification of these two purposes derives from the understanding that scientific communication is

necessarily the result of the interaction between the communicative and cognitive dimensions defined above—as was proposed by Moirand (2003) in her analysis of science communication in the French media. According to Moirand (2003), scientific data constitute the object of knowledge which is extracted from its original scientific context to undergo a set of modifications, fit in a new context (i.e., be recontextualised) and become an object of media. These data, which constitute “the linguistic output of the scientific community” (Moirand, 2003, p. 179), are described and reformulated (two cognitive processes) depending on the standpoint of the mediator recontextualising it and the receiver consuming it (participants considered from the communicative dimensions). Based on this, it could be argued that the object of knowledge Moirand (2003) conceives is in fact the scientific data presented in the tweets of the WWF and Greenpeace accounts for informative purposes. The shape these data take in the tweet would thus depend on the standpoint of the organisations with regard to these data—in this case, an eagerness to denounce the environmental issues addressed and engage the audience in them. Once filtered through that denouncing standpoint, the data becomes an object of the media (in this case Twitter) reformulated in a new digital environment.

As regards the instantiation of informative and engagement purposes, while the former (present in 92% of the tweets) are materialised through the introduction of scientific facts and data (such as the percentage “~30%” shown in Figure 1), the latter (in 87%) tend to be present through diverse functions. These functions, which can also be referred to as secondary purposes for engagement, are 1) awareness-raising (found in 76% of the tweets), i.e., making the audience realise the urgency to address certain issues, 2) persuasive (in 46%), i.e., prompting the reader to perform specific actions and feel in a certain way, and 3) self-promoting (in 34%), i.e., posting organisation-related content to promote their campaigns and members.



Figure 1: Tweet from the Greenpeace account

The high frequency of both informative and engagement purposes might indicate there is a significant interdependence between the two. In other words, it seems that in the same way that the audience needs to be well-informed about an (in this case environmental) issue to feel engaged in it, it seems necessary for the topic to be presented in an appealing format for the reader to feel eager to learn more about it. Apart from this, the analysis of the corpus demonstrates that, given their need to craft fully informative, engaging messages, this type of Twitter



account conveniently exploits the technological affordances available on the platform. As was anticipated before, this includes producing highly multimodal and hypertextual tweets whose features, shared by the two accounts analysed, might represent a new set of recontextualisation patterns for digital scientific communication on Twitter.

With regard to the exploitation of multimodality, in all the tweets gathered for this analysis there is a convergence of two specific modes, the verbal and the visual, which points to the importance of diversifying the means for information conveyance in science communication. By crafting the same message in different formats (e.g., through visual and verbal elements) experts are more likely to make scientific data comprehensible, and thus, fit the expertise level of their diversified (Twitter) audience.

Concerning the visual mode, the analysis of the corpus shows that there are two main types of visuals: emoji (found in 51% of the tweets) and images (in 78%). These visual elements contribute to the achievement of engagement purposes by making tweets more visually appealing and, thus, enhancing reader's interest in them. Their most significant functions are visually indicating either the topic of the tweet (as in Figure 2 in which both the emoji and the image file show a tiger because the tweet deals with the Tiger Chinese Year) or the emotion the reader should be feeling—this latter option fulfilling a persuasive function.



Figure 2: Tweet from the WWF account

Apart from conveying emotion and highlighting the topic of a tweet, visual elements (namely emoji) also fulfil the function of discourse markers (in 18% of the tweets). This is generally achieved by introducing conceptual emoji such as arrows (as in Figure 3), with which the accounts manage to, first, make reading easier by organising discourse, and second, make new navigation paths more appealing by encouraging the reader to click on the hyperlink to which these arrow-emoji point. Hence, they contribute to informing and awareness-raising (engagement) purposes as they give prominence to information sources such as the websites linked. In this line, another type of visual element favouring these informative and awareness-raising purposes are hyperlinked images (found in 33% of the tweets) attached to outbound links—which are different from individual image files as the one in Figure 2. These hyperlinked images carry out a summarising function as

they indicate the main topic of the article linked via hyperlinking, thus promoting direct access to them.

Concerning the verbal mode, it primarily contributes to the achievement of informative purposes since all specialised scientific data are conveyed via conventional alphanumeric elements. Nonetheless, the Greenpeace and WWF accounts also introduce specific reader-engagement features in their tweets to make the audience participant in their discourse (see Table 1).

Reader-engagement features	Frequency of use in the total corpus
In-group identity markers (“we”, “our”, “us”, “ourselves”)	31%
Direct addresses to the reader (“you”)	30%
Rhetorical questions	34%
Imperative forms	36%

Table 1: Types of reader-engagement features identified in the WWF-Greenpeace corpus and their frequency of appearance in it.

Out of these features, there stands out the reliance on rhetorical questions (used in 34% of the selected sample) which seems to primarily respond to persuasive and awareness-raising purposes. As a linguistic strategy traditionally used to encourage reflection (Swasy & Munch, 1985), it is likely that this type of question is introduced in these denouncing tweets to cause an attitudinal change in their audience towards the environmental dangers addressed. Apart from this, it is also significant how direct addresses to the audience (in 30% of the tweets) and directive speech acts in the form of imperatives (in 36%) are introduced as means to make an impact on the reader. This impact is further reinforced by the diverse realisations of the inclusive “we” used as in-group identity markers in the tweets analysed (in 31% of them). From a pragmatic perspective, the use of first-person plural forms is perceived as a linguistic strategy used by speakers to reduce the interpersonal distance with their interlocutors and thus establish a closer bond with them (Brown & Levinson, 1987). Nonetheless, these originally deictic plural forms might also fulfil an exclusive function if the group which they designate differs from their addressee. Sometimes this seems to be the case with the WWF and Greenpeace accounts here analysed, whose sampling tweets present a 14% of first-person plural forms used to refer to actions carried out by the organisations. For example, in figure 3 the WWF account distinctively uses “our” to refer to (everyone’s) world leaders and “us” to refer to the organisation themselves, marking a distance between what needs to be collectively addressed (i.e., the environmental inaction of world leaders) and what is already being done with regard to it (i.e., the initiative of the organisation against plastic pollution).





Figure 3: Tweet from the WWF account

Overall, despite some examples of linguistic exclusivity, this type of account seems to be more prone to introducing the audience in their discourse of environmental activism. It might be understood that, by doing this, the readers will more likely assume an active role in denouncing and fighting against anti-environmental policies—hence the awareness-raising and persuasive function of these linguistic strategies.

Lastly, the analysis demonstrates that this type of account greatly relies on hyperlinking to reach informative purposes. This is attained by introducing diverse types of links (see Table 2 for these types and their frequency of appearance in the corpus) which redirect the audience to other information sources in which extended and more detailed information on the topic addressed in each tweet is provided. As a consequence, there is a change in the knowledge production and sharing dynamics between experts and non-experts, as the latter become active participants in the disseminating process. It is up to the reader to decide whether they want to extensively inform themselves about the topic to they are exposed with the tweets—hence the importance of engaging the audience with, among other strategies, visually salient elements.

Hyperlinks	Percentage of use
Hashtags	68%
Outbound links	62%
Tags	25%
Quoted tweet	12%

Table 2: Hyperlink types and their frequency of appearance in the corpus

Table 2 shows that there is a predominance of hashtags and outbound links, which in addition to their informative functionality, also contribute to fostering engagement. In the case of hashtags, this is generally achieved by giving visual saliency to the keywords of the tweet (as in Figure 1 with #NatureEmergency, #LessMeat and #LessHeat; and Figure 2 with #Tiger and #Yearofthetiger)—a strategy which further enhances the multimodal nature of this type of tweet. Apart from the visually enhancing function of single- and two-word hashtags, three-word or longer hashtags seem to perform a slogan-like function. As it is

typical of slogans (Denton, 1980), these hashtags outline specific (environmental) concerns in an attempt to vindicate the lack of or the little action taken with regard to them. For example, in Figure 3, the imperative form #StopPlasticPollution is introduced in an attempt to mobilise the audience against it.

As regards tags, these regularly serve self-promotional purposes as they facilitate access to Twitter profiles affiliated with or related to both Greenpeace and WWF. These profiles can either belong to members of the organisations or to subordinated organisational departments/affiliations (such as @GreenpeaceNL in Figure 1 and @wwf\_tigers, @WWFmy in Figure 2). Similarly, outbound links also fulfil a self-promotional function as they redirect the reader to the official website of these companies (with an external link), hence contributing to increasing their views and range of influence. The relatively high-frequency of outbound links, as well as that of the visually salient hyperlinked images attached to them, suggest that Twitter primarily operates as a recontextualisation platform. This hypothesis derives from the idea that the scientific content shared by these accounts is not originally generated in the platform but extracted from external websites linked and recontextualised and remediatised into Twitter.

## 4. Conclusion

Social networking sites such as Twitter have become a key means for the transmission of specialised knowledge as they allow experts to reach diverse specialised and non-specialised audiences. In this context, this paper has demonstrated that science communication on Twitter requires the remediation of specialised knowledge into highly hypertextual and multimodal texts (tweets). Indeed, the results reached in this preliminary study as regards this multimodal and hypertextual exploitation of the affordances of the platform point to the idea that Twitter is primarily conceived by the scientific community as space for knowledge recontextualisation. This entails that Twitter—and perhaps other social media sites—would not be used as spaces to generate scientific knowledge, but rather just as means to transmit it. Therefore, its functionality would only apply to the latest stages of scientific research: transmitting, sharing, and eventually, democratising knowledge.

As regards the recontextualisation practices, these are materialised in the crafting of multimodal tweets (namely verbal-visual ensembles) and the introduction of platform-specific digital features such as the use of hashtags and outbound links. The main consequence of these digital practices seems to be the increasingly active role given to the reader, who stops being perceived as a mere passive receiver of scientific knowledge and becomes an agent able to interact with knowledge in various ways—either responding to the tweets, giving them a like or a Retweet or following the hyperlinks included. Because of these new interactive dynamics, the reader becomes a fundamental part not only in the transmission of knowledge, but also in its eventual democratisation.

Conclusively, the strong reliance on multimodality and hyperlinking found in the two accounts here analysed points to the emergence of scientific communication trends that seem to be characteristic of science recontextualisation on social media and social networks. However, it is still yet to be confirmed whether these patterns –and thus Twitter as a dissemination platform– are means intended only for knowledge recontextualisation or if they are also used for knowledge production.

## 5. References in the text

- Askehave, I. (1999). Communicative purpose as genre determinant. *HERMES-Journal of Language and Communication in Business*, (23), 13-23
- Askehave, I., & Nielsen, A. E. (2005). Digital genres: a challenge to traditional genre theory. *Information technology & people*, 18(2), 120-141.
- Bednarek, M., & Caple, H. (2017). *The discourse of news values: How news organizations create newsworthiness*. Oxford University Press.
- Belcher, D. D. (2023). Digital genres: What they are, what they do, and why we need to better understand them. *English for Specific Purposes*, 70, 33-43.
- Bezemer, J., & Kress, G. (2008). Writing in Multimodal Texts: A Social Semiotic Account of Designs for Learning. *Written Communication*, 25(2), 165–195.
- Bondi, M., & Cacchiani, S. (2021). Knowledge communication and knowledge dissemination in a digital world. *Journal of Pragmatics*, 186, 117-123.
- Brown, P., & Levinson, S. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.
- Calsamiglia, H., & Van Dijk, T. A. (2004). Popularization discourse and knowledge about the genome. *Discourse & society*, 15(4), 369-389
- Denton Jr, R. E. (1980). The rhetorical functions of slogans: Classifications and characteristics. *Communication quarterly*, 28(2), 10-18.
- Insall, R. (2023). Science Twitter—navigating change in science communication. *Nature reviews molecular cell biology*, 24, 305-306.
- Jayes, L. T., Fitzsimmons, G., Weal, M. J., Kaakinen, J. K., & Drieghe, D. (2022). The impact of hyperlinks, skim reading and perceived importance when reading on the Web. *PLoS ONE*, 17(2): e0263669.
- Kress, G. (2000). Multimodality. In B. Cope and M. Kalantzis (Eds.) *Multiliteracies: Literacy Learning and the Design of Social Futures* (pp. 182–202). Routledge.
- Kress, G. (2010). *Multimodality. A social semiotic approach to contemporary communication*. Routledge.
- Kress, G., & Van Leeuwen, T. (2006). *Reading images: The grammar of visual design*. Routledge.
- Kurtuluş, F. (2021). The democratization of science. In Ludwig, D., Koskinen, I., Mncube, Z., Poliseli, L. & Reyes-Galindo, L. (Eds.) *Global epistemologies and philosophies of science* (pp. 145-154). Routledge.
- Luzón, M. J., & Pérez-Llantada, C. (2019). *Science Communication on the Internet: old genres meet new genres* (Vol. 308). John Benjamins Publishing Company.
- Maireder, A., & Ausserhofer, J. (2014). Political discourses on Twitter: Networking topics, objects, and people. In Weller et al., (Eds.) *Twitter and society* (pp. 305-318). Peter Lang.
- Moirand, S. (2003). Communicative and cognitive dimensions of discourse on science in the French mass media. *Discourse studies*, 5(2), 175-206.
- Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. Routledge.
- Norris, S., & Maier, C. D. (2014). *Interactions, images and texts: A reader in multimodality*. Walter de Gruyter.
- Siever, C. M., Siever, T., & Stöckl, H. (2020). Emoji-text relations on Instagram: Empirical corpus studies on multimodal uses of the iconographic mode. In Stöckl, H., Caple, H., & Pflaeging, J. (Eds.) *Shifts towards image-centricity in contemporary multimodal practices* (pp. 177-203). Routledge.
- Squires, L. (2016). Twitter: Design, discourse, and the implications of public text. In Georgakopoulou, A., & Spilioti, T. (Eds.) *The Routledge handbook of language and digital communication* (pp. 239-256). Routledge.
- Swasy, J. L., & Munch, J. M. (1985). Examining the target of receiver elaborations: Rhetorical question effects on source processing and persuasion. *Journal of consumer research*, 11(4), 877-886.
- Tang, Y., & Hew, K. F. (2017). Using Twitter for education: Beneficial or simply a waste of time?. *Computers & education*, 106, 97-118.
- Yılmaz, F., Elmas, T., & Eröz, B. (2023). Twitter-based analysis of anti-refugee discourses in Türkiye. *Discourse & Communication*, 17(3), 298-318.
- Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New media & society*, 13(5), 788-806.

# Can I Publish my Social Media Corpus? Legal Considerations for Data Publication

Oliver Watteler, Ulrike Schneider

Leibniz Institute for the Social Sciences (GESIS) Köln, University of Mainz

oliver.watteler@gesis.org, ulrike.schneider@uni-mainz.de

## Abstract

This paper is intended as an aid for linguists and other researchers wishing to compile and publish collections of social media data. Based on the example of the Corpus of Political Tweets by Trump and US Senators (PoTTUS), it demonstrates potential technical steps for obtaining social media data and discusses the requirements social media data needs to fulfil to make publication possible, especially the legal basis for publication. The discussion reveals that oftentimes a compromise between researchers' wishes and the legal limitations imposed by social media companies is the only viable solution. In the case of Twitter data, this might mean publication of Tweet IDs instead of the content of the tweets. The downside of this is so called 'data rot', i.e. loss of parts of the data.

**Keywords:** Twitter, social media, data publication, data sharing

## 1 Introduction

Every day, new social media corpora are being compiled for research purposes. At the same time, it has become good scientific practice to make data publicly available to allow subsequent use by other researchers (e.g., DFG, 2015). Therefore, many researchers wish to publish their social media data. This, however, poses some unique challenges, not least because social media constitute a rapidly changing organizational and technical environment which leads researchers to operate in a context in which not all legal and ethical issues have been resolved in satisfying ways and best practice workflows are yet to be established. Twitter is a very good point in case.

The present paper is intended as an aid for linguists and other researchers wishing to compile and publish collections of social media data. It demonstrates which factors must be taken into consideration when assessing whether and how publication of social media data is possible under German/EU law. The use case is the PoTTUS Corpus of Political Tweets by Donald Trump and US Senators compiled at the University of Mainz in 2020/2021, and the options for archiving and publication offered by GESIS – Leibniz Institute for the Social Sciences.

The rest of the paper is structured as follows. Section 2 details the current state of affairs concerning the publication of social media data. Section 3 introduces the PoTTUS Corpus and provides information on how it was compiled. Section 4 introduces GESIS and its data services, before Section 5 demonstrates which questions need to be asked in order to determine whether publication is a viable option. Section 6 presents available options for publication for the PoTTUS corpus. The paper concludes with Section 7 which provides suggestions and guidelines for researchers currently compiling social media corpora.

## 2 Publication of Social Media Data

The Social Media Research Group (2016) defines social media as “web-based platforms that enable and facilitate users to generate and share content, allowing subsequent online interactions with other users (where users are usually, but not always, individuals)”. Thus, social media data comprises user-generated content, like images, photographs, or texts, as well as information about the users and their interactions with others. All of this can be retrieved from platforms that operate as private enterprises. The result is a legally complex situation.

In the next section, we therefore look at the legal basis for collecting social media data before we take a brief look at the differences between web scraping and the use of APIs for accessing the data. Finally, the challenges for data publishing which arise from the contractual relation between the researcher and the platform provider are outlined.

### 2.1 The Legal Basis for Collecting Social Media Data

We usually have to consider at least three legal areas to determine whether and how research data can be published: data protection, copyright law, and contractual agreements.

Let us first consider data protection. It is needed whenever we are dealing with personal data. The latter, in turn, is defined very broadly as

any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (GDPR, Article 4)

If we think of typical contents of social media, posts, opinions, personal preferences, or images of the account

holder come to mind. As all of these are likely to reveal details about the user's gender, age, social status etc., we have to treat social media data as personal data (Watteler, 2022). Therefore, the user as well as any others about whom information is revealed need to be protected from unintended or negative consequences resulting from the use of their data. In Europe, this is governed by the General Data Protection Regulation (GDPR), which demands (Article 6 (1)) a legal basis for the processing of personal data, such as informed consent by 'research subjects', a contractual agreement or proof of a legitimate interest by the researchers.

The second legal area to consider is copyright.

Copyright legislation is part of the wider body of law known as intellectual property (IP) which refers broadly to the creations of the human mind. IP rights protect the interests of innovators and creators by giving them rights over their creation. (WIPO, 2016: 3)

These creations can be literary and artistic works, like (longer) texts, photographs, drawings, illustrations or construction plans. Their use – even for research purposes – is legally restricted by copyright legislation (WIPO, 2016).

When thinking about user-generated content on social media platforms which might fall under copyright protection, posted visual creations readily come to mind. Yet, theoretically, the postings themselves, even if no longer than 140 (or 280) characters, could be protected by copyright. However, postings communicating facts are not protectable and, as a general rule, we can assume that "tweets will not be protected by copyright law, and such protected tweets are extremely rare" (Beurskens, 2014).

The third legal area concerns contracts, which are legally binding agreements holding between two or more parties (Martin, 2003). Even consenting to terms of service by means of affirmative confirmation like clicking a checkbox ('click wrap') constitutes such a contract (Brehm & Lee, 2015: 5; Vogel & Hilgendorf, 2018). Researchers are not exempt and have to adhere to the rules and regulations set out in those agreements.

Issues with copyright and data protection can be circumvented by means of users giving consent to their data being used for analysis. Yet, depending on the nature of a social media corpus, it can become virtually impossible to get consent from all users included, as their numbers may run into the thousands.

Before we discuss these in detail, we need to briefly look at the distinction between 'web scraping' and the use of application programming interfaces (APIs) as choosing one over the other can legally make a difference.

## 2.2 Web Scraping versus Using an API for Data Collection

Web Scraping consists of the three interrelated tasks: First, the underlying structure of a website is examined to determine how the data is being stored. Secondly, the website is crawled, which involves developing and running a script that automatically browses the website and

retrieves the needed data. Finally, the data must be cleaned, pre-processed and organized in a way that enables further analysis (Krotov et al., 2020: 556–557).

Scraping is explicitly permitted by the German Act on Copyright and Related Rights (UrhG 1965, 2018) and does not hinge on or constitute a contractual agreement with the website provider (Vogel & Hilgendorf, 2018). Rights of the individual, like data protection or copyright, nevertheless still apply. Thus, in this scenario, the legal basis of data processing is formed by laws.

An API, on the other hand, is a set of functions which allows software to communicate with the application or service for which the API is provided (de Souza et al., 2004: 64). The legal basis for data processing is the same as for scraping, but users usually have to agree to terms of service before being able to use a specific API. These then constitute contractual agreements for data accessing, which in turn means that platform providers and their interests may have an impact on data processing even beyond the end of the research project.

## 2.3 Challenges of Providing Platform Data for Secondary Use

Once researchers have legally obtained social media data and want to publish it, they might face further obstacles. One rests again in the contractual relation between them and the platform providers, another one is the lack of best practice guidelines for data publishing.

When it comes to the obstacles imposed by contractual relations, Twitter is a case in point; two events highlight the challenges. In the aftermath of the riots of January 2021, which followed Trump's defeat in the presidential elections, Twitter suspended the president's account. As a consequence, his tweets were no longer accessible and even the US National Record Administration (NARA), in charge of preserving documents and communication, for example, of outgoing presidents, was unable to obtain permission from Twitter to access this data (Forgey, 2021). Thus, historically important data is at risk of being lost.

Two years later, on February 2<sup>nd</sup>, 2023, Twitter announced the end of free access to its API, which would not only impact commercial enterprises, but also endanger public services like that of libraries. It led the German National Library (DNB) to put out a distress call to researchers to assemble an emergency corpus of German language tweets on February 20<sup>th</sup> of the same year.

The challenges imposed by terms of service and the fast pace at which the platforms' services, regulations and technology evolve have seriously hampered archiving services aiming at making research data available for secondary use. Weller & Kinder-Kurlanda (2016) were among the first to call for concerted action in this matter and Williams et al. (2017) devised a first decision flow chart for the possible publication of Twitter data.

Events like the ones mentioned and many complaints by researchers about the overall unsatisfactory situation have led the European Commission to change European legislation and to initiate the European Digital Media

Observatory (EDMO). The Digital Service Act (DSA), for example, now regulates the responsibilities of digital services within the EU with the aim to connect consumers with goods, services and content, but it also affects researchers' use of platform data. A core idea behind the DSA is that institutions like GESIS should act as data trustees managing access, for example, to social media data on behalf of platform providers. And the EDMO Working Group on Platform-to-Researcher Data Access has worked on a code of conduct for data access under Article 40 of the GDPR (EDMO, 2022).

Nevertheless, there is still no universally agreed-upon standard path for the publication of social media data in general and of Twitter data in particular. This is why PoTTUS and other projects are so important, because they are piloting the development of such a standard.

### 3 The PoTTUS Corpus

We will now offer a brief account of the reasons behind the compilation of the PoTTUS corpus as well as of the methodology employed for its collection.

During the Trump presidency, Twitter played an unprecedented role in American politics. Trump joined Twitter in 2009, initially using tweets to promote his businesses, but the content of his tweets changed over time. He began using it for political commentary and as a campaign tool, before eventually making his private account the official presidential one. The language and style of his tweets have garnered a lot of linguistic attention. At the time of compilation, most studies exclusively focused on Trump's account (cf. e.g., Ott, 2017; Clarke & Grieve, 2019), yet we felt that certain points which were being discussed, such as whether Trump's language was particularly negative or emotional, required comparative analyses – after all, Trump is not the only politician who has taken to Twitter and his might merely be a prominent specimen of a common tweeting style. In order to make such comparative analyses possible (cf. e.g. Schneider 2021<sup>1</sup>), in November 2020, we started compiling the PoTTUS Corpus, which combines the tweets sent from Trump's account @realDonaldTrump with tweets sent by all US senators in office during the Trump presidency. The following subsections provide a step-by-step description of how the corpus was compiled.<sup>2</sup>

#### 3.1 Choice of Accounts and Time Frame

We ensured that only accounts with a blue tick mark became part of the corpus. At the time, this mark indicated “active, notable, and authentic accounts of public interest that Twitter had independently verified” (Twitter Help Center). As it was not possible to determine at which point the account had been verified, all tweets since the registration of the account and the day of compilation were included. Neither retweets of the senators' or Trump's

tweets by other users, nor messages retweeted by the politicians' accounts were included in the corpus.

#### 3.2 Data Collection

Previously, the most common method of collecting data from Twitter had been by means of the Twitter API. In November of 2020, however, Twitter was in the process of updating the API. Version 1 no longer had full functionality and Version 2 was only available in a beta version for which no scripts existed yet. We therefore pursued a different route, using the Python script *snsrape* (JustAnotherArchivist 2020). As suggested by its name, this application scraped data from social media pages. The script produced a list of URLs including the unique ID of each tweet.

Twitter described these IDs as ‘dehydrated’ versions of the tweets, which could later be ‘rehydrated’ to obtain the full tweet including meta-information. Crucially, Twitter requires third-party databases to reflect the state of a tweet as it currently appears on Twitter. Therefore, after rehydration, a tweet which was edited will appear in its changed form and deleted tweets cannot be rehydrated, which leads to so-called ‘data rot’ (Walker, 2017). The rehydrated corpus thus might look different from the original data.

Rehydration is possible with the help of a Twitter Developer Account. By creating a Developer Account for the project, we agreed to the Twitter ToS. We furthermore used one of the third-party hydrators recommended by Twitter at the time (Documenting the Now 2020).

The result of rehydration is a csv file containing the tweets as well as their IDs and meta information, such as the Twitter handles of the senders, the version of Twitter used by the senders (e.g., Desktop, Android), geo-coordinates (if the user provided them) etc.

#### 3.3 Legal Basis for Data Collection

As legal basis for data collection we relied (a) on the text and data mining exemptions for research as laid out in the German Copyright Act, and (b) on our legitimate interest as researchers as indicated in the GDPR.

The German Copyright Act at the time permitted the automatic and systematic reproduction of a large number of works (original material) for scientific research in order to create a corpus for analysis. It also allowed making the corpus publicly accessible to a specifically delimited group of persons for the purpose of joint scientific research and to individual third parties for the purpose of reviewing the quality of scientific research. (§60d (1) UrhG, 1 March 2018). From what we saw in the tweets, we did not consider the material collected as protected by copyright.

Concerning data protection we did not consider informed consent a feasible option for data collection and based our activities on our legitimate interest as researchers following Article 6 (1) f of the GDPR.

project at the time, who undertook a lot of the research and technical steps.

<sup>1</sup> Based on a predecessor of the PoTTUS Corpus.

<sup>2</sup> Our thanks go to Julia Schilling, assistant to the

### 3.4 Data Clean-Up and Annotation

We removed a lot of the meta information, particularly points which related to the account rather than the individual tweet. To make the tweets easier to handle with different software applications, we removed line breaks within the tweets, changed the format of the dates in the meta information and reinserted special characters, which had been replaced with their HTML-code by the hydrator (e.g., `&amp;`). We also temporarily removed all meta-information to obtain word-counts (the corpus totals over 25 million words).

Most importantly, we knew that Trump had only begin using the official retweet function in 2016 and had previously used various other methods of indicating that a tweet was a retweet, such as adding *Via @account* to the retweeted material. We wanted to make it possible for users to exclude these tweets as well as others which constitute mostly of quotations. Using regular expressions we annotated these tweets as quotes or manual retweets.

In 2022, we contacted GESIS to explore the possibilities to publish our corpus and make it accessible for secondary use.

## 4 GESIS Services for Publishing and Archiving Data

GESIS is Germany's largest social science data infrastructure. It covers the entire life cycle of empirical research and offers a broad spectrum of research-based services for empirical social research. Its data holdings mostly consist of quantitative data gathered in surveys, but GESIS increasingly covers so-called 'digital behavioural data', which comprises social media and other born-digital data types.

GESIS' general approach to data preservation and publishing is shared by many similar repositories which form part of the German Data Forum (RatSWD) or the Consortium of European Social Science Data Archives (CESSDA). In brief, researchers wanting to publish their data via GESIS determine for whom the data shall be accessible, transfer the data as well as related metadata and documents to the repository, and grant GESIS basic usage rights (see GESIS website at <https://www.gesis.org/en/data-services/home> for access to archiving contracts). Only data which fulfils the legal requirements can be made available for secondary use. This approach, amongst others, formed the basis for the newly drafted model contract for data ingest by the National Research Data Infrastructure – NFDI (Schallaböck et al., 2023).

GESIS is currently expanding its workflow to incorporate social media data with the help of projects like PoTTUS or PEP-TF on monitoring social media use during the campaigns for the 2013 Bundestag elections in Germany (Kaczmarek et al., 2014).

## 5 PoTTUS as a Use Case

In the following, we will use PoTTUS as a use case to demonstrate how we can determine whether data meets criteria for publication.

### 5.1 Documentation

Each study entering the GESIS holdings is documented in a highly structured way using the Data Documentation Initiative (DDI) standard. In this way, GESIS helps to ensure that the data are easy to find and thus meet the FAIR criteria (*F*indable, *A*ccessible, *I*nteroperable, *R*eusable). Furthermore, each study is registered and receives a persistent identifier in the form of a DOI. Furthermore, version control is implemented, which tracks potential changes in the data and documentation.

Along with the data, GESIS preserves all documents necessary to trace the origins of the data. In the case of surveys, these include questionnaires and method reports; for social media data, these could be scripts for scraping the data online, coding schemes, methodological descriptions, and, if applicable, the Terms of Service (ToS) under which data was collected and which regulate its processing. If not downloaded at the time of signing, they might be retrieved later. Twitter, for instance, provides older versions of their ToS online (Twitter 2023).

In the case of PoTTUS, data selection, retrieval method, software and scripts as well as further processing steps were well documented. It therefore meets the criteria set by GESIS.

### 5.2 Data Format

To guarantee the longevity of the data, Microsoft Excel spreadsheets, for instance, have to be converted to CSV files, which can easily be done. This transformation along with adequate documentation distinguishes mere bitstream preservation from true long-term preservation, i.e. it is the difference between merely keeping files 'alive' to keeping data readable and usable.

### 5.3 Legal Basis for Processing

Firstly, we need to assess whether the tweets contained in the corpus fall under copyright law. In the present case, Twitter was used for political communication. Images were not retrieved from Twitter. Thus, we can reasonably expect that no part of the corpus is copyright protected.

Secondly, we need to look at data protection issues. The GDPR does not make a distinction between 'regular' individuals and those that might be considered 'public figures'. And since Donald Trump as well as the Senators reveal their political and private views, we considered the data personal. We relied on legitimate interest as researchers to process the data. This needed to be assessed.

For this assessment we need to weigh the project's interest against the interest of the subjects. This can be done on the basis of a checklist provided by the Article 29 Working Group (2014). As a result of this balance check, we can say that PoTTUS as a research project followed a

lawful and current interest. The data collection was necessary to reach the purpose connected to this interest. No fundamental rights or interests of the subjects were found which would override the researchers' interest, and the project undertook extra safeguards like data minimization and mere publication for scientific re-use to protect the data subjects. The project therefore had a legitimate interest when collecting the tweets.

Thirdly, since the PoTTUS corpus was generated using a Twitter Developer Account and thus the Twitter API, the project had to agree to the Twitter Terms of Service (ToS) and we are looking at a contractual agreement for the processing of the data. In the case of PoTTUS, the Twitter ToS valid from January 1<sup>st</sup>, 2020 apply (version 14, current version no.16). These ToS explicitly granted academic researchers the right to restricted redistribution of so-called 'Twitter Content', more specifically to distribute an unlimited number of Tweet IDs and/or User IDs if they were doing so on behalf of an academic institution and for the sole purpose of non-commercial research (Developer Agreement and Policy, Content redistribution, March 10th, 2020). Yet they did not allow to share tweet contents.

While at the time of retrieval Twitter distinguished regular from so-called 'verified accounts' (with the now infamous blue tick mark), this distinction did not carry over to the ToS, which made no special allowance for the use of tweets from verified accounts. And even the @realdonaldtrump account's status as an official outlet of the US President at the time did not give leeway for official archiving. Instead, Twitter stated that they expect any use of Twitter content by third parties to be consistent "with peoples' reasonable expectations of privacy" (Developer Agreement and Policy, Privacy and control are essential, March 10th, 2020).

Publication of the IDs without the content of the tweets circumvents these issues: Although the IDs directly link to existing tweets and can be related to the account holders, neither the postings nor any copyright-protected material are being directly shared by the researchers.

It goes beyond the scope of this paper to discuss the question whether Twitter's ToS violate European law, for example, by being overly complex or by permitting data processing for research, which a lot of users are not aware of (Kennedy et al., 2017).

## 6 Conclusion

Based on the considerations listed above, we conclude that publication of the PoTTUS corpus is subject to restrictions which also apply to many similar endeavours. The crucial points are the legal limits concerning the use of Twitter data as determined by the legal framework as well as Twitter's ToS.

Although it seems of general interest that announcements made by public figures like the Ex-President of the United States should be free to use at least for research purposes, Twitter's requirement for third party databases to reflect the state of a tweet as it currently

appears on Twitter means that, even if full-text publication were possible, the corpus would need constant maintenance. This restriction could have possibly been circumvented by seeking consent from the research subjects to use their data. But it seems doubtful that the project would have obtained this consent.

The fact that only Tweet IDs and not entire tweets can be distributed solves this issue: When rehydrated, a changed tweet will appear in its new form and a deleted one cannot be rehydrated at all. Yet this also causes a range of undesirable issues: Firstly, the database is fluid and changing, making it impossible to replicate studies exactly. Secondly, it may lead to so-called 'data rot', i.e. to Tweet IDs which are 'dead ends' as they link to non-existing tweets, preventing rehydration. The rot could be considerable. While, for example, Trump has been re-admitted to Twitter under new owner Elon Musk, any tweet he sent prior to the ban is no longer accessible on Twitter, rendering the PoTTUS Corpus a corpus without the ex-POTUS. Finally, it makes the database difficult to access. The introduction of the new API pricing scheme in March of 2023 included revoking special privileges previously granted to researchers and led to many applications losing functionality. At the time of writing (early August 2023), the last of the academic accounts seem to have ceased to function and a new solution for academia, hinted at by Twitter in March of 2023 via their account @TwitterDev (recently @XDevelopers), has not (yet) materialised. As a result, rehydration – or "tweets lookup", as it is now referred to – not only requires technical expertise but also a paid subscription. The senators' section of the corpus alone runs up to over one million tweets. Under current conditions, rehydrating such a large database would incur considerable costs.

## 7 Suggestions for Future Researchers

We would advise everyone who is about to compile a social media corpus to carefully consider the legal conditions before collecting data. And those who are currently compiling social media data to document each step well. Keep record of third-party software used (and which version), when it was downloaded and (if applicable) which search strings were used. Download and save Read Me files, legal agreements and terms of use. Archive websites by making them offline-available, downloading the html code or by screenshotting important pages. Accurately date these records. More available information can speed up later assessments concerning options for publication, can offer sound arguments in favour of publication, or – if all else fails – can provide justification to funding bodies why (full) publication is not possible.

## References

- Akdeniz, E.; Borschewski, K.E.; Breuer, J. and Voronin Y. (2023). Sharing Social Media Data: The Role of Past Experiences, Attitudes, Norms, and Perceived Behavioral Control. *Front. Big Data*, 5: 971974.



- Article 29 Working Group (2014). *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC* (= WP 217). [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf) (last access: 28.07.2023).
- Beurskens, M. (2014). Legal Questions of Twitter Research. In K. Weller; A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (Eds.), *Twitter and Society*. New York: Peter Lang, pp. 123--133.
- Brehm, A.S. and Lee, C.D. (2015). Click Here to Accept the Terms of Service. *Publication of the Forum Committee on Communications Law, American Bar Association*, 31(1), pp. 4--7.
- DFG (Deutsche Forschungsgemeinschaft) (2015). *Leitlinien zum Umgang mit Forschungsdaten*. [https://www.dfg.de/download/pdf/foerderung/grundlagen\\_dfg\\_foerderung/forschungsdaten/leitlinien\\_forschungsdaten.pdf](https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/leitlinien_forschungsdaten.pdf) (last access 05.03.2023).
- Clarke, I. and Grieve J. (2019). Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018. *PLoS ONE*, 14(9), e0222062.
- de Souza, C.R.B.; Redmiles D.; Cheng, L.-T.; Millen, D. and Patterson J. (2004). Sometimes you Need to See through Walls: A Field Study of Application Programming Interfaces. *Proceedings of the 2004 ACM conference on Computer supported cooperative work (CSCW '04)*. Association for Computing Machinery, New York, NY, USA, pp. 63--71.
- Documenting the Now (2020). Hydrator, Computer Software. <https://github.com/docnow/hydrator> pdf (last access 05.03.2023).
- EDMO (2022). Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. 31 May 2022. <https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf?ref=static.internetfreedom.in> (last access 24.07.2023).
- Forgey, Q. (2021). National Archives can't Resurrect Trump's Tweets, Twitter Says. *POLITICO*, 04. July 2021. <https://www.politico.com/news/2021/04/07/twitter-national-archives-realdonaldtrump-479743> (last access: 24.07.2023).
- JustAnotherArchivist (2020). snsrape: A social networking service scraper in Python, v0.3.4, Computer Software. <https://github.com/JustAnotherArchivist/snsrape> (last access: 24.04.2023).
- Kaczmirek, L.; Mayr, P.; Vatraru R.; Bleier, A., Blumenberg, M.S.; Gummer, T.; Hussain, A.; Kinder-Kurlanda, K.; Manshaei, K.; Thamm, M.; Weller, K.; Wenz, A. and Wolf C. (2014). Social Media Monitoring of the Campaigns for the 2013 German Bundestag Elections on Facebook and Twitter. *GESIS Working Papers* 31. Cologne: GESIS. Available online at <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-381955> (last access: 24.04.2023).
- Kreutzer, T., and Lahmann, H. (2021). *Rechtsfragen bei Open Science: Ein Leitfaden*. Hamburg University Press. <https://doi.org/10.15460/HUP.211> (last access 24.07.2023).
- Krotov, V.; Johnson, L. and Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, 47, pp. 539--566.
- Martin, E.A. (2003). *A Dictionary of Law*. Oxford: Oxford University Press.
- Ott, B.L. (2017). The Age of Twitter: Donald J. Trump and the Politics of Debasement. *Critical Studies in Media Communication*, 34(1), pp. 59--68.
- Schallaböck, J.; Kreutzer, T.; Hoffstätter, U. and Buck, D. (2023). Mustervertrag Datenaufnahme KonsortSWD (1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.7648898> (last access 24.07.2023).
- Schneider, Ulrike. (2021). How Trump Tweets: A Comparative Analysis of Tweets by US Politicians. *Research in Corpus Linguistics* 9(2), pp. 34--63.
- Social Media Research Group (2016). Using Social Media for Social Research: An Introduction. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/524750/GSR\\_Social\\_Media\\_Research\\_Guidance\\_-\\_Using\\_social\\_media\\_for\\_social\\_research.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf) (last access 24.07.2023).
- Twitter (2023). Previous Terms of Service. <https://twitter.com/de/tos/previous> (last access: 24.04.2023).
- Twitter Help Center. What does the Blue Checkmark Mean? Available online at <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts> (last access 03.05.2023).
- Vogel, P. and Hilgendorf E. (2020). Big Data in Social, Behavioural, and Economic Sciences: Data Access and Research Data Management. *RatSWD Output* 4(6). Berlin. <https://doi.org/10.17620/02671.52> (last access: 24.04.2023).
- Walker, S. (2017). The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts, Dissertation University of Washington. <http://hdl.handle.net/1773/40612> (last access: 27.04.2023).
- Watteler, O. (2022). GDPR and Research Data in the European Union. In A. Dreiser & C. Samimi (Eds.), *Frontiers in African Digital Research: Conference Proceedings*, University of Bayreuth African Studies Online 9, pp. 99--128. Bayreuth: Universität Bayreuth.
- Weller, K. and Kinder-Kurlanda, K.E. (2016). A Manifesto for Data Sharing in Social Media Research. *Proceedings of the 8th ACM Conference on Web Science (WebSci '16)*. Association for Computing Machinery, New York, NY, USA, pp. 166--172. <https://doi.org/10.1145/2908131.2908172> (last access 24.07.2023).
- Williams, M.L.; Burnap, P. and Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, 51(6), pp. 1149--1168.

# Collecting Health Memes for a Subcorpus of Peer Health Discourse

Laurel Smith Stvan

University of Texas at Arlington

E-mail: stvan@uta.edu

## Abstract

A discussion of a social media subcomponent of CADOH, the Corpus of American Discourses on Health, a corpus focusing on how health information in English is conveyed among non-specialists. Previous online data in the corpus is now supplemented by a collection of 200 food and health-focused internet memes collected between 2018 and 2022 using Google image searches. Query terms were chosen to align with the other corpus components: *calories*, *cold*, *Covid*, *cholesterol*, *colonoscopy*, *diet*, *fat*, *flu*, *health*, *nutrition*, *shots*, *vaccination*, *germs*, and *sanitizer*. Metadata was collected to track the URL, author, posting date, wording of the caption, keywords of the topic, topic of the image macro, and a jpeg of each meme. The paper discusses the benefits of thematic dataset collection. Applications are suggested for linguistic analyses such as contrasting the performance of online speech acts and for public health investigations into lay beliefs about causality and health outcomes.

**Keywords:** memes, health communication, speech acts, corpus linguistics

## 1. Introduction

This paper discusses a social media component of CADOH, the Corpus of American Discourses on Health. The corpus is a work in progress, but as an overview of its goals I note that while much corpus-linguistic work on health focuses on the language of medical providers or medical researchers (Atkinson & Valle, 2012 *inter alia*), this corpus focuses instead on examining how health information is conveyed among non-specialists. The materials, therefore, share a theme of health and nutrition, but are based on lay or vernacular level discussions.

The current corpus of approximately 340,000 words contains written texts (including newspaper and magazines articles), spoken dialog transcripts (from fictional excerpts, TV and radio broadcast panels, and multi-speaker student focus group discussions) and genres of the informal written language often found online (cf. McCulloch, 2019), including blog post and their comments, and asynchronous forum postings. This online data section is now supplemented by a collection of food and health-focused internet memes.

## 2. Memes as Vernacular Data

The genre of the internet meme fits this corpus's goals well because of the following meme characteristics: a) they use an informal, vernacular register; b) they are used in peer-to-peer communication; c) they occur in a wide set of social media outlets; d) they are fast spreading in their transmission; and e) they have uses that can be both general in topic but can also tightly reflect the time and setting of their creation. This means that they are often culturally specific, reflecting, for example, the local pop culture, customs, and politics of a users' speech community.

## 3. Previous Analyses of Memes

As CMC corpus components, two aspects of memes that have been noted in other studies are relevant here: their

structural components and their use as tokens of topic-specific material.

### 3.1 Meme Structure

Work in the past ten years has emerged on the specific format of the image-based internet meme, including their structural characteristics and the range of places that they occur. Memes show a contrast to the more conversational structure of other online exchanges. Structurally, a meme can be used as “reaction image” to respond to a previous post, similar to the placement of moving gifs. But a meme is often a single-move conversational contribution which can stand alone. For example, on meme generating sites they may simply receive up or down votes from the public. In closed SMS conversations, too, they may receive only tapback reactions. Memes can capture a response to a general social trend or can be offered as an individual's opinion. The current work discusses reasons for collecting and annotating a range of memes on one specified topic. Building on this, future work is planned to look more fully at the real-life contextual uses, as either isolated jokes or as ways to initiate or cap off topics in conversations, i.e., at how they might be used with regards to adjacency pairs.

Because meme posts can be a solo move, they don't immediately follow the shape of other instances of CMC, which Beißwenger & Lungen (2020) describe collectively as a “dialogic, sequentially organised interchange between humans.” But while they are often a single move, the meme unit itself can be made of one, two, and even up to five panels, as in the vertical array of captioned arguments laid out around the American Chopper characters, for example.

In describing the occurrence of memes, Shifman (2014:341) observes that “the unique features of the internet turned the diffusion of memes into a ubiquitous, highly visible, and global routine.” Phillips and Milner note that memes “depend on multimodality... including written words and static images, as well as audio and video”, and that they

offer “the remix and recombination of existing cultural materials,” and rely on “strong personal affinity,” “social creation and transformation,” and circulation through mass networks” (2017:31). That is, the potential trail of the creation of memes entails that they are not purely single voiced. For example, Dynel (2021) details the multiple voices leading to each repost—which involves the subject of the photo, the original photo taker, a later captioner, the meme poster, and many re-posters. For the current project, I am focusing on the meme as a finished combination, which will allow investigation of what aspects a re-poster could be taken as intending to signal by using it.

### 3.1 Thematic Collections of Memes

There are also precedents for collecting thematic datasets of memes, used in more social, anthropological, or discursive investigations of natural online exchanges as opposed to the computational or NLP approaches (such as Wang & Wen, 2015) which aim at automating the generation of memes.

Thus, researchers working in cultural studies have focused on datasets of memes on particular topics. For example, work by Dynel (2021) on memes related to Covid masking, and work by Malik & Tehseen Zahra (2022) on memes depicting responses to distant learning. In collections with broader ranging topics, Piata (2020) examines classical art memes, and Gasparini et al. (2022) worked on transcriptions of a dataset of memes identified as having misogynistic content. In short, memes are starting to be seen as ways that subject-based content is shared and discussed within different disciplines.

## 4. Uses in Linguistic Analysis

There are, of course, uses of memes that would be of interest to linguistic analysis. I will demonstrate with two of my own research interests in pragmatics. One is as a way to investigate the use of memes as speech acts. And another topic is the ways memes can convey ideas about discussing causality. Both studies enable a contrast with how these functions are expressed in other types of discourse.

### 4.1 Speech Acts

For example, as uses of indirect speech acts of advice, below are two meme tokens which I suggest are used to give guidance about health. In Figure 1, the post advises that the quantity, not just the ingredients, affect what one should eat. And in Figure 2, a two-part image that is verbally lacking any syntax, yet is interpreted as guidance. In this case, suggesting the ineffectiveness of essential oils in protecting against catching diseases. The bottom panel shows a fragile Cheeto which serves to latch a door against the metaphorical invaders above.

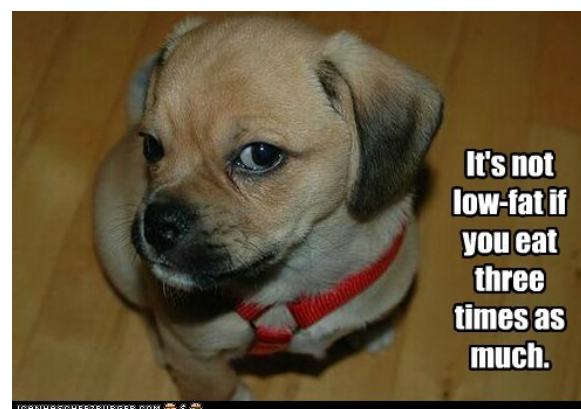


Figure 1: Advice about dieting.



Figure 2: Advice about essential oils.

In sentence structure and lay out, Figures 1 and 2 are quite different. Both rely on not only the caption’s wording, but also on the images and camera angles to present the force of directives.

### 4.2 Causality

Another linguistic analysis of the functions of memes explores how the message of the meme endorses or denies causality, often through contradicting or using the myth-busting of common health beliefs. These are illustrated by three myth busting memes in figures 3-5:



Figure 3: Myth busting about wet hair.



Figure 4: Myth busting about eating fruit.



Figure 5: Myth busting about eating fat.

Figure 3 suggests evidence to counter a prevalent myth about how going outside with wet hair leads to catching a cold. Figure 4 suggests that not all foods containing sugar should be considered unhealthy. And Figure 5 emphatically clarifies that eating fat is not what makes one fat.

## 5. Data Sources and Methods

In seeking memes on health-related topics, I used two sources. A first round of the health meme dataset was collected in the fall of 2018. I used a Google image search, which involved a text query including the word *meme*. Later I also obtained access to the Reddit Meme Dataset, which contained over 3000 memes from 2018. In this case I could do a text search of the title field of each entry. Neither of these methods, however, provided access to the captions of the memes, or a guarantee of the image content, requiring the material to be manually vetted. This restriction showed up in other existing sources of meme collections. The American Folklife Center and the Library of Congress, for example, created a site called “Meme Generator: collected datasets” where users can “create and share image macros” which also serves “as a searchable collection of user-created images” (Library of Congress, 2018). Their material was first gathered in May of 2018 from crawling the Library of Congress's Web Cultures Web Archive, creating a dataset of 57,652 unique memes. The Meme Generator site is searchable by choosing from a list of image macros named in the side bar (Willy Wonka, Success Kid, etc.), so for comparing uses of a certain macro, this could provide a more controlled search than Google image queries. But once again, searches match the name of the macro, rather than the words in the caption.

With the arrival of Covid in 2020, I added pandemic-related memes to my Google image searches, ending up with a total of 200 memes in this pilot set. For consistency, I gathered only two image types, those which Shifman (2014) categorizes as “reaction Photoshops” and “stock character image macros,” leaving out the more eclectic range of annotated screenshots, still-life food photos, medical illustrations, charts, and lists that also can be found as memes. Based on the topics from the other genres in the CADOH corpus, I initially used the search terms *calories*, *cold*, *cholesterol*, *colonoscopy*, *diet*, *fat*, *flu*, *health*, *nutrition*, *shots*, *vaccination*, *vax*, *germs*, and *sanitizer*. I later added variants of *Covid-19*, and *mask* as well as uses with the purposely mis-spelled variant of *helth*, which is used in a popular meme template. A version of the latter is seen in Figure 6, which shows the character Meme Man, or Stonks, who is described as “a reaction image to joke about making poor health decisions” (Know Your Meme 2007).





Figure 6: A variant of the Stonks meme template.

When my undergraduate research assistant and I began collecting the items in 2018, we found it useful to spell out in the headers of our shared spreadsheet the types of metadata to track for each meme. These included a) the item number, b) the listed author or copyright holder, c) the date it was posted, d) the exact words of the caption, e) keywords of the topic, f) who or what is in the image—with assistance from the Know Your Meme: Internet Meme Database (2007), g) the item's URL, and h) a thumbnail jpeg of each meme. Later the jpeg's dimensions were also tracked. These categories will form the basis of the markup in an XML file using Beißwenger & Lungen (2020)'s proposed components of CMC interactions as made up of utterances, posts (including written and multimodal pieces), and nonverbal activities.

For the purposes of compiling and annotating a corpus, it was valuable to note each meme creator, if they were listed. But since the use cycle of a meme involves sharing, any number of people could re-post each meme. Thus, when examining later actual online conversations that contained memes, researchers might choose to anonymize the posters and their interlocutors.

## 6. Potential Analyses

As applications, I will first point out two areas of qualitative exploration that could use this form of health discourse data. One relies on how memes are not just multimodal, but can be found with the same image macros appearing with a variety of captions. The mix-and-match aspects invite some teasing apart as to their separate influences on the overall message. This offers a way to check for the ability of the text and picture to either contradict or reinforce an advice giver's stance.

A related aspect is the interpretation of memes as conveyors of backgrounded, common-sense information, representing an authoritative voice—due in part to the prevalence of memes, but also due to the familiarity of the characters

portrayed. However, while memes can be presented as authoritative, they feature both pros and cons of divisive topics (such as vaccinations, masks, diet foods, causes of catching a cold, etc.). So, exploration is called for in tracking the rhetorical effects that occur when the same meme ingredients are used to argue for one side or another of divisive topics.

But beyond context-independent analyses of the meme form, a larger pragmatic analysis would be able to examine how different forms are responded to in their posted contexts—intersecting with studies on texting that look at responses such as tapbacks, as well as larger conversational follow-up moves of agreement, neutrality, or disagreement.

Themed data sets have potential applications outside of linguistics as well. For example, for the type of memes I collected, I suggest a use by public health workers who might monitor this modality to observe which health issues people are aware of. It would be possible to use meme information to track the advice tactics that are used by lay people online, as a way to gauge a community's understanding of expected health outcomes. That is, what do peers tell each other to do? Or to avoid doing? How does this align with other sources of health information? And how might that advice change as we move through various epidemics?

## 7. Conclusions

I have shown the rationale for gathering meme components as part of a CMC corpus and detailed the issues encountered in this task. The goal was to capture examples of an under-analyzed vernacular form of online interaction. Because these posts communicate through multimodal signals, such a dataset enables comparison with speaker positioning used on the same topic in other text-based forms of CMC. Metadata issues for identifying the corpus components concerned distilling text captions, image descriptions and dimensions, as well as any author name, creation date, and URL. When gathered by theme, a meme dataset also allows qualitative drilling down into discourse practices in discipline-specific topics. In short, this paper describes finding the data to build the health discourse dataset and spells out the methodology for annotating and storing it as a corpus, thus paving the way for other works intending to use the meme content as a basis for questions explored in public health research or linguistic analysis.

## 8. References

- Atkinson, D. and Valle, E. (2012). Corpus analysis of scientific and medical writing across time. *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: John Wiley & Sons.
- Beißwenger, M. and Lungen H. 2020. CMC-core: a schema for the representation of CMC corpora in TEI. *Corpus*, 20. <https://doi.org/10.4000/corpus.4553>.
- Dynel, M. (2021). COVID-19 memes going viral: On the multiple multimodal voices behind face masks. *Discourse & Society*, 32(2), pp. 175--195.

- Gasparini, F; Rizzi, G; Saibene, A. and Fersini, E. (2022). Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44, pp. 1--8.
- Goswami, S. (2018). Reddit Memes Dataset. <https://www.kaggle.com/datasets/sayangoswami/reddit-memes-dataset>.
- Highfield, T. and Leaver, T. (2016). Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*, 2(1), pp. 47--62.
- Know Your Meme: Internet Meme Database. (2007). <https://knowyourmeme.com/>.
- Library Of Congress, Collector, and Sponsor American Folklife Center. *Meme generator: Collected datasets*. (2018). <https://www.loc.gov/item/2018655320/>.
- Malik, A. and Zahra T. (2022). Pragmatic analysis of internet memes on distant learning. *Pakistan Journal of Society, Education and Language*, 8(2), pp. 303--317.
- McCulloch, G. (2019). *Because Internet: Understanding the New Rules of Language*. New York: Riverhead Books.
- Phillips, W. and Milner R. M. (2017). *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online*. Cambridge, UK: Polity Press.
- Piata, A. (2020). Stylistic humor across modalities: The case of classical art memes. *Internet Pragmatics*, 3(2), pp. 174--201.
- Shifman, L. (2014). The cultural logic of photo-based meme genres. *Journal of Visual Culture*. 13(3), pp. 340--358.
- Wan, Y; Liu, X. and Chen, Y. (2011). Online image classifier learning for Google image search improvement. In *2011 IEEE International Conference on Information and Automation*, pp. 103--110. IEEE.
- Wang, W. Y. and Wen, M. (2015). I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Red Hook NY: Curran Associates, pp. 355--365.

# Specific behaviours in Wikipedia talk pages: some insights from extreme cases

Ludovic Tanguy<sup>(1)</sup>, Céline Poudat<sup>(2)</sup>, Lydia-Mai Ho-Dac<sup>(1)</sup>

(1) CLLE: CNRS & University of Toulouse, France

(2) BCL: CNRS & University of Nice Côte d'Azur, France

Email: [ludovic.tanguy@univ-tlse2.fr](mailto:ludovic.tanguy@univ-tlse2.fr), [celine.poudat@univ-cotedazur.fr](mailto:celine.poudat@univ-cotedazur.fr), [lydia-mai.ho-dac@univ-tlse2.fr](mailto:lydia-mai.ho-dac@univ-tlse2.fr)

## Abstract

Based on a dataset of 3.4 million threads from English Wikipedia talk pages, we specifically focus on extreme cases. We propose a qualitative analysis of the most prolific message authors, the longest threads in terms of messages, contributors and durations, as well as the longest monologues (single-user threads). These case studies allow us to identify a number of behaviours that can significantly differ from the typical discussions between Wikipedians. If some threads do not have a real dialogic status (polls, monologues, logbooks and diaries), some of them push online communication to its limits across time. These sometimes unexpected behaviours can help us get a more precise understanding of this unique source of computer-mediated communication data.

**Keywords:** Wikipedia talk pages, online interaction, extreme behaviours

## 1. Introduction

The study presented in this paper is part of a larger project that explores various dimensions of Wikipedia talk pages. Talk pages have been extensively studied as they provide a unique means to examine the dynamics of interaction between Wikipedians (Laniado et al. 2011). They also serve as a valuable source of computer-mediated communication data which is abundant, multilingual and freely accessible, making them suitable for large-scale studies on generic online interactions (Gomez et al. 2011, Lügen & Herzberg 2019). The main practices in Wikipedia talk pages have already been studied and described with a focus on the topics discussed (Schneider et al. 2010) or local interaction patterns (Kopf 2022).

The case study presented here focuses on marginal, or even extreme behaviours in Wikipedia talk pages. We have selected a number of outlier cases that exhibit unexpected characteristics at the thread or user levels. These include highly prolific users, excessively long threads (in terms of duration, number of posts or users involved) and monologues. We assume that the analysis of such extreme cases can help to better understand expected and unexpected interactions between Wikipedians. This will also allow us to highlight practices which are generally neglected although they may be found in more typical configurations.

## 2. Dataset: English Wikipedia talk pages

We base our study on a dataset, which consists of threads extracted from the August 2019 dump of Wikipedia. At that time the English version of Wikipedia contained 14,856,106 article pages and 7,903,148 talk pages, including archives. Among these, only 2,025,888 contained at least one posting with at least 2 words.

It is worth noting that talk pages on Wikipedia are produced on the same infrastructure as the articles, using

*wikicode* formatting. This means that a talk page is fully editable by any user and that its layout and organisation can be freely modified, in spite of strong recommendations from the Wikipedia community. Talk pages typically feature a section-based structure, with each section representing a distinct discussion having its own heading and clear boundaries. Individual messages are organised along a tree structure which follows the example of the more traditional online discussion platforms. However, the *wikicode* allows freeform editing which may lead to unusual structures in discussion threads, such as the re-sectioning of existing talk pages (used for archival purposes for example), the writing of non-contiguous answers to a previous long message (similar to emails), or postings appearing in a non-chronological order. This situation has dire consequences on the parsing of Wikipedia talk pages, which requires additional efforts to identify the network of interactions.

Despite these challenges, we segmented each talk page into sections, with each section representing a thread. Each thread was segmented into posts (or comments or messages) following an heuristic based on signatures and indentations. The whole structure was then converted into XML format following the TEI-CMC guidelines, so that each post is associated with its author's name and date. Finally, threads containing a post written by a bot were discarded. In the end our corpus contains 3,385,583 threads and 8,873,620 messages (Ho-Dac, to appear).

Table 1 gives an overview of the dataset characteristics that were considered relevant for identifying extreme behaviours. The large differences between means and medians suggest highly skewed distributions with numerous outliers for each variable. In the following section we focus on the outlier cases corresponding to the highest values for each variable in the table.



Feature	Maximum	Median	Mean
Number of posts per user	25,078	1	20.06
Number of posts per thread	651	1	2.62
Number of users involved	97	1	1.85
Duration of threads with 2 or more posts (N=1,688,939)	16.6 years	5.3 days	260 days
Longest duration between 2 posts in the thread	16.1 years	4.1 days	233 days
Number of posts per single user thread (N=1,812,457)	150	1	1.08

Table 1: Overview of features used to identify extreme behaviours

### 3. Extreme behaviours

While identifying outliers is a common initial step in data analysis, its primary objective is to remove atypical individuals which can skew the study of the central tendencies. Here, although we initially targeted outliers in order to exclude them from the dataset and facilitate discourse analysis studies, the qualitative analysis of these outliers allows us to identify behaviours that are made possible by the Wikipedia device, and that may even be typical of Wikipedia interactions.

#### 3.1 The most prolific message authors

Our first investigation targets Wikipedia users who have produced a significant number of posts on talk pages. In our dataset, we found a total of 499,137 different usernames in the signatures of all talk pages (without including the bots or the unregistered users who are only identified by their IP addresses). As expected, the number of posts per user follows a Zipfian distribution, meaning that while a majority of users have written a single comment, a few Wikipedians are the authors of a very large number of messages. The user ranking first posted 25,078 messages, the user ranking #10 14,281, and the user ranking #100 5,900.

To compare message-posting behaviour with actual Wikipedia editing activity, we gathered data on the number of edits (i.e. the modifications made on any page of the Wikipedia, including posts in any kind of talk page) and the number of posts in the article talk pages for the 1000 most productive Wikipedia editors as indicated in the official leaderboard<sup>1</sup> (as of July 2019), shown in Figure 1. We measured a weak positive correlation ( $\rho=0.09$ ) between the number of edits and the number of messages. As an example, the most active editor of the English Wikipedia (Steven Pruitt, who was responsible for more than 3 million edits in 2019, and over 5 million in 2023) has never participated in a discussion in any article talk page (although he did post some messages in a few users' personal talk pages, not included in our dataset).

<sup>1</sup> <https://en.wikipedia.org/w/index.php?title=Wikipedia:WBE>

Similarly, several of the most prolific authors on the articles talk pages rarely modify the articles themselves, limiting their role to commenting or proofreading the text written by others, or to enforcing Wikipedia policy and rules through discussion.

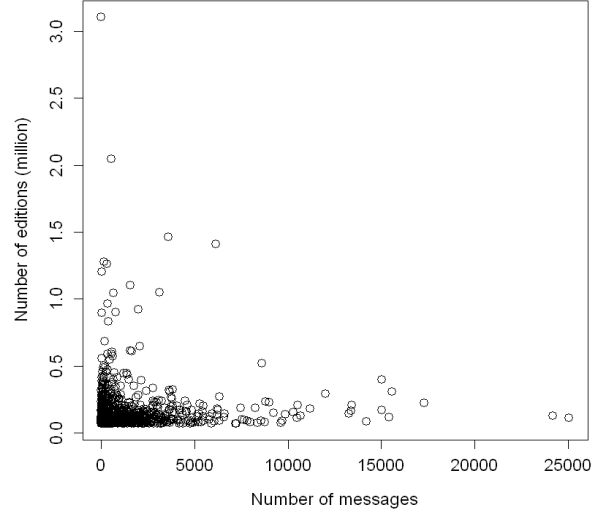


Figure 1: Number of editions versus number of messages for the 1000 most productive Wikipedia editors

These first observations would clearly show that taking part in a Wikipedia discussion can to some extent be considered as a specific activity, decorrelated from article writing, at least for a subset of the Wikipedia users.

#### 3.2 Threads with the highest numbers of posts

The second phenomenon we investigated is the number of posts per thread. If 53% of the threads consist of a single post, some of them contain several hundred posts.

We examined the 100 longest threads in our dataset (threads with more than 90 posts, up to 651). Surprisingly, these very long threads rarely imply a large number of participants (median of 14 different users) and they can even be written by a single user (this particular category is examined more closely in §3.5).

If we only consider their organisation and structure, these long threads can be classified as follows:

- 68 of the 100 examined threads can be qualified as *standard discussions*. Indeed, these threads follow the conventional organisation where users exchange their views and arguments, following a tree-like structure where the replies and reactions to previous posts are indicated through cumulative indentations. However, due to the extensive size and depth of the threads, indentation can hinder their readability. To address this, some users (most of the time participants to the discussion) sometimes use the flexibility of the talk pages (based on the same wikicode used for article pages) to organise them into sections. When appropriate, subtopics can be identified and used to start a new nested thread in a subsection, while remaining in the same section and therefore related to

the same topic. When not, *arbitrary breaks* are introduced to reset the indent level when it becomes too deep<sup>2</sup>.

- 26 of them are *polls* or *series of polls*. In these threads a user collects the position or opinion of others on specific topics. As such, every single vote by the polled users counts for a message. The length of these threads can be attributed to the high number of participants (up to 97), multiple related polls grouped together (with the same users posting a message for each subtopic), or one or more nested threads developing inside the poll. For example, a user may explain his or her position, eliciting reactions from others. These threads are further described in §3.3.
- 6 are long *lists*, the items of which are expressed as separate messages, and are initially posted by the same user. As these threads only marginally contain posts by different users we study in more detail this specific type of thread in §3.5.

To summarise, our findings indicate that only two thirds of the 100 longest threads can be classified as discussions, highlighting the diverse uses of talk pages.

### 3.3 Threads with the most users

The 100 threads with the highest numbers of different participants are all polls or series of polls. Polls are a common practice in Wikipedia talk pages as they represent the pursuit of consensus (Kopf 2022). Polls can cover various decisions related to the article page, such as article deletion, merging with another related article, changing the article's title, deleting a whole section, choosing between different pictures etc. These polls may be created after inconclusive discussions or as a first intent when dealing with a new issue. The questions asked can be binary (support/oppose a suggestion) or open-ended (propose a new title, picture etc.). As we focus here on the number of different users, our sample is limited to threads with a single poll.

Due to the flexibility of the underlying wikicode, polls may be organised in two different ways. Messages can be in chronological order, with each user expressing her opinion in sequences. Alternatively, messages can be grouped based on their position, so that all messages, users and arguments in support or opposing the initial proposition are in the same section.<sup>3</sup>

Some of the polls are both spontaneous and local, and can be organised inside a discussion: they are qualified as straw polls. Others are qualified as Request for Comments (RfC) and follow a more sophisticated organisation. RfC polls are indexed in the Wikipedia space and therefore receive much more attention. This increased attention can lead to some problems when high stakes motivate certain

users to manipulate the voting process with additional or fake accounts (*puppetry*), leading to their abandonment.<sup>4</sup> Several of our most massive threads show such cases that are explicitly flagged, but all expressed votes and comments remain available.

### 3.4 Longest-lasting threads

The temporal dynamics of Wikipedia discussions has been studied in (Kaltbrunner & Laniado 2012) but, as seen in Table 1, some threads can last more than 15 years, nearly the timespan of our dataset. In 2019, the 100 longest-lasting threads covered a duration of over 14.5 years. 8 of the threads we examined are false positives: the prolonged duration is merely a consequence of some messages being placed in a generic section of the talk page (labelled as “*Comments*” or similar). Therefore the messages simply do not constitute a discussion; but the 92 other cases are clear instances of communicating occurring over an extended period of time.

About 10% of these threads exhibit a continuous spread over a significant period, with regular postings and no extended periods of silence exceeding a couple of years. However, the majority of threads demonstrate a single notable jump across time, with a message being posted in response to a comment made over a decade ago, such as the example in Figure 2.

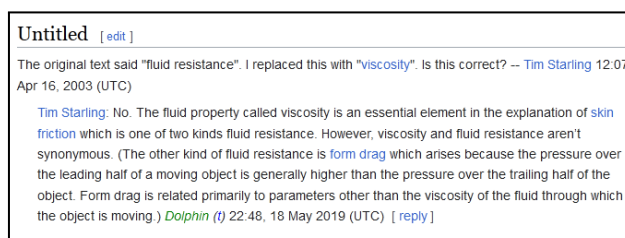


Figure 2: sample thread with a 16-year gap

[https://en.wikipedia.org/wiki/Talk:Charles-Augustin\\_de\\_Coulomb#Untitled](https://en.wikipedia.org/wiki/Talk:Charles-Augustin_de_Coulomb#Untitled)

Surprisingly, most of these dialogues (72) contain no explicit mention of the temporal specificity. Users write their comments as if the message they are reacting to was posted just a few minutes ago. A wide range of dialogue acts can be observed in such situations: answering a simple factual question (as in Figure 2), providing a reference, commenting on a statement<sup>5</sup>, etc. In a few of these cases however we found that the answerer addresses the author of the first message in the third person, which may seem unusual in online communications (“Related to why that was put by *an earlier editor*, the reason is [...]”<sup>6</sup>, “I have to wonder what *this IP user* imagined [...]”). This may indicate that the more recent author acknowledges the fact that his interlocutor has long departed from the talk page and that the response is directed toward present

<sup>2</sup> [https://en.wikipedia.org/wiki/Talk:Gamergate\\_\(harassment\\_campaign\)/Archive\\_12#KotakuInAction\\_moderators\\_misogynist/anti-feminist/interested\\_in\\_female\\_subjugation\\_porn](https://en.wikipedia.org/wiki/Talk:Gamergate_(harassment_campaign)/Archive_12#KotakuInAction_moderators_misogynist/anti-feminist/interested_in_female_subjugation_porn)

<sup>3</sup> [https://en.wikipedia.org/wiki/Talk:Campaign\\_for\\_the\\_neologism\\_%22sanctum%22/Archive\\_6#Proposal\\_to\\_rename,\\_redirect,\\_and\\_merge\\_content](https://en.wikipedia.org/wiki/Talk:Campaign_for_the_neologism_%22sanctum%22/Archive_6#Proposal_to_rename,_redirect,_and_merge_content)

<sup>4</sup> [https://en.wikipedia.org/wiki/Talk:K.\\_P.\\_Yohannan#Keeping\\_the\\_controversy\\_section\\_in\\_this\\_article](https://en.wikipedia.org/wiki/Talk:K._P._Yohannan#Keeping_the_controversy_section_in_this_article)

<sup>5</sup> <https://en.wikipedia.org/wiki/Talk:T-shirt#Capitalisation>

<sup>6</sup> <https://en.wikipedia.org/wiki/Talk:Brondesbury#Place>

and future readers. But this particular behaviour has to be studied more precisely; Herzberg & Lügen (to appear) studied the different ways a user addresses the author of a previous message, and found that a second person address occurs in less than 30% of replies.

If the late response is sometimes justified by a change in the world or an advancement of knowledge, it can also deal with atemporal topics. All these efforts to provide answers and additional information across time, even in the absence of the original participant, reflects the global dynamics and objective of the Wikipedia project.

In the remaining cases, users also take advantage of the flexibility of Wikipedia talk pages. Some users explicitly modify the timestamp of their message, pre-dating them to several years in the future to prevent their automatic archival. This is a move similar but somewhat more drastic to “bumping” a thread in online forums (i.e. adding empty messages to an existing thread to keep it visible).

In two cases we found what can be qualified as *talk page archaeology* (see example in Figure 3). A user re-posts an old message or discussion that had been deleted or lost in the restructuring of Wikipedia. The reason for this is apparently not to answer the initial question or to correct a statement, but simply to preserve a trace from previous efforts. This preservative attitude has even led to keeping the very first versions of Wikipedia accessible in *Nostalgia Wikipedia*<sup>7</sup>.

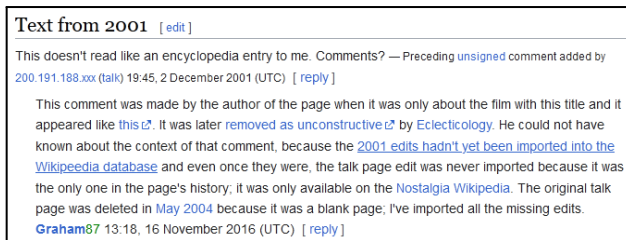


Figure 3: sample thread restoring a previous comment  
[https://en.wikipedia.org/wiki/Talk:Casablanca#Text\\_from\\_2001](https://en.wikipedia.org/wiki/Talk:Casablanca#Text_from_2001)

Although these temporal behaviours have not been formally described before, they confirm the specific position of the Wikipedia project as a global memory as expressed by Pentzold et al. (2017).

### 3.5 Longest single-user threads

Our last study focuses on single-user threads. In our dataset, 53% of all threads are authored by a single user, primarily due to them consisting of a single post. However, 6.9% of threads with 2 or more posts are written entirely by a single user. These “monologues” can grow to be quite extensive, reaching up to 150 messages. Similar to our previous analyses, we examined the 100 longest single-user threads (with 12 or more posts) and identified two main configurations.

A significant majority of these threads (88) are *lists*, as we

had observed in some of the longest threads (§3.2). The messages within these threads can take the form of paragraphs that include comments, remarks or suggestions<sup>8</sup>. These cases typically result from a review of the article, or a series of proposals and suggestions for rewriting or expanding it. Of course, these items can sometimes receive comments or extensions in the form of nested messages by other users as noted in §3.3.

But long lists of another kind contain only simple informational elements relevant to the article, such as products, dates, characters, users... In most cases, the thread lacks an explicit communication goal and appears to function as a logbook or to-do list for the author. A thread of such “grocery list” type can include check marks or crossed out items, indicating that they have been processed (e.g. proofread, referenced, integrated into the article...). In only 12 cases of such lists we could find explicit invitations from the author to others to contribute by extending, commenting or correcting the items, although in our sample these remained unanswered.. Figure 4 shows such an explicit checklist with the author giving potential helping hands precise instructions.

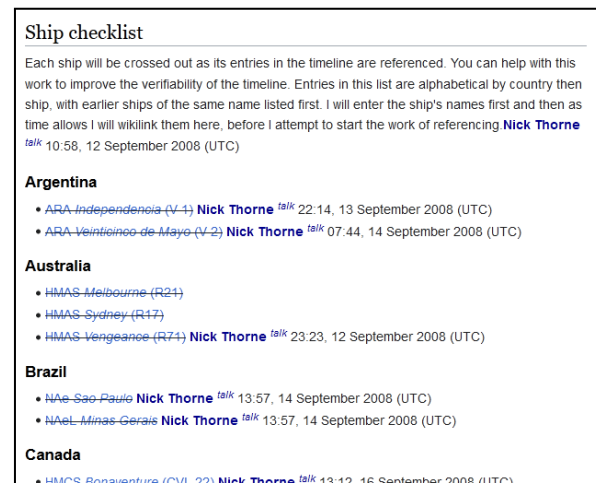


Figure 4: sample list thread (extract)  
[https://en.wikipedia.org/wiki/Talk:Timeline\\_for\\_aircraft\\_carrier\\_service/Archive\\_1#Ship\\_checklist](https://en.wikipedia.org/wiki/Talk:Timeline_for_aircraft_carrier_service/Archive_1#Ship_checklist)

The 12 remaining long monologues contain heterogeneous posts, which can consist of larger text segments such as problem analyses, reviews, suggestions, hypotheses, reports of actions taken, steps in an investigation and more, to various combinations of such messages within the same thread<sup>9</sup>. In all cases these monologues lack explicit indicators of dialogue such as the use of second-person pronouns or explicit calls for reactions. Instead, they can be considered as some kind of diary, following a Wikipedian’s work and thoughts on a topic, spread out over time.

<sup>7</sup> <https://nostalgia.wikipedia.org/>

<sup>8</sup> [https://en.wikipedia.org/wiki/Talk:Timeline\\_of\\_the\\_Irish\\_War\\_of\\_Independence#Doubtful\\_edits](https://en.wikipedia.org/wiki/Talk:Timeline_of_the_Irish_War_of_Independence#Doubtful_edits)

<sup>9</sup> [https://en.wikipedia.org/wiki/Talk:CMB%20cold%20spot#Professor\\_Mersini\\_Radio\\_Broadcast](https://en.wikipedia.org/wiki/Talk:CMB%20cold%20spot#Professor_Mersini_Radio_Broadcast)

#### 4. Conclusion

Our study of the outlier threads in a dataset of over 3 million discussions from the English Wikipedia talk pages has allowed us to identify several specific behaviours.

The flexibility of the platform plays a crucial role in enabling these behaviours, as users can reshape and reorganise the posts in ways which are not possible in the other online discussion environments. The ability users have to freely (re-)order messages in a thread facilitates the emergence of new forms such as organised polls, sectioned long threads and the use of threads as checklists. In some cases, these possibilities may induce a shift away from the central communicational goal of the talk pages, such as monologues and threads used as log books or diaries. However, interaction remains possible even in these cases.

Our observations of long-lasting discussions confirm the objective of the Wikipedia project to create a cultural monument and testament. Talk pages, as the main articles of the encyclopaedia, are considered permanent documents. Therefore, it is not a problem for a Wikipedian to respond to a message even 15 years later, with the response being primarily directed towards the community rather than the original user.

It was not our aim to investigate the specific topics or domains in which certain types of discussion take place. During our observations we did not identify any particular area of knowledge that would correlate with specific behaviours. However, it is evident that popular topics such as pop culture, sports and geopolitics tend to attract a larger number of participants. Nevertheless, impressive efforts to gather information from a single individual can be found across various subjects, including niche areas.

On the methodological front, our approach needs further completion by exploring the extent to which these phenomena appear in less extreme cases. Preliminary surveys have shown, for instance, that polls and single-author lists appear at much smaller scales (2-3 voters, a few items in a list, short monologues) and, therefore, occur more frequently.

This naturally calls for further investigations, including a more systematic corpus search of local configurations in order to estimate the frequency of these behaviours, and to enable cross-lingual comparisons. It should be noted, however, that Wikipedia talk pages cannot be regarded as typical CMC data without taking these specificities into account.

#### 5. References

- Gómez, V., Kappen, H. J., & Kaltenbrunner, A. (2011). Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (pp. 181-190).
- Kaltenbrunner, A., & Laniado, D. (2012). There is no deadline: time evolution of Wikipedia discussions. In *Proceedings of the 8<sup>th</sup> Annual International Symposium on Wikis and Open Collaboration* (pp. 1-10).

- Kopf, S. (2022). *A Discursive Perspective on Wikipedia: More than an Encyclopaedia?* Palgrave Macmillan.
- Herzberg, L. & Lügen, H. (to appear). Investigating reply relations on Wikipedia talk pages to reconstruct interactional strategies of Wikipedia authors. In C. Poudat, H. Lügen & L. Herzberg (Eds.), *Investigating Wikipedia: linguistic corpus building, exploration and analyses*. John Benjamins.
- Ho-Dac, L.-M. (to appear). Building a comparable corpus of online discussions in Wikipedia: the EFG WikiCorpus. In C. Poudat, H. Lügen & L. Herzberg (Eds.), *Investigating Wikipedia: linguistic corpus building, exploration and analyses*. John Benjamins.
- Laniado, D., Tasso, R., Volkovich, Y., & Kaltenbrunner, A. (2011). When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *Fifth international AAAI conference on weblogs and social media*.
- Lügen, H. & Herzberg, L. (2019). Types and annotation of reply relations in computer-mediated communication. *European Journal of Applied Linguistics* 7 (2). Berlin/Boston: de Gruyter, 2019. S. 305-331.
- Mehler, A., Gleim, R., Lücking, A., Uslu, T., & Stegbauer, C. (2018). On the Self-similarity of Wikipedia Talks: a Combined Discourse-analytical and Quantitative Approach. *Glottometrics*, 40, 1-45.
- Pentzold, C., Weltevrede, E., Mauri, M., Laniado, D., Kaltenbrunner, A., & Borra, E. (2017). Digging Wikipedia: The online encyclopedia as a digital cultural heritage gateway and site. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(1), 1-19.
- Schneider, J., Passant, A., & Breslin, J. G. (2010). A content analysis: How Wikipedia talk pages are used. In *Proceedings of the 2nd International Conference of Web Science* (pp. 1-7).



# “Don’t be afraid of Greeklish”: Adolescent students’ transliteration practices

Ralia Thoma

University of Thessaly

E-mail: [rthoma@uth.gr](mailto:rthoma@uth.gr) / [rallouthoma@gmail.com](mailto:rallouthoma@gmail.com)

## Abstract

Greeklish, the Latin-alphabet Greek used for the past 30 years in Computer-Mediated Communication (CMC), has sparked much debate in Greek society. However, previous research has mainly recorded adults’ transliteration practices. This study is concerned with the Greeklish transliteration practices used nowadays, mainly by adolescent users, and reports on any differences observed compared to those of adults. The analysis of the Greeklish corpus that was built for this purpose shows that adolescents transliterate some graphemes differently; however, consistency is observed in the transliteration of Greek graphemes that can be transliterated with more than one Latin grapheme, except for the grapheme <y>. Adolescents use mainly the mixed transliteration type, the combination of phonetic and orthographic transliteration, but prefer the orthographic transliteration of vowels in verbs and nouns, especially when those are positioned on the suffix of those word classes.

**Keywords:** Greeklish, adolescent students, transliteration practices

## 1. Introduction

Greeklish, the Latin-alphabet Greek (LaG) used in CMC environments by both adolescents and adults during the last three decades, has been a practice that sparked a lot of debate in Greek society as it was perceived as a threat to the Greek writing system and language (Tseliga, 2007; Androutsopoulos, 2009; Lees, 2017; Tzortzatou et al., 2018). Research on Greeklish has mainly examined adults’ transliteration practices and their attitudes toward the phenomenon, although in the last 10 years, with the popularisation of smartphones, Greeklish has been mainly used by adolescent students (Kavvadia, 2015; Lees, 2017; Tzortzatou et al., 2018).

This corpus-based study adopts the Computer-Mediated Discourse Analysis approach (CMDA) (Herring, 2004) to investigate adolescent students’ Greeklish transliteration practices. It aims to identify the Greeklish transliterations that are used by adolescent students aged 12-15 years old and examine whether these transliterations differ from those used by adults. The paper is also concerned with the three Greeklish transliteration types: phonetic, orthographic (keyboard-based and visual), and mixed transliteration (Androutsopoulos, 1999; Androutsopoulos, 2009; Panteli & Maragoudakis, 2011; Lees, 2017) and the changes that may be observed through the years.

## 2. A short history of Greeklish

The use of Greeklish – a composite word derived from the words Greek and English, although there is no use of the English language – dates to the Middle Ages (Panteli & Maragoudakis, 2011; Lees, 2017), a period when the official Greek language had not yet been established (Traintafyllidis, 1930). According to recorded evidence, LaG has been used in publications such as Byzantine texts, folk songs from Crete and Cyprus, Greek religious texts, books, letters, and contracts, by Greeks who lived abroad, by the population of Levantines who inhabited Smyrna, and whose mother tongue was Italian or French, due to the difficulties imposed by the Greek historical orthography

that resulted in numerous misspellings (Androutsopoulos, 1999; Tseliga, 2007).

Even in the recent past (around 1930), scholars argued in favor of the LaG, as it would solve the problem of misspellings that arises from the use of Greek’s historical orthography, help the dissemination of the Greek language, facilitate economic transactions, and strengthen relations with European countries (Fillidas et al., 1980). Nevertheless, the LaG was never officially adopted.

With the spectacular rise of the internet and the emergence of CMC, LaG has regained much attention under the name of Greeklish. The technological constraints of ASCII encoding eliminated non-Latin-alphabet languages like Greek, and Latinization of those scripts appeared to be the only viable alternative for digital communication (Tseliga, 2007; Lees, 2017).

Greeklish is characterized by a lack of consistency and spelling variance, as Greek-speaking users employ a range of non-standard ways that differ from person-to-person (Androutsopoulos, 2009; Lees, 2017).

Previous research highlights that although generalizations and individual transliteration choices are observed, Greeklish users show a systematicity in their transliteration choices (Androutsopoulos, 2009), resulting in the following three main types of transliteration (Chalamandaris et al., 2006; Androutsopoulos, 2009; Panteli & Maragoudakis, 2011; Lees et al., 2017):

- **Phonetic:** Latin graphemes are used to represent Greek phonemes based on sound correspondence, i.e., <ω>-<o>, <ι> -<i>, <φ>-<f>, resulting in the simplification or even elimination of the historical spelling: the graphemes <ι, η, οι, ει, υ>, <ο,ω> and the digraph <ου> are replaced by the corresponding Latin ones, i.e., <i> for the first group of vowels, <o> for the second group of vowels, and <u> for the third case, i.e., *ακολουθώ* (to follow) -*akolutho*.
- **Orthographic:** There are two sub-cases: a) keyboard-based transliteration: based on the QWERTY keyboard layout for Latin script languages, users type on their keyboard Latin

characters as they would if they were typing in Greek script, i.e., *ακολουθώ-akoloyuv*, b) visual transliteration: users use graphs in Latin which are similar in form to their Greek counterparts, i.e., *ακολουθώ: akolou8w/akolou9v/akolou0w*.

- **Mixed:** Users mix phonetic and orthographic (independent of its sub-cases) transliteration.

Table 1 below shows all types of Greeklish transliteration collected from previous research.

Greek Grapheme	Phonetic value	Greeklish transliteration		
		phonetic	orthographic	
			keyboard	visual
η	/i/	i	h	n
υ	/i/	i	y	u
ει, οι	/i/	i	ei, oi	ei, oi
ω	/o/	o	v	w
ου	/u/	u	oy	ou
β	/v/	v	b	b
γ	/i/ /y/ /ji/	g	y	y
θ	/th/	th	u	8, 0, 9
ξ	/ks/	x, ks	j	3
φ	/f/	f	f	ph
χ	/x/	ch, h	x	x
ψ	/ps/	ps	c	ps
αι	/e/	e	ai	ai
ευ	/ef/ - /ev/	ef, ev	ey	eu
αυ	/af/ - /av/	af, av	ay	Au
μπ	/b/	b	mp	mp
ντ	/d/	d	nt	nt
γκ	/g/ /j/	g	gk	gk

Table 1: Greeklish transliterations from previous research

### 3. The corpus of students' Greeklish transliterations

The Greeklish corpus was built by the author of the study as part of her Ph.D. research regarding eye movements of adolescent students when reading Greeklish transliterations, as there was no Greeklish transliteration corpus that could be used, only observations on adolescent data samples. This corpus served as a data sample of Greeklish transliterations from which we would choose the single words and sentences that would serve as stimuli in the eye-tracking tasks during the second research stage. Thus, the basis of content gathered for the analysis was limited to text-based CMC.

#### 3.1 Collection process

A group of seventy students ( $M$  age = 13.4  $SD$  = 1.0) from 2 different junior high schools, with the valuable help of their ICT teachers, participated in a three-step research task. Only students whose parents had signed the consent form participated in the following tasks.

In the first task, students were asked to complete an online questionnaire concerning their frequency of using Greeklish and their preferences in social media platforms and instant messaging apps. The researcher used this information to create the fake accounts - to protect students' personal data- in the students' favorite social media platforms (Facebook and Instagram) and instant messaging apps (Viber, FB Messenger, and WhatsApp) that would be used in the following tasks. Only students who reported using Greeklish "sometimes or all the time" participated in the transliteration activity.

Following, each school's ICT teacher supplied each participant with a digital device (smartphone or tablet), asked them to open their favorite social media platform or instant messaging app and to transliterate the 40 words that they would hear into Greeklish as if they were to use them in a message in an instant messaging app or a social media post. For this task, we selected 20 verbs and 20 nouns from the specialized corpus "Glossa" (Thoma, 2021) of high and medium frequency, consisting of 4-6 or 7-12 characters, considering eye-tracking reading research regarding lexical variables such as a word's class, length, frequency, and morphology that strongly influence fixation's duration (Rayner et al., 2015; Conklin & Pellicer-Sanchez, 2016).

In the third task, students were asked to select a classmate who participated in the research, were given a smartphone or a tablet, and were asked to chat using their favorite instant messaging app about anything they wanted using either Greek, Greeklish, or both.

Finally, we collected students' posts from their Facebook walls. The researcher contacted students whom she had "friends" on the social network Facebook and noticed that they used Greeklish in some or all of their posts. The initial contact was made with the students through Fb Messenger, followed by telephone contact with their guardians. Data from nine students were collected by searching the posts of the last two weeks from when the students and their guardians were informed of the survey and consented.

#### 3.2 Processing steps

In order to clean the data, we excluded words written in Greek script, emojis, emoticons, stickers, punctuation marks, and numbers that were not part of a word, e.g., prices and hours. However, we didn't delete abbreviations, such as *smr*, which stands for word *σήμερα* pronounced [simera] (today), or *8elo*, which stands for the verb *θέλω* [Télo] (I want), expressive texting such as *geiaaaaaaaaa* that stands for the word *γεια*-[ja] (hello). The size of the corpus is 7,837 tokens.

### 4. Annotation process

As stated, the corpus contains Greeklish tokens written using phonetic, orthographic, mixed, or unclassified transliteration types. Two annotators (the author of the study and an undergraduate student whose thesis was related to Greeklish), have independently annotated the corpus manually, based on previous research regarding Greeklish transliteration practices (Chalamndaris et al., 2006; Androutsopoulos, 2009; Panteli & Maragoudakis,



2011; Lees, 2017). The goal of the annotation is to identify whether the Greeklish tokens could be sorted into one of the three transliteration types: phonetic, orthographic, or mixed. Greeklish tokens that could not be assigned to one of the transliteration types are marked as unknown so that they can be identified and excluded from the analysis.

#### 4.1 Annotation labelling

To ensure the annotators followed the same annotation steps, we used the following annotation labels:

- **Phonetic transliteration:** For tokens where Latin graphemes are used to represent Greek phonemes based on sound correspondence, resulting in the simplification or even elimination of the historical spelling, i.e., *υπερβολές* (exaggerations)-[ipervolés]- *ipervoles*.
- **Orthographic transliteration:** To simplify this label, we included and labelled as “orthographic transliteration” the QWERTY keyboard layout sub-case, the visual transliteration, and their combinations, i.e., *ακολουθώ* (to follow) – [akoluTó]: QWERTY keyboard layout: *akoloyuv*, visual: *akolou8w/akolou9v/akolou0w*, combination: *akoloy8v/akoloy8w/akolou8w/akolou8w* and all the other combinations that may occur.
- **Mixed transliteration:** In cases where a combination of phonetic and orthographic transliteration is observed, visual or keyboard, i.e., *ακολουθώ* (to follow) – [akoluTó]: *akolou8o/akolu8w/akolou8v/* and all the other combinations that may occur.
- **Unknown transliteration:** In cases where the transliterated token could not be sorted into one of the above labels. In this case, we assumed that the students didn’t know the token’s correct orthography and the token was excluded from the analysis, i.e., *καταναλώνει* (to consume) - [katanalóni] was transliterated as *katanalonoï/katanalwnoi* while according to previous research regarding Greeklish transliterations the suffix *-ei* that in this verb encodes 3<sup>rd</sup> Person Singular should be transliterated as <i> in phonetic transliteration or as <ei> in the orthographic one.

The total number of tokens included in the analysis (see paragraph 5) was 3,469, as those were the only ones that could be sorted into the three transliteration categories – phonetic, orthographic, and mixed. AntConc’s N-Gram tool (Anthony, 2022) was used to verify the annotators’ labelling mentioned above. A sample of Greeklish transliterations before and after annotation is shown in Table 2.

Task	Tokens	Annotated Tokens
Word transliteration	2,426	2,156
Chatting	1,489	525
Fb posts	498	263
Total	7,837	3,469

Table 2: Greeklish transliterations before and after processing

## 5. Analysis and Results

The analysis is conducted by adopting the CMDA approach (Herring, 2004; Herring, 2019) regarding the level of structure (micro-linguistic) for examining phenomena such as typography, orthography, morphology, and formatting conventions. Table 3 describes the CMDA research process applied (Herring, 2004, p. 24).

CMDA research process	Application to our research
Articulate research questions	RQ1: How do adolescent students transliterate Greek graphemes? RQ2: How does those transliterations differ from adults’ transliterations?
Select computer-mediated data sample	Corpus of students’ Greeklish transliterations
Operationalize key concepts in terms of discourse features	Structure level: typography, graphemes’ transliterations, transliteration categories
Select and apply method(s) of analysis	Structural analysis: graphemes’ transliterations
Interpret results	Conclusions’ paragraph

Table 3: Application of CMDA research process

The transliterations were characterized according to data from previous research (Chalamndaris et al., 2006; Androutsopoulos, 2009; Panteli & Maragoudakis, 2011; Lees, 2017). In cases where a combination of phonetic and orthographic transliteration is observed, visual or keyboard, the transliteration is characterized as mixed. In cases where a combination of orthographic and visual transliteration was observed, the transliteration is characterized as orthographic.

### 5.1 Students’ transliterations

The transliterations used by the students are the same regardless of the activity (word transliterations, communication simulation, Fb posts). Some differences are observed in comparison to the transliterations recorded by previous research in adult Greeklish users (Table 4).

Greek grapheme	Phonetic value	Adults' transliterations (previous research)	Students' transliterations
γ	[i] [y] [ji]	g, y	g
θ	[θ]	th, u, 8, 9, 0	th, u, 8
ευ	[ev]	eu, ey, ev	eu, ey, ev, eb
αυ	[av]	au, ay, av	au, ay, av, ab
ξ	[ks]	ks, x, j, 3	ks, x, j, 3, 4
ντ	[d]	nt, d	nt

Table 4: Differences in Greeklish transliterations between adults and students

The results demonstrate that consistency is observed in the transliteration of Greek graphemes that can be transliterated with more than one Latin graphemes., i.e., students who transliterated <θ> as <th> choose only this transliteration, e.g., *thermokrasia* – *θερμοκρασία* (temperature)-[Termokrasía] with only one student's exception of <θ> depending in its position in the word, i.e., <8> at the beginning of the word but as <th> in any other place. The same case was observed for the transliteration of the grapheme <β> as <b> or as <v>, e.g., *yperboles-uperbolés* (exaggerations)-[ipervolés] but *vouliázame*-*βουλιάζαμε* (we are sinking)- [vuLázame] (2 students, 2.53%), or for the grapheme <ξ> that was transliterated as <3> when in word's stem, but as <j> when in word's middle or suffix (3 students, 3.80%).

However, the same is not true for the transliteration of the grapheme <y>. Thus, while in several cases students transliterated the grapheme <y> as <y>, in the cases of diphthongs, they used <u> in <au>-<av> and <y> in <ey>-<ev> regardless of their pronunciation, [af]-[av] and [ef]-[ev], respectively (15 students, 18.99%).

This finding seems coherent with the observation regarding the Greek grapheme <υ> -<i>-[i] when alone or when found in the digraph <ov>-[u]-for which most students (65%) used different transliteration types even in their data, i.e., a students used the following transliterations of <υ>: *lugiseis* – *λυγίσεις* (to bend) – [lijísis], *mousikh* – *μουσική* (music) – [musicí].

## 5.2 Greeklish transliteration types used by students.

Regarding the transliteration type/s that students use, the analysis of data show that the mixed transliteration type is the most popular one, as most students (67 students, 84.81%) combine phonetic and orthographic transliteration in the graphemes of the same word, i.e., in the word *autokinitodromos* – *αυτοκινητόδρομος* (highway)-[aftocinitóDromos] we observe the orthographic transliteration of the diphthong vowel <av>-<au>- [af], but the phonetic transliteration of the grapheme <η>- <i>-[i] or in the sentence *Tous bare9hka na milane enw den akoune* - *Τους βαρέθηκα να μιλάνε ενώ δεν ακούνε* (I'm tired of them talking while not listening)- [tus varéTika na miláne enó

Dén akóne]. Few students (15.19%) transliterated words using only the orthographic transliteration type (keyboard and/or visual), whereas no student used the phonetic transliteration type exclusively.

Going further with the analysis, the picture described so far gets clearer if vowel transliterations are examined. The majority of students used only one type of transliteration, phonetic or orthographic, with the orthographic one to be used in most cases, i.e., *Pou kollaei auto twra* - *Πού κολλάει αυτό τώρα* (literally: where does this stick now, acceptable colloquial speech: this is not relevant to our discussion) - [pú kolai aftó tóra], in which the bolded vowels (including diphthongs) are orthographic transliterations.

However, specific attention should be paid to the transliteration used in some vowels regarding their position in the word (stem or suffix) and the word's class. The analysis of data collected from the “word transliteration task” showed that students prefer the phonetic transliteration of grapheme <ω>-<w>-/o/ when positioned in verbs' stem, while there is little difference among the phonetic and the orthographic transliteration when <ω> is positioned in the verbs' suffix. The orthographic transliteration type is preferred for the digraphs <αι>-<ai>-[e] and <ει>-<ei>-[i], regardless of their position in the verb's stem or suffix. Accordingly, regarding nouns, most students prefer the orthographic transliteration of the grapheme <η> - <h> - [i] and the digraph <οι> - <oi> -[i], regardless of their position in the stem or the suffix of the noun (Table 5).

Words' class	Grapheme/digraph	Stem		Suffix	
		Orth.	Phonetic	Orth.	Phonetic
verbs	<ω> - [o]	18.06%	80.56%	48.61%	50.00%
	<ει> -[i]	59.72%	38.89%	72.22%	26.39%
	<αι> - [e]	44.44%	54.17%	86.11%	12.50%
nouns	<οι> -[i]	83.33%	15.28%	94.44%	4.17%
	<η> -[i]	66.67%	31.94%	51.39%	47.22%

Table 5: percentages of students that used orthographic or phonetic transliteration for graphemes/digraphs when positioned in the suffix/stem of verbs/nouns

## 6. Conclusions

In this paper examined adolescent students' transliteration practices when using Greeklish in CMC. The results show that students use the same transliteration practices as adults, with few exceptions that we assume may be due to the digital device used, i.e., smartphones or tablets instead of laptops or desktops. The Greeklish corpus analysis also showed that most students use mainly the mixed transliteration, a practice underlined from previous research but not in the same frequency, as adults preferred orthographic or phonetic transliteration (Androutsopoulos, 2009; Lees, 2017). Although Greeklish lacks consistency (Androutsopoulos, 2009), we observed that most students prefer the orthographic transliteration for vowels and digraphs in verbs and nouns, with some exceptions depending on their position in the stem or the suffix of the

word, an observation supported by previous research (Androutsopoulos, 2009; Lees, 2017). The use of orthographic transliteration by most students can be explained by previous research that underlines the development of morphological awareness in this age group (Carlisle, 2003).

Greeklsh has been the focal point of discussion examined mainly from the ideological approach of orthography that “views orthography as a set of social practices in specific social and cultural contexts” (Androutsopoulos, 2009:222). Most participants in Greeklsh studies, including its users, express their concern about Greeklsh’s detrimental consequences on the Greek language; teachers and parents have voiced a negative view toward Greeklsh, arguing that its use affects the spelling and written language of its users, or that they would even forget the Greek language (Koutsogiannis, 2015; Tzortzatos, et al., 2018). Thus, a campaign (social media, TV posts, popular TV shows, TV and radio publications) took place in Greece against the use of Greeklsh, which considering the results of this research regarding the use of orthographic transliteration, we assume that has fulfilled its purpose.

The limitation that needs to be discussed and restricts the meaningfulness of this study is mainly its sample size and the percentage of Greeklsh users that seems to decrease (Lees, 2017). Nowadays, people, especially students, invent new communication practices such as using Greek abbreviations, emojis, emoticons and some Greeklsh transliterations. A longitudinal study regarding children’s exposure to social media and digital devices that aims to collect and create a corpus of CMC practices would help in their in-depth study.

## 7. Copyrights

Proceedings will be published under a [Creative Commons Attribution 4.0 International license](#).

## 8. References

- Androutsopoulos, J. (2009). “Greeklsh”: Transliteration Practice and Discourse in the Context of Computer - Mediated Digraphia. In A. Georgakopoulou & M. Silk (Ed.), *Standard Languages and Language Standards: Greek, Past & Present*. Hampshire: Ashgate Publishing Ltd, 221-249.
- Androutsopoulos, J. (1999, September 5). *Από τα φραγκοχιώτικα στα greeklsh* {From Fragochiotika to Greeklsh}. To Vima. <https://www.tovima.gr/2008/11/24/opinions/apo-ta-fragkoxiwtika-sta-greeklsh/>
- Anthony, L., (2022). AntConc (version 4.2.0) [Computer Sftware]. Tokyo, Japan. Available from <https://www.laurenceanthony.net/software>
- Chalamandaris, Ai., Protopappas, Ath., Raptis, Sp. (2006). All Greek to me! An automatic Greeklsh to Greek transliteration system. Proceedings of the *5th International Conference on Language Resources and Evaluation (LREC’06)*, 1226-1229.
- Carlisle, J.F., Fleming, J. (2003). Lexical Processing of Morphologically Complex Words in the Elementary Years. *Studies of Reading*, 7(3), 23-253.
- Conklin, J., Pellicer - Sanchez, A. (2016). Using eye - tracking in applied linguistics and second language research. *Second Language Research*, 32(3), 453-467.
- Filintas, M., Glinos, D., Sideris, G., Gifyllis, F., Chatzidakis, N., Prousis, K., Karthais, K., Benekos, G. (1980). *Φωνητική γραφή* {Phonteic script}. Athens: Kalvos
- Herring, S.C. (2019) The Coevolution of Computer-Mediated Communication and Computer-Mediated Discourse Analysis. In: P. Bou-Franch, B.P. Garcés-Conejos (Ed.), *Analyzing Digital Discourse*. USA: Palgrave Macmillan, Cham, 25-67.
- Herring, S.C., (2004). Computer - mediated discourse analysis: an approach to researching online communities. In S.A. Barab, R. Kling, J.H. Gray (Ed.), *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press, 338-376.
- Kavvadia, A. (2015). Technolect, Greeklsh και ορθογραφία: έρευνα σε μαθητές /τριες δευτεροβάθμιας εκπαίδευσης { Technolect, Greeklsh and orthography: research with students in secondary education}. *Studies in Greek Linguistics*, 35, 632-642. [http://ins.web.auth.gr/images/MEG\\_PLIRI/MEG\\_35\\_6\\_32\\_642.pdf](http://ins.web.auth.gr/images/MEG_PLIRI/MEG_35_6_32_642.pdf)
- Koutsogiannis, D. (2015). Translocalization in Digital Writing, Orders of Literacy, and Schooled Literacy. In Sc. Bulfin, N.G., Johnson, Ch. Bigum (Ed.), *Critical Perspectives on Technology and Education*. New York: Palgrave Macmillan, 183-202
- Lees, Ch. (2017). *Γλωσσικές πρακτικές των νέων σε τόπους κοινωνικής δικτύωσης: Η περίπτωση του Facebook* {The language practices of young people on social networking sites: the case of Facebook} (Publication ND 41809) [Doctoral thesis, Aristotele University of Thessaloniki]. National Archive of PhD Theses. <http://hdl.handle.net/10442/hedi/41809>
- Panteli, A., Maragoudakis, M. (2011). A Random Forests Text Transliteration System for Greek Digraphia. In L. Iliadis, I. Maglogiannis, H. Papadopoulos (Ed.), *Artificial Intelligence Applications and Innovations. EANN 2011, AIAI 2011. IFIP Advances in Information and Communication Technology*, 364. Berlin, Heidelberg: Springer, 196-201.
- Rayner, K., Schotter, E.R., Masson, M.E.J., Potter, M.C., Treiman, R. (2015). So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? *Psychological Science in the Public Interest*, 17(1), 4-34.
- Thoma, R. (2021). *Η επίδραση των Greeklsh στη γλωσσική επίγνωση των μαθητών* {The impact of Greeklsh on students’ language awareness} (Publication ND 50229) [Doctoral thesis, University of Thessaly]. National Archive of PhD Theses. <http://hdl.handle.net/10442/hedi/50229>
- Tseliga, Th. (2007). “It’s all Greeklsh to Me!” Linguistic and Sociocultural Perspectives on Roman - Alphabetized Greek in Asynchronous Computer - Mediated Communication. In B. Danet & S. Herring (Ed.), *The Multilingual Internet: Language, Culture, and*

- Communication Online*. New York: Oxford University Press, 116-141.
- Trianatafyllidis, M. (1930). *Το πρόβλημα της ορθογραφίας μας* {The problem of our spelling}. Athens: private edition
- Tzortzatos, K., Archakis, A., Iordanidou, A., Xidopoulos, G. I. (2018). Η ιστορική ορθογραφία και τα Greeklish στην εκπαίδευση: Ορθογραφικές Στάσεις στο Δημοτικό {Historical spelling and Greeklish in education: Spelling Attitudes in Primary School}. *Studies in Greek Linguistics*, 38, 227-240.

# *Not an expert, but not a fan either.* A corpus-based study of negative self-identification as epistemic index in web forum interaction.

Eva Triebel

University of Vienna  
E-mail: eva.triebl@univie.ac.at

## Abstract

This study examines the linguistic micro-management of identity in and across online contexts, drawing upon corpus-based pragmatic analysis of a structure with a meaning potential to examine wider questions about identity in digitally mediated social life. The structure analyzed are negative self-identifiers of the type “I + copula + not + indefinite NP” used in UK web discussion forums. It was chosen as it explicitly relates the speaker with the notion of interest, namely identity, and, by negating explicitly stated or presupposed claims, indexes how speakers perceive, and discursively create, the context they are writing into. Qualitatively and quantitatively analyzing the forms and functions of 936 instances of the structure in their co-texts, negative self-identifiers from the fields of expertise and preferences were found to be salient in the examined data, framing co-texts in which speakers linguistically enacted various forms of expertise, pointing to heightened reflexivity regarding the epistemic status and social impact of their utterances and a reconceptualization of expertise as a transient discourse phenomenon rather than a more permanent identity feature.

**Keywords:** corpus pragmatics, expert identity, epistemic management

## 1. Introduction: Why study what forum users say they are not

The pragmatic effects of negative self-identifiers (henceforth NSIs) are noteworthy because they are uninformative unless seen as context-sensitive meaning potentials serving to defeat explicit or implicit identity claims present in the immediate co-text, the situational context or in the wider cultural context of the utterance (Givón, 1993:191; Jordan, 1998: 706). Therefore, they can fruitfully be studied to learn about conceptualizations speakers contrast themselves with, and thus orient towards, in online interaction. Like Barron and Schneider (2014: 1), this study adopts the view that “the pragmatics of discourse and the pragmatics of utterances are two complementary levels of analysis, respectively highlighting more global and more local aspects of human communication”, and takes a micro-pragmatic starting point to empirically examine and critically reflect upon the role of language in contemporary macro-social processes.

<sup>1</sup> It was decided to exclude NSIs from the corpus that appeared in instances of active voicing (e.g. in *He said, “I’m no liar”*) or in embedded clauses with subjects other than the first person singular (e.g. *She can’t argue that I am not an expert*), as referring to someone else’s identity ascription is not the same as negatively identifying with a particular NP oneself.

The overarching research topic to which this study contributes is the digitally mediated, reflexive performance of the (disembodied) self in online discourse (Benwell & Stokoe, 2006: 278; Leppänen et al., 2015: 1) against a background of pluralizing choices and individualization, foregrounding authenticity and, thus, difference and differentiation (Beck & Beck-Gernsheim, 2001). In light of these larger-scale trajectories of change, negative self-identifiers can be seen as instances of local disalignment with situationally relevant categories, but potentially also as indices of orientation towards macro-conceptualizations underlying contemporary ways of being and interacting (van Dijk, 2015: 468).

To explain the relation between authentic instances of negative self-identification and their digital contexts of use, I theorize NSIs as contextualization cues (Ochs, 1995; Aijmer, 2013) which may be used strategically by speakers to modify the interpretation of their utterances, designed based on their evaluations of their audience in collapsed online contexts (Marwick & boyd, 2011; Tagg et al., 2017). In this view, NSIs are the most explicit linguistic means of positioning speakers in relation to what they post online. This, in turn, raises the question of what conceptualizations speakers, by negatively identifying with them, linguistically index as contextually relevant frames for interpretation of their postings, and whether certain conceptualizations can be found to be linguistically foregrounded in particular co-texts in patterned ways.

## 2. Study design and data

The corpus for this study was created to represent variants of the formally defined structure “I + copula + not + indefinite NP” in their utterance-internal and sequential co-texts as used in a particular type of digitally enabled interaction, namely web forums. As forums unite users with at least one shared interest, but with potentially very diverse backgrounds, they represent interesting sites for studying linguistic disalignment with particular – and, if considered as patterns – superordinate categories. Transtextually indexed types of negative self-identification could point towards certain concepts standing out as salient and thus characterize self-representation on forums more generally (Spitzmüller & Warnke, 2011: 82). Based on these considerations, corpus compilation was guided by linguistic and platform-related criteria.

As for the linguistic criteria for including instances of NSIs, customized Google searches were employed to find the following formal variants of the matrix clause:

- **Tenses:** present simple, present perfect simple (*I’m/am not, I’ve/have never been*)<sup>1</sup>
- **Contraction:** *I’m not, I am not*
- **No-negation:** *I am no, I’m no*
- **Constructions with *never*:** *I have never been*<sup>2</sup>

<sup>2</sup> The reason why constructions with *never* are listed as a variant in their own right is that they occurred significantly more often in the data than present perfect tense NSIs with no adverbial modifiers (such as *I haven’t been a basketball player for two years*).

- **Adverbs:** e.g. *I'm not really, I'm definitely not*
- **Indefinite article:** *I'm not a/an*

Regarding the platforms from which data was taken, the search was limited to publicly available UK websites in English language featuring the word “forum” or “thread” in their domains. The data was not controlled for topic, purpose or user characteristics, representing a wide variety of contexts in which variants of the formally defined structure appeared. As for corpus size and balance, data collection was systematically randomized by retrieving the same number of instances for each formally defined variant from each page of results of the respective Google searches until a target of 100 occurrences was reached. In cases where the searched variant occurred fewer than 100 times in total, all instances were included. While this has the disadvantage of the corpus not reflecting the actual proportions of the respective variants’ frequencies, sampling proportionately would have entailed the exclusion of infrequent variants, which are now (over-)represented in the corpus. To obtain a sufficiently large sample of the structure in use, and in light of the potentially asynchronous nature of forum interaction, a time span for data collection was defined as the criterion for including NSIs in the corpus (July–September 2019). This means that the corpus is a snapshot of NSIs as they appeared on web forums at the time of compiling it, capturing online interaction in which the structure appeared over a longer period. The reason for focusing on adequately representing the form, while leaving sufficient room for contextual variation, is the study’s micro-pragmatic approach, which implies an emphasis on linguistic details to identify patterns as potential indicators of longer-term, gradual and inherently fuzzy phenomena.

As for the question of what kind of co-text, and how much of it, was included in the corpus, NSIs were collected together with their utterance-internal co-text and their sequential, utterance-external co-text, i.e. together with the posting(s) to which they replied or which followed them. Thus, a corpus of 936 instances of the matrix clause in their contexts of use was gathered by searching for, reading, selecting, copying, pasting, and storing postings in text files. As a backup, the entire websites from which the data was taken were also downloaded and stored separately.

Then, in an iterative process, metadata about textual and contextual aspects, namely the meaning of the identifying NP, the formal appearance and functions of the immediate and wider co-texts in which the structure appeared, the topic of the thread and the forum featuring the NSI, were manually added to the data by using tags. The annotation, and thus the qualitative analysis, began at the conceptually most important and syntactically narrowest level, namely with a semantic profile of identifying NPs, and proceeded in structurally ascending steps by formally and functionally categorizing the sentence-internal and sentence-external co-texts of NSIs. This was done by adding information about the data in the form of corpus annotation and by using MS Excel. The frequencies of the identified categories, and their patterned relations, were

analyzed using the concordancing function of WordSmith 5.0 (Scott 2008) as well as Excel’s sorting and calculating functions.

To close the gap between the micro-pragmatic study around the focal structure and questions about the relations between NSIs and their online contexts of use, high-frequency NSIs were also qualitatively examined in their situational contexts. In section 3, the following key findings of the study will briefly be presented:

- The conceptual profile of identifying NPs
- The formal-functional profile of the clause-external co-textual relations of attested instances of the structure
- The transitivity analyses of clauses formally related to NSIs
- The functional analyses of high-frequency processes represented by clauses formally related to NSIs
- The functional analyses of co-texts at or beyond sentence level preceding NSIs

In section 4, then, the implications of the findings of these analyses regarding functional patterns and the conceptualizations about self-presentation they construe in online interaction will be critically discussed.

### 3. The empirical study

#### 3.1. The meanings and co-textual relations of NSIs

Figure 1 below shows a conceptual profile of identifying NPs, in which the domains of expertise and professionalism (notably constructions with *expert*) on the one hand and preferences (notably constructions with *fan*) on the other are most prominently represented.

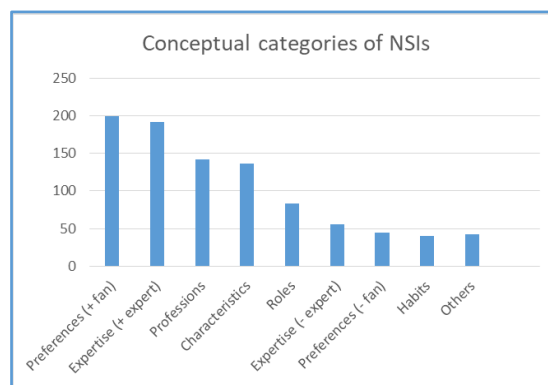


Figure 1: Overview of conceptual categories of identifying NPs

Table 1 gives an overview of the relations of instances of NSIs with their clause-external co-texts. As can be seen, NSIs appear as part of complex clauses in 717 of 936 cases, with cases where the structure precedes a clause marked as contrast being by far most frequent. This, coupled with the fact that NSIs were found to be often lexically fixed and, thus, formulaic (appearing as variants of the form “I’m not an expert” and “I’m not a fan”, respectively), indicates that they tend to be strategically used as framing devices, pre-emptively negating anticipated implications of utterances following them.



Relationship	1L <sup>3</sup> co-text	1R co-text	No. of NSIs	% of all instances
<b>Contrast and concession</b>				
Contrast	86	279	365	
NSI = concessional clause	7	16	23	
1R/1L = concessional clause	4	5	9	
<b>Total</b>			<b>397</b>	<b>42%</b>
<b>Cause and consequence</b>				
NSI as cause	41	93	134	
NSI as consequence	4	20	24	
<b>Total</b>			<b>158</b>	<b>17%</b>
<b>Addition</b>				
Coordination	24	138	162	
<b>Total</b>			<b>162</b>	<b>17%</b>
<b>TOTAL</b>	<b>166</b>	<b>551</b>	<b>717</b>	

Table 1: Relationships between NSIs and their immediate co-texts

## 3.2. Functionally profiling textual environments

### 3.2.1. Transitivity analysis of immediate co-texts

To learn what co-textual meanings NSIs interact with, the transitivity framework (Halliday & Matthiessen, 2014) was employed to categorize the textual surroundings of the matrix clause. It was found that of the 717 clauses with formal links to the matrix clauses, 443 have a first-person participant as thematic subject<sup>4</sup>. These most often represent speakers' thoughts and ideas, either through mental processes (e.g. *think*, *find*) or through relational processes (most of which relate the speaker to emotional or cognitive traits and abilities, e.g. *capable*, *interested*). In mental process co-texts, identifying NPs from the areas of expertise and professionalism occur in more than half of the examined clauses (in 103 of 192 cases, i.e., 53%). In these co-texts, preference disclaimers – the most frequently instantiated conceptual category of NSIs in the entire corpus – appear in only 39 cases (i.e., in 20% of examined clauses). These findings indicate the patterned use of disclaimers of expertise in conjunction with clauses indexing the subjectivity of claims they project.

According to Myers (2006: 77–78), using the phrase *in my opinion* when expected to take a stance anyway signals awareness of the potential of opinion statements to serve multiple epistemic and social functions and to be subject to context-specific constraints. Accordingly, the use of NSIs to modify expressions of opinion can be considered to make explicit those aspects of a speaker's identity felt to be constraining the appropriateness of their claims, with expertise being the identity category linguistically foregrounded. In other words, epistemic hedging is one, but certainly not the only or even most important function of negative self-identification with NPs from the field of expertise, just like framing claims as 'opinions' does not only mark them as potentially contestable. Scrutinizing examples of NSIs from the corpus suggests that the structure also fulfils functions relating to face management. For example, in *I'm not an expert but I believe I have a*

*good grasp of the laws of the game*, the NSI signals awareness of the potential face threat involved in the speaker's claim to expertise.

As for the use of preference disclaimers in contexts presenting speakers' views, such NSIs tend to appear in textual environments featuring linguistic indices of authority. The self-confidence and 'sassy' rhetoric of the corpus example below, containing a preference disclaimer, is a case in point. Despite framing their assessment as *personal opinion*, the speaker expresses certainty when (mockingly) assessing the appearance of the product in question (*it certainly does not stand out*), and implicitly addresses the designers with suggestions for improvement (*Perhaps just a little metallic band across it*), representing themselves as undisguisedly subjective, but situationally authoritative:

Yes, I can see that, it certainly does not stand out. [...] Perhaps just a little metallic band across it, in a similar tone to the fabric. [...] To be totally honest, **I'm not a fan of the Home Max Speaker** for the same reason. It's just lacking something, just an element to stop it looking like a fat grey lump :) All personal opinion of course :)

Judging from the results of these analyses, it seems that speakers use NSIs in contexts of linguistically performing expertise, with the two salient phrases *I'm not an expert* and *I'm not a fan* indexing orientation towards different notions of expertise deemed relevant in the surrounding posting, namely lay (as opposed to formal) expertise in the first case, and preference as an indicator of experiential expertise in the latter. To get a fuller picture of the communicative functions of NSIs in their clause-internal co-texts, mental, relational and material process clauses with *I* as Role-1 participant conjoined with the matrix clause were selected for finer-grained analysis, presented in the next section.

### 3.2.2. The functions of sentence-internal co-texts

For the functional analysis of the sentence-internal co-texts of NSIs, a framework was devised that considers both the meaning of the verb phrases and their projected clauses, as well as aspects such as tense, aspect and polarity. It differentiates, e.g., between emotive verbs with a complement referring to the addressee (e.g. *I highly appreciate your reply*), verbs of perception with a complement referring to a contextually relevant object or question (e.g. *I can see small teeth at the front of the lower jaw*), and verbs of perception in the past tense, describing an experience rather than an in-situ impression (e.g. *I have not experienced many changes in medication*).

Table 2 presents the categories of communicative functions identified in the examined clauses formally related to the matrix clause. The analysis revealed that NSIs are characteristically used in textual environments expressing speakers' beliefs, opinions and experiences,

as actors, experiencers, carriers, sayers, existents and behaviors under the label of "Role 1"-participants.

<sup>3</sup> 1L and 1R refer to the co-text immediately left and right to the matrix clause.

<sup>4</sup> Because thematic roles do not necessarily coincide with grammatical subjects, the study subsumed participants appearing

which, as the examples show, are linguistically realized not only by mental process co-texts (e.g. *I'd guess a telemark is something to do with the binding*), but also by co-texts formally representing relational (e.g. *[I] have got the impression that they don't flex their immune protocols much*) and material processes (e.g., *[I] have worked with it a lot*).

In all three types of co-text, NSIs from the conceptual domain of expertise occur relatively most frequently (in 61, 29, and 54 instances, respectively). This further supports findings of the previous analyses, which pointed to a tendency for disclaimers of expertise to modify speakers' claims or references to (different kinds of) knowledge. In co-texts referring to speakers' experiences, habits and principles (e.g. *i usually play solo*), preference disclaimers are relatively most frequent; however, overall, the relation between this type of NSI and this co-textual category cannot be claimed to constitute a pattern so prominent as the one that could be observed for epistemic disclaimers and co-texts presenting and negotiating knowledge.

Functional profile of clauses formally linked to NSIs				
Functional category	Men.	Rel.	Mat.	Total
Knowledge representation/Opinion	72	20	1	93
				61 expertise NSIs
Knowledge/Understanding reference	31	26	5	62
				29 expertise NSIs
Experience	18	22	22	61
				54 expertise NSIs
Preferences/Habits/Principles	24	10	19	53
				31 preference NSIs
Others <sup>5</sup>				104
TOTAL	192	125	84	401

Table 2: Functional profile of clauses formally related to NSIs

Considering NSIs in these categories in relation to their wider discourse contexts revealed that interestingly, they do not only serve to justify potential limitations of expertise speakers share on web forums, but also index that speakers, despite not being formally accredited experts, are aware of their knowledge and skills and, thus, project epistemic self-confidence.

### 3.2.3. The functions of sentence-external co-texts

To learn about the relations between NSIs and their (most frequent) clause-external co-texts beyond sentence level, the 376 declarative sentences preceding NSIs, and, if applicable, the discourse unit they were part of, were analyzed in terms of their content and pragmatic function.<sup>6</sup> The identified functions of these co-texts were then, again, cross-categorized with conceptual categories of NSIs.

The three most prominent categories identified are discourse units representing users' experiences with

products (67 instances), representations of, and reflections upon, knowledge and information (38), and advice (34). Product experience stories are most often followed by preference disclaimers (26/67 instances); factual claims and reflections upon speakers' understanding as well as instances of advice predominantly precede disclaimers of expertise (27/38 and 23/34 instances, respectively).

This means that in contexts where knowledge is shared and negotiated, disclaimers of expertise are used in patterned ways. While speakers are hesitant to represent what they know as unproblematic, using expertise disclaimers to epistemically hedge factual claims, they are more self-confident when discussing consumption choices.

## 4. Critical discussion and conclusion

Formal expertise, on the one hand, and informed choice-making, on the other, figure as key identification paradigms in the examined data, manifesting themselves in speakers' micro-management of what they post online. They point to struggle around two superordinate notions structuring meaning-making in forum interaction, namely epistemic certainty, which is challenged in contexts marked by the absence of cues about speakers' 'real' identity and expertise, and appreciation and relational work, which plays a pivotal role on media defined by their sociality (Petroni, 2019).

In online contexts, where the risk of emotional disagreement is high (Langlotz & Locher, 2012), negatively identifying as an expert could function as a strategy of cancelling, at least formally, the power differential implied by metadiscursive processes of explaining, rationalizing and assessing information (Silverstein, 2003). As Au and Eyal (2022) put it, "presenting oneself as 'not an expert' is a useful strategy to bypass the crisis of expertise that would shut down lines of communication when the contested identity of the credentialed expert is invoked". In the following corpus example, an NSI follows an otherwise unmitigated piece of advice, but precedes an invitation for others to voice their views, thus illustrating the tension between enacting and disclaiming expertise:

Whatever oil you use change it at the recommended times and keep the air filter clean. I repeat that **I am not an expert** and welcome other opinions.

Preference disclaimers, conversely, make reference to a less problematic, because inherently subjective, identification category – that of *fans*, *lovers*, and so on – and frame experience accounts, which could be associated with a lower risk of offending others, while indexing speakers' awareness and liberty of choice.

What counts as expertise, and how speakers hence use NSIs to position themselves in relation to it, appears to depend on the speech situation, i.e., drawing upon pragmatically appropriate registers is what construes

<sup>5</sup> For reasons of space, categories with fewer than 30 instances assigned to them were not included in this table.

<sup>6</sup> The analysis loosely followed the BCU approach for top-down corpus-based analysis of texts by iteratively segmenting the text into functional units and analyzing these functional paradigms to

define their lexical and grammatical characteristics (Upton & Cohen, 2009). The criteria for considering stretches of text as unit were an identifiable macro-topic on the ideational level, a discernible pragmatic purpose on the interpersonal level, and textual cohesion on the textual level.

credibility online (Mey, 2001: 220). Speaking like an expert, rather than identifying as one, seems more important in post-Panoptic online sociality, where the system of sayability, and not (just) visibility, is what is appreciated or sanctioned (Caluya, 2010). Meticulous linguistic analysis of identity work in corpora of digitally mediated discourse can provide empirical evidence for patterns of speaking that make today's online experts; after all, the "big things reside in the small things, and the most inconspicuous and uniquely situated social action is, in that sense, 'systemic' and 'typical'" (Blommaert et al., 2018: 5).

## References

- Aijmer, K. (2013). *Understanding Pragmatic Markers: A Variational Pragmatic Approach*. Edinburgh: Edinburgh UP.
- Au, L. & G. Eyal (2022). Whose Advice is Credible? Claiming Lay Expertise in a Covid-19 Online Community. *Qual Sociol* 45: 31–61. <https://doi.org/10.1007/s11133-021-09492-1>
- Beck, U. & E. Beck-Gernsheim (2001). *Individualization: Institutionalized Individualism and its Social and Political Consequences*. London [etc.]: SAGE.
- Benwell, B. & E. Stokoe (2019). *Discourse and Identity*. Edinburgh: Edinburgh UP.
- Blommaert, J., L. Smits & N. Yacoubi (2018). Context and its complications. *Tilburg Papers in Culture Studies* 208: 1–21.
- Caluya, G. (2010). The post-panoptic society? Reassessing Foucault in surveillance studies. *Social Identities* 16 (5): 621–633.
- Van Dijk, T. (2015). Critical Discourse Analysis. In D. Tannen, H. Hamilton & D. Schiffrin (Eds.). *The Handbook of Discourse Analysis*, 2<sup>nd</sup> ed. Chichester: Wiley Blackwell: 466–485.
- Givón, T. (1993). *English Grammar: A Function-Based Introduction*. Amsterdam: Benjamins.
- Halliday, M.A.K. & C. Matthiessen (2014). *Halliday's Introduction to Functional Grammar*. 4<sup>th</sup> ed., Rev. by C. Matthiessen. Oxfordshire, UK/New York: Routledge.
- Jordan, M. P. (1998). The power of negation in English: Text, context and relevance. *Journal of Pragmatics* 29: 705–752.
- Langlotz, A. & M. A. Locher (2012). Ways of communicating emotional stance in online disagreements. *Journal of Pragmatics* 44: 1591–1606.
- Leppänen, S., Möller, J. S., Nørreby, T. R., Stæhr, A., & Kytölä, S. (2015). Authenticity, normativity and social media. *Discourse, Context and Media* 8: 1–5. <https://doi.org/10.1016/j.dcm.2015.05.008>
- Marwick, A. E. & D. boyd (2011). 'I tweet honestly, I tweet passionately': Twitter users, context collapse, and the imagined audience. *New Media & Society* 13 (1): 114–133.
- Mey, J. L. (2010). Reference and the pragmeme. *Journal of Pragmatics* 42: 2882–2888.
- Myers, G. (2001). 'In My Opinion': The place of Personal Views in Undergraduate Essays. In M. Hewings (Ed.). *Academic Writing in Contexts: Implications and Applications*: 63–78.
- Ochs, E. (1996). Linguistic resources for socializing humanity. In J. Gumperz & S. Levinson (Eds.). *Rethinking Linguistic Relativity*. Cambridge: CUP: 407–437.
- Petroni, S. (2019). How Social Media Shape Identities and Discourses in Professional Digital Settings: Self-Communication or Self-Branding? In P. Bou-Franch Blitvich & P. Garcés-Conejos (Eds). *Analyzing Digital Discourse: New Insights and Future Directions*. Cham: Springer.
- Scott, M. (2008). *WordSmith Tools* Version 5. Liverpool: Lexical Analysis Software.
- Silverstein M. (2003). Indexical order and the dialectics of sociolinguistic life. *Lang. Comm.* 23: 193–229
- Spitzmüller, J. & I. Warnke (2011). Discourse as a 'linguistic object: methodical and methodological delimitations. *Critical Discourse Studies* 8 (2): 75–94. <https://doi-org-10.1080/17405904.2011.558680>
- Tagg, C., P. Seargeant & A. Brown (2017). *Taking Offence on Social Media: Conviviality and Communication on Facebook*. Cham, Switzerland: Palgrave Macmillan.
- Upton, T. A. & M. A. Cohen (2009). An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies* 11 (5): 585–605. <https://doi-org.10.1177/1461445609341006>

# Towards a more inclusive approach of digital literacy: social media writing at an older age

Reinhild Vandekerckhove, Sarah Bernolet, Astrid De Wit & Tanja Mortelmans

University of Antwerp, Department of Linguistics

E-mail: reinhild.vandekerckhove@uantwerpen.be

## Abstract

We present two complementary pilot studies on older adults' social media literacy. The first pilot discusses a survey among two generations of older adults, the second is based on family WhatsApp conversations between young adults and their parents. While the survey results show a restricted command of abbreviation strategies and emoji pragmatics, in spite of a clear predilection for emoji, the WhatsApp conversations point to a more elaborate exploitation of emoji functions by the parent generation. Still, older adults' practices clearly do not always align with those of the younger generations. Both lack of knowledge and dislike of specific online practices seem to be determining factors. The pilots constitute the starting point for a more extensive research project on seniors' social media literacy which in the end should lead to a more inclusive approach of present-day digital literacy.

**Keywords:** social media literacy, older adults, inclusion

## 1. Introduction

In spite of the rising importance of informal CMC for older generations, it has hardly been investigated how seniors interact via social media and to what extent they acquire and appropriate the conventions of informal online communication. While quite a lot of studies focus on the importance of online social networks for elderly people (e.g. Leist, 2013), (sociolinguistic) research on online writing has mainly focused on younger generations' practices, leaving older adults and seniors underrepresented or completely out of the picture. The present paper wants to address this gap by presenting two pilot studies<sup>1</sup> that serve as a run-up to a more extensive research project on senior's social media literacy designed by the authors of the present paper.

## 2. Pilot 1: digital literacy of 50+

The first pilot (Heremans, 2022) is based on a survey conducted in December 2021 and January 2022 among two generations of Flemish adults, i.e. people in their fifties and seventies. The study compares the two age groups in terms of their familiarity with informal online communication. More specifically, it investigates to what extent both generations know and actively use prototypical markers of the genre and to what extent they are aware of age and gender related preferences for these markers. In addition, participants' attitudes towards genre specific features were also examined.

Both groups were recruited via snowball sampling, starting with some acquaintances of the researcher. The survey was distributed through the online survey tool Qualtrics (participants who were less acquainted with digital tools received personal assistance when filling it out). Table 1 presents the number of participants for the two age and gender groups. The representation of both genders is perfectly comparable: women slightly outnumber men both

in the younger and the older group (resp. 53,3% and 53,8%). However, in view of the scope of the present paper, we exclusively focus on the variable age.

People in their	fifties	seventies	TOTAL
men	21	12	33
women	24	14	38
TOTAL	45	26	71

Table 1: participants survey 2022

Apart from a general section that questioned participants' use of digital devices and digital tools/media, the survey mainly comprised questions on the prototypical features and the socio-pragmatics of social media writing. For the identification of the typical markers of the genre, we rely on the generally acknowledged (implicit) maxims or principles of informal interactive online writing (e.g. Androutsopoulos, 2011: 149): i.e. the principles of (1) expressive compensation (which relates to the use of expressive markers like emoji and letter repetition), (2) orality (which entails the use of speech-like features) and (3) brevity or economy (which explains the use of all kinds of abbreviations and elliptic constructions).

First of all, participants had to report on their own use of these prototypical markers of the genre and their appreciation of (people using) these features. Furthermore, they had to analyze the functions of emoji in a range of social media posts, they had to decode typical acronyms (e.g.: *omg*) and they performed a gender and age detection task based on social media posts. Finally, they had to make up a fictitious happy birthday message for a young person they were close with (e.g. grandchild, niece or nephew) and send it via sms to the researcher. This small additional task was intended to supplement the reported language behavior with some actual language behavior, albeit not in an authentic setting. The survey was pretested on transparency and feasibility by three people in their fifties who were not

<sup>1</sup> Both pilot studies are based on MA-theses supervised by the first author of the present paper.

included in the final sample of participants. They needed about half an hour to complete the questionnaire. Nearly all the participants in their 50s turned out to have a smartphone and so did 92% of the seniors. 15% of the seniors, however, had never used a computer. WhatsApp was by far the most popular social media app for both age groups: it was used by 96,6% of the youngest group and 80,1% of the seniors. In terms of familiarity with the pragmatics of social media writing, the emoji task revealed some interesting patterns. For each emoji (in context) people could tick several potential functions described in layman's terms (e.g.: "the chatter makes clear his message should not be taken literally", "the chatter wants to express that they are on good terms with each other"). Answers were considered 'correct' if the emoji's main function, as identified by the researcher and her supervisor, was among the options that were ticked by the participants. While the younger group scored significantly better for this task ( $X^2=30.96$ ,  $p<0.0001$ ), we see a comparable pattern for both groups: participants score high when emoji are used in a most basic way, i.e. with a purely referential function (e.g., emoji picturing a dog when a dog is actually being referred to) or for the expression of emotions. However, whenever they are used as "indicators of illocutionary force" (Dresner & Herring, 2010) that serve for tone modification and face work (see e.g. Beißwenger & Pappert, 2019), both groups often have no clue. At the same time both the responses to the attitudinal questions and the birthday messages show that emoji are the only chatspeak features older adults really embrace. Strikingly, all participants of the youngest group added emoji to the birthday messages they had to make up, and so did 60% of the seniors. Conversely, the attitudinal responses show that both groups tend to dislike the use of speech-like features (e.g., final *t*-deletion in function words like *ni* 'not' and *da* 'that' instead of *niet* and *dat*). The birthday messages reflect the lack of appeal of these features: only in 13,6% and 5% of the messages produced by respectively the younger and the older group one or more markers of colloquial speech can be found. Moreover, seniors are unfamiliar with most chat abbreviations or acronyms. For the latter features we see a major discrepancy between both groups: seniors managed to decode only 21,2% of them, whereas the younger group scored 74,8% ( $X^2=112.95$ ,  $p<0.0001$ ). In line with this, the birthday messages hardly contain any abbreviations of whatever kind (two exceptions: one fifty-plus participant produced *ly* 'love you', another one *B-day* 'birthday').

### 3. Pilot 2: use of emoji in family WhatsApp

Rihoux (2021) investigated the frequency of emoji and the exploitation of their functional potential by two generations

in Flemish family interactions. The case study aimed at finding out whether age patterns persisted or levelled out in intimate intergenerational online communication. Unlike the first pilot, this study is based on spontaneous and authentic online communication produced outside a research context and focuses exclusively on just one marker of the genre: emoji.

In spite of the differences in the respective research designs, these studies are to some extent complementary, especially because the oldest generation in the data of Rihoux corresponds to the youngest generation of Heremans (i.e. people in their fifties). Rihoux collected a small but unique corpus of private WhatsApp group conversations of five families. The corpus contains posts produced by the five mothers and five fathers, aged 51-60, and their adult children, aged 18-25 (seven daughters and eight sons).<sup>2</sup>

	posts	tokens <sup>3</sup>
parents	1169	9110
children	1331	9816
TOTAL	2500	18926

Table 2: composition of the family WhatsApp corpus 2021

Surprisingly, the relative frequency of emoji is significantly higher in the posts of the parents than in those of their children ( $X^2=17.1$ ,  $p<0.0001$ ). Gender is no interfering variable.<sup>4</sup> Emoji represent respectively 3,12% and 4,26% of the tokens in the child versus parent corpus. While this suggests overaccommodation on the part of the parents, their children most probably also display accommodative behavior by suppressing the use of emoji to a certain extent when communicating with the older generation. The latter would be in line with Hilte et al. (2021), who found that the use of expressive markers (including emoji) in adolescent online chat is significantly less frequent in intergenerational communication with adults compared to intragenerational communication with peers.

For the analysis of the pragmatic exploitation of emoji, elaborating on Pappert (2017), seven emoji functions were distinguished, i.e. a distinction was made between emotionally expressive, evaluative, socio-empathetic (building relationships), tone modifying, emphatic, decorative and (purely) referential emoji. Most of these functions are well-represented both in the parent and in the child corpus, which means that, overall, parents and children seem to exploit the functional potential of emoji in similar ways. However, a closer analysis lays bare subtle differences that are most telling. More specifically, parents use significantly more emoji with an evaluative function ( $X^2=16.27$ ,  $p<0.001$ ) and they produce significantly more 'naked emoji' ( $X^2=8.35$ ,  $p<0.01$ ), i.e. isolated emoji that are not combined with a verbal reaction. The two of them are related, since parents mainly use naked emoji to make clear that their children are doing well and that they approve of

<sup>2</sup> The participants belong to the personal network of Anton Rihoux. They all gave consent for secure storage of and research on their anonymized data. Data are stored in research group CLiPS, University of Antwerp.

<sup>3</sup> Tokens are the result of splitting the text on whitespace. A token

can be a word, an emoji or isolated punctuation marks, e.g.: *hi !!!* contains two tokens.

<sup>4</sup> No gender differences were attested. This did not match our expectations, since generally speaking women tend to use more expressive markers than men, see e.g. Hilte et al., 2018.

their ‘actions’. While this seems to reveal an eagerness to connect with youths’ social media practices, these isolated evaluative emoji (especially the thumbs up) are not popular amongst youngsters. Research by Hilte et al. (2019: 28) revealed that posting an isolated 👍 is perceived as ‘unfriendly’ by adolescents. Youngsters tend to interpret this practice as a display of indifference rather than of enthusiasm, while the parents clearly want to convey the latter emotion. As such this practice of the parent generation presents a case of subjective accommodation and objective divergence, since parents most probably “perceive their behavior as convergent when, in fact, it is objectively divergent” (Dragojevic et al., 2016: 41).

#### 4. Conclusion: A call for further research

When comparing survey results mainly based on reported language behavior in social media contexts and related attitudes (pilot 1) with authentic social media data (pilot 2), some caution obviously is in order. Still, interestingly, the family conversations confirm the survey findings in two respects: they show that while older adults have appropriated emoji, they do not always seem to align with younger generations when it comes to emoji pragmatics. But then the question is to what extent they really want their online practices to align with ‘youthful conventions’. The survey results for instance clearly reveal that participants explicitly disapprove of speech-like writing, which is extremely common in adolescent CMC (see e.g. Hilte et al. 2020). Correspondingly, the senior survey participants generally do not integrate colloquial speech markers in the writing task, even though the intended addressee was said to belong to the youngest generations. This is in sharp contrast with the positive attitudes towards and eager use of emoji. Since it has been observed before that senior language users make linguistic novelties fit into their own systems (Anthonissen & Petré 2019), it seems not unlikely that they develop a kind of intermediate style by integrating particular genre conventions selectively and moderately, while at the same time relying on classical and more formal writing practices. Therefore, we plan more extensive research on how seniors reconcile their firmly entrenched writing habits with the potential of a ‘new’ genre. In the end this should lead to a more inclusive approach of social media literacy. The small-scale pilots presented here are but a modest first step towards achieving this.

#### 5. 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.
- Anthonissen, L. and Petré, P. (2019). Grammaticalization and the linguistic individual: new avenues in lifespan research. *Linguistics Vanguard* 5(2), <https://doi.org/10.1515/lingvan-2018-0037>.
- Beißwenger, M. and Pappert, S. (2019). How to be polite with emojis: a pragmatic analysis of face work strategies in an online learning environment. *European Journal of Applied Linguistics*, 7(2), pp. 225--253.
- Dragojevic, M.; Gasiorek, J. and H. Giles (2016). Accommodative Strategies as Core of the Theory. In H. Giles (Ed.), *Communication Accommodation Theory: Negotiating Personal Relationships and Social Identities Across Contexts*. Cambridge: Cambridge University Press, pp. 36--59.
- Dresner, E. and Herring, S.C (2010): Functions of the Non-Verbal in CMC: Emoticons and Illocutionary Force. *Communication Theory*, 20(3), pp. 249--268.
- Heremans, B. (2022). *De digitale geletterdheid van 50-plussers. Een vergelijkend onderzoek bij twee generaties naar de vertrouwdeheid met de principes van informele online communicatie*. MA thesis, University of Antwerp.
- Hilte, L.; Vandekerckhove, R. and Daelemans, W. (2018). Expressive markers in online teenage talk: a correlational analysis. *Nederlandse Taalkunde*, 23(3), pp. 293--323.
- Hilte, L.; Vandekerckhove, R. and Daelemans, W. (2019). Adolescents’ perceptions of social media writing: Has non-standard become the new standard? *European Journal of Applied Linguistics* 7(2), pp. 189--224.
- Hilte, L.; Daelemans, W. and Vandekerckhove, R. (2021). Interlocutors’ age impacts teenagers online writing style: accommodation in intra- an intergenerational online conversations. *Frontiers in Artificial Intelligence* <https://doi.org/10.3389/frai.2021.738278>
- Hilte, L.; Vandekerckhove, R. and Daelemans, W. (2020). Modeling adolescents’ online writing practices: the sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik*, 87(2), pp. 173--201.
- Pappert, S. (2017). Zu kommunikativen Funktionen von Emojis in der WhatsApp-Kommunikation. In M. Beißwenger (Ed.), *Empirische Erforschung internet-basierter Kommunikation*. Berlin&Boston: De Gruyter, pp. 175--212.
- Rihoux, A. (2021). *Emoji-gebruik in familiegesprekken. Hoe gaan verschillende generaties om met emoji en hun functioneel potentieel?* MA thesis, University of Antwerp.



# Phonetic Metaphor of Chinese Emojis: An Approach of Neologism Formation

Jiayi Zhou

Institute for Language Sciences (ILS), Utrecht University

E-mail: zhoujiayi0523@gmail.com

## Abstract

In Internet-mediated communication, emoji has gradually become a non-negligible element, and the visual writing system of language is also experiencing the impact of emoji. Neologisms have also emerged as a result, and one interesting way of creating new words is through phonetic metaphors. Chinese, with its unique character system and one-character-one-syllable feature, is more likely to produce emoji-related phonetic metaphors. For example, through phonetic metaphors, 🍌🍌 acquires the phonetic sound of 垃圾 “trash” *laji* and is used to refer to trash-like useless people. This paper explains that the instantiation of this phonetic metaphor approach is a two-way result; on the one hand, the need for expression cannot be directly satisfied by the existing emojis, and on the other hand, it is possible to extract phonetic materials from emojis. Moreover, the paper also argues with semantic and pragmatic evidence that these emoji expressions are neologisms rather than new calligraphic forms.

**Keywords:** Chinese emojis, Phonetic metaphor, Neologism

## 1. Introduction

Written languages are regarded as secondary towards spoken languages due to the fact that there are languages without writing systems (Radford et al., 2009). Therefore, the writing system is just a visual system in the service of the oral system. In recent years, emojis coexist more and more frequently with written characters in Internet-mediated communication, enriching the expression style of the writing system. However, the role and status of emojis in this specific context are still open to discussion.

One common view is that emojis in online texts act as gestures in spoken scenarios. According to Yu & Qin (2011), although sign languages are not taken into consideration by them, real communication is spoken language with gestures, and online communication is written language with emojis. They mapped emojis to gestures based on their common functions like exchanging feelings, attitudes and meanings and regulating the communicative atmosphere of real or virtual contexts. Furthermore, in terms of visual expression systems, emojis are going further than characters. They are generated from images rather than language. Even if we could describe emojis in words, there are actually no fixed syllables corresponding to these symbols. All interpretations are based on the understanding of the image or the conceptual reinforcement of particular input method editors (IME), and this understanding is also influenced by individual comprehension and cultural background. For example, when describing 🙏, there are possible divergent

interpretations such as *please, the symbol of please, the pattern of clasping hands, prayer, a person wanting someone to forgive him, etc.*

Yet, in the interaction between online emojis and specific languages, netizens, more or less depending on the language, do not stay in the zone of taking emojis only as gesture-like symbols. They are trying to embed them into characters. This phenomenon is particularly evident in Chinese social media. One notable tendency is that Chinese netizens use homophonic relations, i.e., phonetic metaphors, to give emojis fixed auditory properties, thus giving birth to neologisms. For example, 🍌🍌 has the same pronunciation as 垃圾 “trash” *laji*. This article aims to explore the generation mechanism and the effect of these emoji expressions, as well as to argue from a linguistic perspective that they are indeed neologisms.

The article mainly adopts a descriptive and explanatory approach and is structured as follows. Section 2 illustrates the rationale for the formation of these new words through examples and explains the special role played by Chinese syllables. Section 3 focuses on their semantic pragmatic effects. Section 4 summarizes and points out possible methods for subsequent research.

## 2. The Mechanism of Word Formation: High Adaptability of Phonetic Metaphor

### 2.1 The Process of Employing Phonetic Metaphor

First of all, emojis are functionally restricted in directly interpreting all the concepts and objects on their own. Han

(2017) uses the principles of Peirce's semiotics to classify emojis by iconicity, indexicality and symbolicity. Iconic forms have immediate relatednesses, such as 😊, 😢 and 🤖. Indexical forms emphasize the relationship between objects, such as ➡, ➡ and ✖. Symbolic forms are those established according to sociocultural customs or pre-existing rules, such as ❤, 🔥 and 💬. However, compared with all possible concepts and objects in one language, such as *genius*, *Manchurian Tiger* and *Metaverse*, existing groups of emojis are relatively straightforward and inevitably limited in number. If users cannot find an emoji that directly corresponds to a concept, i.e., iconicity, indexicality, and symbolicity do not work, but they still want to use emojis to create specific effects (as explained in later sections), they will have to find other ways to do so. One of the common methods is to turn to phonetic metaphors.

The use of phonetic metaphors is not accidental, especially when digging into the history of language evolution. In Ancient Egyptian, "phonetic metaphor can be said to have been the medium through which the language of new abstractions, utilizing phonetic metaphor for the representation of specifications such as proper names - the name of the king Narmer in his palace, the name of his sandal-bearer, and the name of his vanquished enemy." (Goldwasser, 2015, p.18) As for Chinese, one of the six main methods of Chinese character formation is the phonetic loan principle, which means that phonetic loan characters originally have no characters but get the characters from the homophonous ones (Xu, 1984[1815]).

In terms of the feasibility of applying phonetic metaphors to linguistic materials, emojis presented as images need to have relatively stable concepts first, and these concepts stand for an inventory of sounds. For example, there are possible ways to describe the emoji 🐵 such as *monkey*, *brown monkey*, *squatting*, etc. The most central concept can be summarized as *monkey*, and the relatively marginal concepts are *brown* and *squatting*. These concepts make up the concept cluster of this emoji. In Chinese, these concepts can correspond to some pronunciations, such as 猴子 "money" *houzi*, 猴 "money" *hou*, 棕 "brown" *zong*, 蹲 "squat" *dun*, etc. When a new concept 好 "good" *hao* requires an emoji but cannot be directly linked to existing ones, users can take advantage of phonetic metaphor to find

an emoji with similar pronunciation. Since *hou* is pronounced similarly to *hao*, although not exactly the same, it is possible to associate 🐵 with 好 *hao*. This emoji thereby acquires a new meaning, namely *good*, by metaphorical mapping.

Another example can be 🌶🗑, which means 垃圾 "trash" *laji*. The first emoji 🌶 can be described as chilli, spicy, red pepper, red, etc., establishing an inventory of sounds in Chinese with its focus on 辣 "spicy" *la*, 椒 "pepper" *jiao* and 红 "red" *hong*. The second emoji 🗑 can be described with Chinese pronunciations as 公鸡 "cock" *gongji*, 鸡 "chicken" *ji*, 鸡头 "the head of the chicken" *jitou*, 鸡冠 "cockscorn" *jiguan*. Apparently, 🌶🗑 represents 垃圾 "trash" *laji* by taking 垃 *la* from 辣 *la* in 🌶 and taking 圾 *ji* from 鸡 *ji* in 🗑.

Hence, it can be seen that the realization of phonetic metaphor is a two-way process. On one hand, there is a lack of emojis that are directly related to a concept. This prompts people to search for indirectly related emojis, leading them to use phonetic metaphors. On the other hand, most emojis have a relatively fixed set of concepts with which their pronunciations can be associated.

## 2.2 Chinese Syllables and the Adaptability of Phonetic Metaphors


Li (2005) suggests that compared with English, there are more phonetic metaphors in Chinese, due to the difference in their syllabic structures. This view is also supported by Hu (2021), who argues that Chinese has mostly monosyllabic characters and that polysyllables are more difficult for phonetic metaphors than monosyllables. This article continues their view on Chinese syllables, but also attempts to further refine the logic chain of interpretation in terms of phonology and prosody and highlights the connection of emoji to this aspect.

Chinese is regarded as a tone language and every syllable has its intrinsic lexical tone (Chao, 1930), which makes it much easier to discriminate one syllable from another. The writing system of Chinese characters also exhibits this syllabic feature to a great extent. Its syllabic independence and recognizability play an important role in triggering phonetic metaphors. Here, we break it down into several specific characteristics.

First, Chinese syllables and characters are generally in the relationship of one-to-one mapping. This correspondence also helps Chinese netizens to find the targeted emoji. *Laji* is a disyllabic word, so they do not need to find an emoji that corresponds to *laji* as a whole directly, they can split *laji* into two syllables, *la* and *ji*, for mapping separately.

Second, there is a limited number of syllables in Mandarin Chinese. While English has about 15,831 possible syllables (Barker, 2009), Chinese has only 418 syllables regardless of 4 tones (Su & Lin, 2006). Since Chinese has more than 90,000 characters and about 7,000 commonly used characters, there are inevitably many homophones in Chinese. This provides a favourable ground for phonetic metaphor, which is also the case for emoji's phonetic metaphor, as they need to resort to the Chinese syllable database for the purpose of realizing phonetic metaphors.



Third, the similarity between Chinese characters and emojis. Both Chinese characters and emoji are encoded by Unicode and occupy the same glyph width in the Internet text. Although Chinese characters are abstract linguistic symbols, they still have more graphic and symbolic qualities than Latin letters. This also allows Chinese netizens to treat emojis as pseudo-Chinese characters from a cognitive point of view.


Forth, syllable matching has a high degree of flexibility in Chinese phonetic metaphors. Informal forms of phonetic expressions are recognized in the online context, thus close but not identical pronunciation is quite feasible. Tones, vowels, and consonants can all be different, as long as the similarity is maintained. In addition, dialectal pronunciation is also popular on the Internet. For example, the mapping of  *hou* and *hao* is mediated by Cantonese rather than Mandarin.




### 3. Semantic and Pragmatic Evidence for Emoji-based Neologisms

Such emoji expressions can easily be seen as different ways of writing the same word, such as different calligraphic systems of Chinese characters like Xing Shu, Kai Shu, etc., which are only artistic expressions. However, this paper argues that they are not calligraphic, but they indeed constitute neologisms. This section attempts to provide semantic as well as pragmatic evidence for the proposal.

### 3.1 Semantic Narrowing

The phonetic metaphorically mapped emoji does not replace the full semantic meaning of the original Chinese character, but only a part of its meaning, i.e., semantic narrowing happens. The emoji , although mapped with *hao*, can only function as *hao* in the case of an affirmative response, as in example (1). Xiaowang sends a text message to Aming asking him if he wants to watch a movie tomorrow, and when Aming wants to watch it, he can reply with either a character 好 *hao* or an emoji  on social media.

However, when *hao* is used as an adjective or an adverb of degree, using  instead of *hao* is abrupt and ungrammatical, as seen in examples (2) and (3).

- (1) - 小王:            明天        看            电影        吗?
- |                 |                 |            |                 |           |
|-----------------|-----------------|------------|-----------------|-----------|
| <i>xiaowang</i> | <i>mingtian</i> | <i>kan</i> | <i>dianying</i> | <i>ma</i> |
| Xiaowang        | tomorrow        | watch      | movie           | INT       |
- Xiaowang: Do you want to watch a movie tomorrow?
- 阿明:            好。 / 
- |              |                |
|--------------|----------------|
| <i>aming</i> | <i>hao/hou</i> |
| Aming        | Yes.           |
- Aming: Yes!
- (2) a. 这棵            树            好            高!
- |              |            |            |            |
|--------------|------------|------------|------------|
| <i>zheke</i> | <i>shu</i> | <i>hao</i> | <i>gao</i> |
| this-CL      | one-CL     | very       | tall       |
- This tree is very tall.
- \*b. 这棵            树                        高!
- |              |            |            |            |
|--------------|------------|------------|------------|
| <i>zheke</i> | <i>shu</i> | <i>hou</i> | <i>gao</i> |
| this-CL      | one-CL     | very       | tall       |
- This tree is very tall.
- (3) a. 小李            人            非常            好。
- |               |            |                 |            |
|---------------|------------|-----------------|------------|
| <i>xiaoli</i> | <i>ren</i> | <i>feichang</i> | <i>hao</i> |
| Xiaoli        | person     | extremely       | one-CL     |
- Xiaoli is a very nice person.
- \*b. 小李            人            非常            .
- |               |            |                 |            |
|---------------|------------|-----------------|------------|
| <i>xiaoli</i> | <i>ren</i> | <i>feichang</i> | <i>hou</i> |
| Xiaoli        | person     | extremely       | one-CL     |
- Xiaoli is a very nice person.

### 3.2 Indirectness and Positivity

The most important pragmatic features in the phonetic metaphor from Chinese characters to emoji are indirectness and positivity.

Indirectness is reflected in the mapping of phonetic metaphors and the speech code-switching from characters to visual images (emojis). This makes expressions more subtle and euphemistic and can avoid profanities, weaken insults, and moderate strong words. The word *laji* “trash” is an insult in itself, and is used in communication to accuse the other person of being useless in some way, like trash, but the presence of 🌶️🍷 avoids the direct use of the original characters. 牛屌 “awesome” *niubi*, which originally means the genitalia of a cow, is used extensively on the Chinese Internet to praise someone for being great, capable and awesome. In order to avoid profanity, 牛屌 is often written 牛逼 *niubi*, 牛 B *niu-B* and 牛 X *niu-X* on the Internet. Its corresponding emoji expression is 🐮🍺, which consists of the icon 🐮 and the symbol 🍺 linked by phonetic metaphor, respectively. One of the core meanings of 🍺 is beer, and netizens took the 啤 “beer” *pi* of 啤酒 “beer” (beer and alcohol, literally) *pījiu* to map to the original word *bi*. It is worth noting that this phonetic metaphor also switches the consonant, changing from the unaspirated [p] to the aspirated [p<sup>h</sup>]. This also exemplifies the euphemistic role played by indirectness.

Positivity is related to indirectness. Emojis create a stronger visual effect while weakening the potential conflict, and their motivation to communicate positively reinforces the positive side of the concept, though this reinforcement of positivity is not absolute. More specifically, first, emojis visualize meaning. On the premise of the same concept, images have the advantage of rich colours and diverse lines compared to characters. The reinforcement of the visual stimulus corresponds to the reinforcement of the emotional concept. Secondly, by rejecting the use of ready-made characters and painstakingly using emojis indirectly, it manifests a positive purpose of enriching forms of communication. The user chooses the expression he finds most interesting in the series of words *trash*, which is 🌶️🍷. This also explains why 🌶️🍷 is more playful than reproachful. The playfulness also limits the situations in which it can be used; for example, it would be inappropriate to use 🌶️🍷 in a formal rebuke.

#### 4. Conclusion

This paper focuses on the mechanisms and specific features of new Chinese emoji words formed by means of phonetic metaphor. The semantic narrowing, indirectness and positivity properties that these expressions carry also prove that they are indeed neologisms. From a more distant picture, the opinion that the writing system is secondary to the speaking system is challenged by the emergence of visual emoji neologisms.

In the subsequent study, a systematic statistical analysis of the Chinese web corpus can be conducted so as to have a more comprehensive grasp and understanding of this phenomenon.

#### 5. Reference

- Barker C. (2009). *How Many Syllables Does English Have?* <https://web.archive.org/web/20160822211027/http://sem.arch.linguistics.fas.nyu.edu/barker/Syllables/index.txt> Retrieved 2023-05-01.
- Chao, Y. (1930). A system of tone letters. *Le Maître Phonétique*, 45, pp. 24–27.
- Goldwasser, O. (1995). *From Icon to Metaphor: Studies in the Semiotics of the Hieroglyphs*. Fribourg: University Press.
- Han, W.W. (2017). Fuhaoxue shiyuxia emoji biaoqing fuhao de biao Zheng yu yi yi [The representation and meaning of emojis in the semiotic perspective]. *Xiju Zhi Jia* [Home Drama], 14, p.286.
- Hu, Z.Q. (2021). Qianxi yuyin yinyu de youguan texing [Analyzing relevant features of phonetic metaphor]. *Zhongguo Waiyu* [Foreign Languages in China], 18(04), pp. 26-31.
- Li, H. (2005). Yuyin yinyu chutan [Some reflection on phonetic metaphors]. *Sichuan Waiyu Xueyuan Xuebao* [Journal of Sichuan International Studies University], 3, pp.70-74.
- Radford, A.; Atkinson, M.; Britain, D.; Clahsen, H. and Spencer, A. (2009). *Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Su, X.C., Lin, J.Z. (2006). Putonghua yinjiesshu ji zaijiliang de tongji fenxi: jiyu xiandai hanyu cidian zhuyin cailiao [A statistic analysis of the counts of Putonghua syllables and the distribution of characters among syllables: A case study of the phonetic notation materials in The Contemporary Chinese Dictionary]. *Zhongguo Yuwen* [Chinese Language]. 3. pp.274-284+288.
- Xu, S. (1984[1815]). *Shuo Wen Jie Zi* [Dictionary of Words Explained and Analyzed]. Shanghai: Shanghai Guji Chubanshe [Shanghai Classics Publishing House].
- Yu, G.W., Qin, Y. (2011). Yuyanxue shijiaoxia de wangluo biaoqing fuhao chutan [A study on Internet emoticons from the linguistic perspective]. *Zhongguo Shehui Kexueyuan Yanjiushengyuan Xuebao* [Journal of Graduate School of Chinese Academy of Social Sciences]. 1. pp.130-135.

## Author index

- Alanzi, Aida 13  
Aminoff, Johanna 13  
Anastasi, Selenia 23  
Babayode, Aminat 3  
Bai, Tianyi 29  
Barta, Mindi 140  
Bartsch, Marie-Louise 7  
Beißwenger, Michael 33  
Bernolet, Sarah 187  
Biemann, Chris 23  
Bosman, Laurens 3  
Bothe, Laura 39  
Chan, Nicole 3  
Coats, Steven 51  
Cotgrove, Louis 55  
Denis, Derek 13  
De Wit, Astrid 187  
Doudot, Liana 96  
Drylie, Daniel 140  
Ehret, Katharina 3  
Elwert, Frederik 15  
Erten, Selcen 60  
Fabian, Annamaria 65  
Ferber, Anne 73,112  
Fernández Polo, Francisco Javier 78  
Finkelstein, Shir 83  
Fischer, Tim 23  
Flinz, Carolina 86  
Fong, Ivan 3  
Frenken, Florian 91  
Gredel, Eva 33,86  
Gupta, Prakhar 96  
Hamdi, Antonia 118  
Harris, Noel 3  
Helenius,, Teemu 102  
Herzberg, Laura 86,131  
Hewton, Alissa 3  
Ho-Dac, Lydia-Mai 171  
Jeon, Sangwan 108  
Kibisova, Elizaveta 9  
Knierim, Aenne 10  
Krause, André Frank 112  
Krause, Andre Frank 73  
Lemnitzer, Lothar 118  
Linardi, Michele 45  
Longhi, Julien 45  
Lorés, Rosa 124  
Loup, Romain 96  
Lüngen, Harald 131  
Machado Carneiro, Bruno 45  
Mäkinen, Martti 136  
McCullough, Rachel 140  
Mortelmans, Tanja 187  
Moshnikov, Ilia 142  
Mostovaia, Irina 7  
Netz, Hadar 83  
Pabst, Katharina 13  
Piroh, Anastasiia 148  
Pitsch, Karola 73,112  
Poudat, Céline 171  
Rebhan, Lena 33  
Reid, Danica 3  
Reimann, Sebastian 15  
Rodenhausen, Lina 15  
Rykova, Eugenia 142  
Sancho-Ortiz, Ana Eugenia 154  
Scheffler, Tatjana 2,15,17  
Schneider, Florian 23  
Schneider, Ulrike 160  
Seemann, Hannah J. 17  
Shahmohammadi, Sara 17  
Smith, Daniel 140  
Stede, Manfred 17  
Steinsiek, Sarah 18,33  
Stvan, Laurel 166  
Taboda, Maite 3  
Tanguy, Ludovic 171  
Tayib, Raisa 13  
Thoma, Ralia 176  
Triebl, Eva 182  
Trost,, Igor 65  
Vandekerckhove, Reinhild 187  
Watteler, Oliver 160  
Wong, Rebekah 3  
Xanthos, Aris 96  
Yudytska,, Jenia 20  
Zang, Yinglei 21  
Zhou, Jiayi 190

## Keyword index

- adolescent students 176
- affordance 20
- annotation guidelines 45
- anonymization 112
- backchanneling 13
- black history 10
- black history month 10
- chat communication 29
- chats 96
- chinese emojis 190
- cmc 2,136
- cmc corpora 23
- computer-mediated communication 3,18,21,33,55
- computer mediated discourse 140
- conase 51
- conspiracy theories 39
- content moderation 102
- cooperative learning 18
- corpora 33
- corpus 96
- corpus analysis 10,73
- corpus-building 2
- corpus compilation 136
- corpus linguistics 3,9,51,55,86,108,140,166
- corpus pragmatics 182
- corpus scale 20
- corpus workflow 73
- cxg 39
- dash 51
- data collection 17,18
- data publication 160
- data sharing 160
- deep learning 45
- de-identification 96
- deontic authority 21
- device 20
- dialect 1
- digital scientific communication 124
- disability discourse 65
- discourse analysis 2,39,65,108,118,140
- discourse annotation 17
- discussion forums 78
- ehrlich 118
- ellipsis' 7
- ellipsis points 33
- email 21
- english language 140
- epistemic management 182
- expert identity 182
- expert image 78
- extreme behaviours 171
- face detection 112
- fake news 9
- finnish-swedish 136
- 'first-person singular 7
- formants 51
- french 96
- geometric multivariate analysis 91
- german language 118
- greeklish 176
- health communication 166
- hyperlinks 154
- im 96
- incels 23
- inclusion 187
- instant messaging 96,136
- intensification 55
- interaction 55
- interactional norms 13
- interactive unit 118
- interdiscursivity 148
- intertextuality 124,148
- karelian 142
- language and gender 140
- language corrections 83
- language of extremism 140
- lexical analysis 65
- linguistic analysis 9
- linguistic practices in cmc 136
- machine learning 45
- manual annotation 60
- memes 166
- metaphor 15
- microlinguistic features 20
- minority language recognition 142
- mixed-methods 10
- mixed methods 20
- mocoda2 29
- multilingual corpora 23,140
- multilingualism 136
- multimodal interaction 73
- multimodality 18,23,102,136,148,154
- neologism 190
- older adults 187
- online commenting 83
- online extremism 23
- online forums 15



online hate speech 108  
 online interaction 171  
 parallel corpus 17  
 parameter optimization 112  
 peer-advice 78  
 phonetic metaphor 190  
 phonetics 51  
 platform affordances 102  
 political discourse 83,108  
 popularization 124  
 pragmatics 18,33,55  
 propaganda 108  
 questions under discussion 17  
 recontextualization 124  
 reddit 15,91  
 referencing strategies 131  
 register analysis 3  
 register variation 91  
 reply function 29  
 reply relations 131  
 rhetorical structure theory 17  
 russian language 9,140  
 science dissemination 154  
 semiotic resources 148  
 semiotics 55  
 sentiment analysis 65  
 social deixis 86  
 social media 1,9,102,160  
 social media literacy 187  
 sociolinguistics 108  
 speech acts 166  
 standard language ideology 83  
 subject pronoun ich ("i") 7  
 sud classification 45  
 telegram 39  
 text dispersion keyword analysis 60  
 text messaging 33  
 thread structure 15  
 tiktok 102  
 topic modeling 10  
 transfer learning 45  
 transliteration practices 176  
 turkish web registers 60  
 twitter 108,142,154,160  
 variation 2  
 variationist sociolinguistics 13  
 videoconferencing 13  
 video film review 148  
 visibility 102  
 wechat 21  
 whatsapp 29,96  
 whatsapp chats 7  
 wikipedia 33,86  
 wikipedia talk pages 131,171  
 workflow 112  
 youth language 1,55  
 youtube 51