

Annamária Fábián/Igor Trost (eds.)

# Impulses and Approaches to Computer-Mediated Communication

## Proceedings of the 12<sup>th</sup> International Conference on Computer Mediated Communication and Social Media Corpora for the Humanities



UNIVERSITÄT  
BAYREUTH

4<sup>th</sup>-5<sup>th</sup> September 2025





Annamária Fábíán/Igor Trost (eds.)

# **Impulses and Approaches to Computer-Mediated Communication**

**Proceedings of the  
12<sup>th</sup> International Conference on  
Computer Mediated Communication and  
Social Media Corpora for the Humanities**

**CMC 2025**

**4<sup>th</sup>-5<sup>th</sup> September 2025**

**University of Bayreuth, Germany**

Impulses and Approaches to Computer-Mediated Communication. Proceedings of the 12th International Conference on Computer Mediated Communication and Social Media Corpora for the Humanities, CMC 2025, 4th-5th September 2025, University of Bayreuth, Germany.

Editors: Annamária Fábián, Igor Trost

Published by University of Bayreuth

Conference website: <https://www.cmc2025.uni-bayreuth.de/en/index.html>

DOI: 10.15495/EPub\_UBT\_00008705

ISBN: 978-3-00-085196-4

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





# Table of Contents

<b>I.</b>	<b>Preface and New Impulses and Approaches to Computer-Mediated Communication</b>	<b>9</b>
<b>II.</b>	<b>Committees</b>	<b>13</b>
<b>III.</b>	<b>Keynotes</b>	<b>15</b>
(1)	<b>Studying Discourse in Social Media: Challenges &amp; Opportunities</b> Stephanie Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)	17
(2)	<b>Studying language and identity in a corpus of computer-mediated communication with (and without) sociodemographic metadata</b> Gavin Brookes (Lancaster University, UK)	22
<b>IV.</b>	<b>Talks</b>	<b>27</b>
(3)	<b>Towards a new Curation Workflow for the CMC Corpora Resource Family</b> Egon W. Stemle, Lionel Nicolas (Eurac Research, Italy), Alexander König (CLARIN-ERIC, The Netherlands)	29
(4)	<b>HopeEmo: A Bilingual Social Media Corpus for Emotion and Hope Speech Analysis</b> Wajdi Zaghouani (Northwestern University, Qatar), Md. Rafiul Biswas (Hamad Bin Khalifa University, Qatar)	33
(5)	<b>Tracking Ephemerality in YouTube Comments: Towards Methods for Building Dynamic CMC Corpora</b> Yining Wang, Katrin Weller (Leibniz Institute for the Social Sciences, Germany)	36
(6)	<b>Deepfakes in Criminal Investigations: Interdisciplinary Research Directions for CMC Research</b> Lorenz Meinen, Astrid Schomäcker, Timo Speith, Lena Kästner, Christian Rückert (University of Bayreuth, Germany), Niklas Kühl, Stefanie Wiedemann (University of Bayreuth /FIM Forschungsinstitut, Germany), Markus Hartmann (ZAC NRW, Germany)	40
(7)	<b>CRIME: The Corpus of Recorded Investigative, Media, and Evidence-based Proceedings</b> Steven Coats (University of Oulu, Finland), Dana Roemling (University of Birmingham, UK)	45
(8)	<b>Dimensions of Drivel in German Telegram Posts: Manual Annotation and Predictive Power</b> Andreas Blombach, Evert, Stephanie, Linda Havenstein, Philipp Heinrich (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)	50

(9)	<b>A Case Study on Annotating and Analysing Situation Entity Types in Reddit Discussions on Democracy</b>	55
	Hanna Schmück, Annemarie Friedrich (University of Augsburg, Germany), Michael Reder (Munich School of Philosophy, Germany), Katrin Paula (Technical University Munich, Germany)	
(10)	<b>Annotating and Extracting Suggestive Language in CMC: A Linguistically Grounded Corpus and NLP Approach</b>	60
	Omnia Zayed, Sampritha Manjunath, Paul Buitelaar (University of Galway, Ireland)	
(11)	<b>Beyond names: how to label gender automatically in CMC data?</b>	66
	Pasi Fränti, Juhani Järviö, Mehrdad Salimi, Irene Taipale, Mikko Laitinen, Rahel Albicker, Chunyuan Nie, Masoud Fatemi, Paula Rautionaho (University of Eastern Finland, Finland)	
(12)	<b>"I expected better from you, Mr. King": Feminist resistance and reader critique in the subreddit r/MenWritingWomen</b>	72
	Marie Flesch (Université de Lorraine), Heather Burnett (Université Paris Cité, France)	
(13)	<b>OMG! Why discourse markers thrive in interactive social media writing</b>	78
	Reinhild Vandekerckhove (University of Antwerp, Belgium)	
(14)	<b>Emoji and Emoticon Use in Online Dating Profiles and Chats: A Corpus Study into Functions and Categories</b>	82
	Lieke Verheijen (Radboud University, The Netherlands), Tess van der Zanden (Utrecht University, The Netherlands)	
(15)	<b>"Tinder is overrated": Neoliberal Affective Economies in an Italian Incel Forum.</b>	88
	Selenia Anastasi (University of Rome La Sapienza, Italy), Maria Natasha Fragalà (University of Catania, Italy)	
(16)	<b>Modelling the Interaction Space of Twitch: A Multimodal Framework for Corpus Structuring and Analysis</b>	94
	Ariane Julie Robert (Università degli studi di Salerno, Italy)	
(17)	<b>Strategic Transparency or Deliberate Ambiguity? A Multimodal Analysis of Airline CSR Communication on LinkedIn</b>	99
	Fabiola Notari (University of Modena and Reggio Emilia, Italy)	
(18)	<b>Emerging digital discourse traditions: A contrastive analysis of ther/todayilearned subreddit and its German and French counterparts</b>	104
	Dominique Dias (Sorbonne Université, France)	
(19)	<b>Evaluating Different Methods for Building Specialized Corpora: A Case Study on the German Discourse on AI</b>	109
	Bruno Brocai, Janine Dengler (University of Heidelberg, Germany)	
(20)	<b>The most common features of the Albanian language used in computer-mediated communication – an overview based on corpus data</b>	115
	Besim Kabashi (Eberhard Karls Universität Tübingen, Germany)	

<b>V.</b>	<b>Poster Abstracts</b>	119
(21)	<b>Augmenting the CoWoYTP1Att Corpus with Emotion and Hate Speech Annotations: A Study on the Relationship with Appraisal Theory</b> Valentina Tretti-Beckles, Adrian Vergara-Heidke (Potsdam University, Germany)	121
(22)	<b>Methodology for Developing a Fact-Checked News Dataset in Norwegian Bokmål for Fake News Detection (The Fakespeak-NOR Corpus)</b> Aleena Thomas, Silje Susanne Alvestad (SINTEF AS, Norway)	122
(23)	<b>Building and querying Wikipedia discussion corpora using KorAP</b> Eliza Margaretha, Harald Lungen, Nils Diewald, Marc Kupietz, Rameela Yaddehige (Leibniz Institute for the German Language, Germany)	123
(24)	<b>“Prompt as Culture”: A Cross-linguistic Analysis of Prompt Engineering Discourse on Chinese and English Social Media</b> Xiaomin Zhang (University of Modena and Reggio Emilia, Italy)	125
(25)	<b>The Biased Language Taxonomy</b> Costanza Marini, Elisabetta Jezek (University of Pavia, Italy)	126
(26)	<b>Diversifying Meaning in a Viral Age: The Case of 'Demure' on Social Media</b> Haruka Nishiyama (Keio University, Japan)	127
(27)	<b>Discursive Polarisation and the (Non-)Binary Spectrum: Social Media Debate on Gender Diversity</b> Andressa Costa (Karlsruhe Institute of Technology, Germany)	128
(28)	<b>Gender differences in Chinese sensory adjectives: A corpus-based study of food videos on Bilibili</b> Mingyu Liu (The Hong Kong Polytechnic University, Hong Kong)	129
(29)	<b>Emotional Expression in Text-Based Communication: An Analysis of Online Mentoring for Girls in STEM</b> Claudia Uebler, Albert Ziegler, Heidrun Stoeger (University of Regensburg, Germany)	130
(30)	<b>Comparative Analysis of Comments on Feminism on Hupu and Xiaohongshu: A Text Mining Approach</b> Mingyu Liu (The Hong Kong Polytechnic University, Hong Kong)	131
(31)	<b>Metapragmatic Perspectives on Autistic Digital Communication: A Corpus-Assisted Analysis of Self-Reported Practices</b> Nelya Koteyko (Queen Mary University of London, UK)	133

(32)	<b>(A)I Can Empathize with You: Analyses of Linguistically Performed Empathy and Human Identity of Large Language Models in Psychotherapeutic Settings</b> Florina Züllli (University of Zurich, Switzerland)	134
(33)	<b>The Positive Pulse: The Hidden Language of Scientific Social Media</b> Cansu Akan, Sasha Genevieve Coelho (Chemnitz University of Technology, Germany)	135
(34)	<b>Science Communication in Science Slams</b> Johanna Vogel (Leibniz Institute for the German Language, Germany)	136
(35)	<b>A Corpus-Based Appraisal Analysis of English-Language Social Media Discourse on Chinese and Italian Operas</b> Lei Liang (University of Modena and Reggio Emilia, Italy)	137
(36)	<b>Decoding Business German: A Corpus-Based Lexical and Morphological Analysis of Contemporary Job Advertisements</b> Kristina Krcmarevic Bogdanovic, Kristina Ilic (University of Belgrade, Serbia)	138
<b>VI.</b>	<b>Training Session with Stephanie Evert</b>	141
(37)	<b>Reading concordances with algorithms</b> Nathan Dykes, Stephanie Evert, Michaela Mahlberg, Alexander Piperski (Friedrich-Alexander-Universität Erlangen-Nürnberg)	143

# **I. Preface**

## **and New Impulses and Approaches to Computer-Mediated Communication**

Following the excellent exchange at prior editions of the CMC-conference series, we are delighted to present the proceedings of the 12th edition of the *International Conference on Computer-Mediated Communication and Social Media Corpora* (CMC2025). The conference mainly focuses on data collection, annotation, and corpus analysis from computer-mediated communication and social media. The conference also provides a framework for scientific exchange on methods of data processing and sustainable data infrastructures.

The CMC 2025 would like to serve the CMC-community to investigate a wide range of language-centered studies in Computer-Mediated Communication and social media, drawing from linguistics, philology, communication sciences and data science, with research questions stemming from corpus and computational linguistics, computational science, language technology, text technology, and machine learning. This year CMC-edition also enables exchange between the aforementioned disciplines on the one hand and data sciences as well as social sciences in general on the other. In addition, keeping up with social and language change, this conference also highlights communication-related questions of social and linguistic-related diversity, participation, and inclusion. The 12th Conference on CMC and Social Media Corpora is held at the Chair of German Linguistics at the University of Bayreuth (Germany) on September 4th and 5th 2025.

This volume includes two keynote papers, 18 accepted talk papers, and the abstracts of the 16 posters presented at CMC 2025 in Bayreuth. Each contribution underwent an anonymous double peer-review process by scientists at the CMC-scientific committee. The contributions will be presented in two sessions (including poster presentations) and in plenum. The talks and the poster presentations discuss a broad range of topics, ranging from CMC-corpus construction to corpus analysis including methodological discussions and inter- and multidisciplinary co-operation with other scientific fields essential to research on Computer-Mediated Communication.

The two keynote talks are held by Prof. Dr. Stephanie Evert (Friedrich-Alexander-University of Erlangen-Nürnberg, Germany) and Dr. Gavin Brookes (Lancaster University).

Prof. Dr. Stephanie Evert is Chair for Corpus- and Computational Linguistics at Friedrich-Alexander-Universität Erlangen-Nürnberg and elected member of several high-profile scientific organizations and institutions such as the Bavarian Academy of Science and the German Research Council. She has published numerous high-quality papers in outstanding international journals across linguistics, e. g. in computational linguistics, corpus linguistics, computer-mediated communication, and discourse analysis. She also leads several excellent scientific projects. One of them is the project *Reading Concordances in the 21st Century* supported by the Arts and Humanities Research Council (AHRC) and the Deutsche Forschungsgemeinschaft (DFG). Information on the project can be found here: <https://www.dhss.phil.fau.eu/research/current-projects/reading-concordances-in-the-21st-century-rc21/>. Stephanie Evert will deliver the first keynote *Studying Discourse in Social Media: Challenges*

& *Opportunities*'. More information on Prof. Evert's research activities is listed as follows: <https://www.linguistik.phil.fau.de/person/prof-dr-stephanie-evert/>

Dr. Brookes (Department: School of Social Sciences at Lancaster University) is Reader in Linguistics and UKRI Future Leader Fellow with an interest in corpus linguistics, discourse analysis and health communication. He is associate Editor of the International Journal of Corpus Linguistics (John Benjamins), co-Editor of the Corpus and Discourse book series (Bloomsbury, with Michaela Mahlberg) as well as co-Editor of the Critical Discourse Studies (Cambridge University Press, with Veronika Koller). In addition, Gavin Brookes is fellow of the Royal Society for Arts, Manufactures and Commerce. (More scientific information on Dr. Brookes' research activities can be found as follows: <https://www.lancaster.ac.uk/social-sciences/people/gavin-brookes.>)

Gavin Brookes is the PI of his research project '*Public Discourses of Dementia: Challenging Stigma and Promoting Personhood*' (funded £1 million by UKRI) and he will deliver one of the two keynote talks on '*Studying language and identity in a corpus of computer-mediated communication with (and without) sociodemographic metadata*'.

In addition to the keynote talks, the presentations of the talks, and the poster presentations, the conference contributes to community-building and training of new methods. Participants are invited therefore to attend a presentation by CLARIN aiming at the construction of a new network as well as a tutorial session by our keynote Stephanie Evert. Prof. Evert will give a tutorial on '*Reading concordances with algorithms*', developed by Nathan Dykes, Stephanie Evert, Michaela Mahlberg, and Alexander Piperski (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany).

Our aim is to provide valuable new theoretical and methodological insights into CMC. We have endeavored to highlight the great social relevance of CMC research and to focus additionally on diversity, participation, and inclusion in Computer-Mediated Communication. The commitment of CMC research to society, science, and social diversity has provided CMC with new impetus for years and will therefore be discussed in many ways including the keynote talks and also numerous lectures and poster presentations this year. The language-related features of diversity, inclusion, and participation in CMC present linguistics with theoretical and methodological challenges that we can only overcome through interdisciplinary and multidisciplinary collaboration with the social sciences. We are very pleased to host such interdisciplinary and multidisciplinary presentations, especially with data science, law, and philosophy, at CMC 2025.

We would like to thank our colleagues who contributed to the conference and to this volume with their talks and posters. We would also like to express our gratitude to the members of the steering committee (Dr. Steven Coats, Prof. Dr. Julien Longhi, Prof. Dr. Reinhild Vandekerckhove, and Dr. Lieke Verheijen) and the international scientific committee. We thank the University of Bayreuth for hosting us and are very grateful to Prof. Dr. Karin Birkner (Chair for German Linguistics at the University of Bayreuth) for her organizational support. Fábíán also thanks the Bavarian Research Institute for Digital Transformation Munich (bidt) at the Bavarian Academy of Science, of which she is appointed member of, as well as the Bavarian Ministry of Science, Research and Art for their financial support since the conference is co-financed by funds that Fábíán received from the two institutions for her

research project ,*The communicative realization of inclusion for people with disability in social media*' and for the support of her academic career as one of the seven post-docs from all research fields across Bavarian universities appointed to the Bavarian Research Institute for Digital Transformation in 2022 (more information here: <https://www.gl.uni-bayreuth.de/de/team/A-Fabian/index.php>). Last, but not at least, we thank Vanessa Tschöртner, student assistant in Fábián's research project, for her high-level of commitment and great organizational support of the CMC 2025.

The CMC 2025-edition would like to foster inspiring exchanges. In addition, it aims at a significant contribution for strengthening the international CMC-community including excellent scientists, who work on corpus building, data collection, data annotation, corpus analysis, and further methodological and theoretical implications using social media corpora and further corpora essential to Computer-Mediated Communication for collaborative research and infrastructures in the humanities and beyond.

Bayreuth, 31st of August 2025

On behalf of the organizing committee

Annamária Fábián and Igor Trost





## **II. Committees**

### **Local Organizing Committee**

Dr. Annamária Fábián (University of Bayreuth/Bavarian Research Institute for Digital Transformation at the Bavarian Academy of Science)

Prof. Dr. Igor Trost (Alpen-Adria University Klagenfurt/University of Passau)

### **Scientific chairs**

Dr. Steven Coats (University Oulu)

Dr. Annamária Fábián (University of Bayreuth)

Prof. Dr. Julien Longhi (CY Cergy Paris University)

Prof. Igor Trost (Alpen-Adria University Klagenfurt/U. of Passau)

Prof. Reinhild Vandekerckhove (University of Antwerp)

Dr. Lieke Verheijen (Radboud University)

### **Steering Committee**

Dr. Steven Coats (University Oulu)

Prof. Dr. Julien Longhi (CY Cergy Paris University)

Prof. Reinhild Vandekerckhove (University of Antwerp)

Dr. Lieke Verheijen (Radboud University)

### **Scientific Committee**

Paul Baker (Lancaster University)

Gavin Brookes (Lancaster University)

Noah Bubenhofer (University of Zürich)

Mario Cal Varela (Universidade de Santiago de Compostela)

Louis Cotgrove (Leibniz-Institut für Deutsche Sprache Mannheim)

Steven Coats (University of Oulu)

Orphée DeClercq (Ghent University)

Dominique Dias (Sorbonne Université)

Stephanie Evert (Friedrich-Alexander University Erlangen-Nürnberg)

Carolina Flinz (Università degli Studi di Milano)

Francisco Javier Fernández Polo (University of Santiago de Compostela)

Annamária Fábián (University of Bayreuth/Bavarian Research Institute for Digital Transformation – Bavarian Academy of Science)

Jenny Frey (European Academy of Bozen)  
Aivars Glaznieks (Eurac Research Bolzano)  
Alexandra Georgakopoulou-Nunes (King's College London)  
Vjosa Hamiti (University of Prishtina)  
Claire Hardaker (Lancaster University)  
Stefan Hartmann (Heinrich-Heine-University Düsseldorf)  
Iris Hendrickx (Radboud University Nijmegen)  
Axel Herold (Berlin-Brandenburgische Akademie der Wissenschaften)  
Besim Kabashi (Friedrich-Alexander University Erlangen-Nürnberg)  
Erik-Tjong Kim-Sang (Netherlands eScience Center)  
Alexander Koenig (CLARIN ERIC)  
Florian Kunneman (Utrecht University)  
Marc Kupietz (Leibniz-Institut für deutsche Sprache Mannheim)  
Mikko Laitinen (University of Eastern Finland)  
Els Lefever (Ghent University)  
Julien Longhi (Cergy-Pontoise Université)  
Harald Lüngens (Leibniz-Institut für deutsche Sprache Mannheim)  
Konstanze Marx-Wischnowski (University of Greifswald)  
Maja Miličević-Petrović (University of Bologna)  
Nelleke Oostdijk (Radboud University)  
Jan Oliver Rüdiger (Leibniz-Institut für deutsche Sprache Mannheim)  
Tatjana Scheffler (Ruhr-Universität Bochum)  
Mirco Schönfeld (University of Bayreuth)  
Steven Schoonjans (Alpen-Adria University Klagenfurt)  
Stefania Spina (Università per Stranieri di Perugia)  
Egon Stemle (Eurac Research)  
Caroline Tagg (The Open University)  
Igor Trost (Alpen-Adria University Klagenfurt/Universität Passau)  
Reinhild Vandekerckhove (University of Antwerp)  
Lieke Verheijen (Radboud University)  
Stefanie Walter (Technical University Munich)  
Katrin Weller (GESIS Cologne)

### **III. Keynotes**



# Studying Discourse in Social Media: Challenges & Opportunities

Stephanie Evert

Chair of Computational Corpus Linguistics, FAU Erlangen-Nürnberg  
Bismarckstr. 6, 91054 Erlangen, Germany  
E-mail: stephanie.evert@fau.de

## Abstract

Corpus-linguistic studies of social media discourses are essential for understanding how socio-political positions are negotiated in the semi-public sphere, how opinions can be manipulated by targeted campaigns, and how disinformation is spread in order to destabilise democracies world-wide. However, systematic large-scale studies in particular face a range of technical and methodological challenges, including data availability, automatic linguistic annotation of non-standard language, the limitations of corpus queries, and the lack of a true integration of quantitative and qualitative approaches. In this contribution, I discuss these challenges in detail and suggest some urgently needed innovations for social media corpus research (as well as corpus-assisted discourse studies in general).

**Keywords:** corpus-assisted discourse studies, social media, corpus linguistics, NLP, digital hermeneutics

## 1. Introduction

Corpus-assisted discourse studies or CADS (Fairclough, 2013; Baker et al., 2008; Mautner, 2009) offer an important window into how socio-political positions are negotiated in our society, especially in the interaction between decision-makers and the general public. In recent years, social media and other forms of computer-mediated communication have become a major platform for such socio-political discourses, shifting the focus from a small number of actors represented in mass media to a complex network of interactions in which nearly everyone can participate.

At the same time, there is a rapid increase in the prevalence of disinformation, toxicity, populism, and conspiracy theories – a phenomenon that is becoming a threat to democratic societies worldwide. Social media platforms offer various technological affordances for targeted manipulation of discourses (e.g. via disinformation campaigns), and platforms such as X and Truth Social even appear to be intended for this very purpose. The “filter bubbles” and “echo chambers” of social media networks create additional breeding grounds for anti-democratic positions. At the time of writing, this development has already reached a point where a successful disinformation campaign against a candidate for the German supreme court orchestrated by far-right media was echoed by prominent members of the clergy and the ruling conservative party CDU/CSU.<sup>1</sup>

It is thus of great importance and urgency to analyse and understand the discourses of populism and disinformation in social media, and to find ways of bringing them under control and fighting back. This endeavour requires an interdisciplinary collaboration between natural language processing (NLP), corpus linguistics, and the humanities and social sciences – and it is an excellent opportunity for exploring the synergies between these fields.

This paper addresses the challenges of studying social media data with a combination of NLP and corpus-linguistic techniques. It offers some suggestions for necessary methodological and technological innovations, which can make valuable contributions to both fields.

## 2. Challenges

### 2.1. Data Availability

For many years, Twitter was perhaps the best-studied social media network in fields such as natural language processing and computational social studies because researchers had free and easy access to large amounts of Twitter data (Mejova et al., 2015). Other popular social media platforms for research into socio-political discourses include Reddit (Blombach et al., 2020) and Telegram (Blombach et al., 2025b), but data have also been collected from other sources such as Facebook posts and user comments on YouTube videos.

Recently, social media data have become much less accessible, especially for academic research. Most prominently, access to Twitter data has been shut down completely after its acquisition by Elon Musk and the renaming to X. The company has even taken drastic measures to prevent data collection via Web client. Similar developments can be observed for other social media platforms, too: for example, Reddit has stopped offering data downloads via PushShift.<sup>2</sup> In order to continue legitimate and much-needed research e.g. into disinformation campaigns, a collaborative effort of researchers will be needed in order to bring together existing data sets (Pfeffer et al., 2023).<sup>3</sup> While not optimal, such historical data sets collected by various research groups are still useful for understanding the general mechanisms of anti-democratic discourses, especially with a combined data set that covers a broader range of topics and actors. For research into current topics, a massive worldwide collaboration of researchers might enable semi-automatic data collection via personal user accounts.

<sup>1</sup><https://www.tagesschau.de/kommentar/brosius-gersdorf-122.html>, <https://www.tagesschau.de/inland/innenpolitik/brosius-gersdorf-katholische-kirche-100.html>, as well as <https://www.lto.de/recht/nachrichten/n/lto-dokumentiert-erklaerung-im-wortlaut> and <https://www.tagesschau.de/inland/innenpolitik/kloeckner-gotthardt-nius-102.html>

<sup>2</sup><https://pushshift.io/signup>

<sup>3</sup>A single research group may well possess more than 10 TiB of Twitter data dumps, usually on specific topics that were of particular interest to the group.

## 2.2. Linguistic Annotation & Normalisation

NLP research and applications rely increasingly on end-to-end learning with large language models (LLMs) that do not require explicit linguistic annotation of input texts (such as dependency parsing, which used to serve as a basis for various information extraction tasks). LLMs have also proven quite robust to spelling variation, non-standard grammar, and other idiosyncrasies of social media data.

Corpus linguists, on the other hand, still work with traditional annotation levels such as part-of-speech (POS) tagging and lemmatisation. These annotation levels are crucial prerequisites for effective corpus queries and frequency analysis, forming a meaningful unit of analysis that connects automatic quantitative methods to hermeneutic interpretation. As an example, consider the important role of keywords and collocations in CADS studies.

Evaluation studies have shown that off-the-shelf automatic annotation tools perform very poorly on non-standard data from computer-mediated communication (CMC), and often even on data from Web pages (Giesbrecht and Evert, 2009; Beißwenger et al., 2016). Additionally, there is a lack of tools for automatic normalisation of non-standard spellings, which would simplify the formulation of corpus queries and help to aggregate frequency counts across spelling variants.

## 2.3. Corpus Queries

Corpus queries often form the starting point of a corpus-linguistic analysis, especially in concordancing tools such as CQPweb (Hardie, 2012) and Sketch Engine (Kilgarriff et al., 2014). In CADS, complex queries arise e.g. when studying rhetoric, persuasion, or argumentation patterns, where they provide a formalisation of linguistic hypotheses that can also be used for automatic data mining (Dykes et al., 2022; Dykes et al., 2024a).

Existing corpus query languages (CQLs) are designed to express flexible lexico-grammatical patterns in a precise formal notation. In most cases, they are either based on regular expressions (finite-state queries) or on tuples of anchor points connected by structural relations (Evert et al., 2025). Such CQLs are not very suitable for the challenges posed by social media data and the needs of discourse analysis:

1. Typographic errors, creative spellings, and non-standard grammar are difficult to capture with precise formal queries. Instead, some form of fuzzy matching would be needed, both at the level of individual tokens and at the level of token sequences. In many cases, the desired patterns can only be approximated, often by adding various heuristic filters to an initial query in order to reduce the number of false positives.
2. Patterns of interest to CADS researchers often involve semantic elements that are difficult to formalise through lexical or structural constraints. Examples are personal attacks in ad-hominem arguments (involving some kind of invective) or metaphorical expressions from a particular source domain. An ideal CQL would therefore need to support matching elements of a query by semantic similarity.

3. In many languages, relevant lexico-grammatical patterns combine both surface sequences and long-distance dependencies (e.g. a prepositional phrase with its governing verb, noun, or adjective; or verbs with separated particle in German). Current CQLs are geared towards either one (finite-state queries) or the other (anchored queries) and fail to offer an effective combination of both approaches.

## 2.4. Multimodal Discourses

Communication in social media is often multimodal, combining text with images or video snippets, or even replacing written text completely by video content e.g. on TikTok. Multimodal posts take many different forms: Sometimes one of the modalities is dominant (and a purely decorative image can be ignored in a linguistic analysis without too much loss). In other cases, the intended message is only created through the interaction of text and image (the most prototypical case being memes), or one modality modulates the interpretation of the other (e.g. if the text creates a misleading framing of an image or vice versa) (Primig et al., 2023; Martinez Pandiani et al., 2025).

While there has been considerable work on studying large collections of images in digital humanities and other fields, often under the label of “distant viewing” (Arnold and Tilton, 2023), no established methodology is available for integrating these approaches into a corpus-linguistic analysis. Nor are there suitable software tools for such research, with concordancing software focused strongly on textual content. Promising starting points for multimodal CADS are multimodal language models for automatic labelling of images beyond mere object recognition (Sharma et al., 2023), as well as work in NLP e.g. on fake news detection (Segura-Bedmar and Alonso-Bartolome, 2022).

## 2.5. Quantitative-Qualitative Integration

Effective analysis of large social media corpora (which can easily scale up to billions of words) requires the use of quantitative methods such as keywords and collocations, topic models, semantic clustering, as well as many other techniques. However, their results are just statistical summaries of observable linguistic patterns. A human interpretation and contextualisation is essential in order to gain a deeper understanding of discursive positions, argumentation strategies, the underlying goals of different actors, etc. in a CADS study.

So far, the combination of quantitative and qualitative aspects is almost always realised in the form of a unidirectional process, which starts with a quantitative analysis that operates without any human input (except for a few parameter settings). The human analyst then has to make sense of the quantitative results, often through visualisations (with the risk of an interpretation guided by aesthetic appraisal) and aided by more or less systematic close reading of individual examples (e.g. via a concordancing software). Crucially, the human insights do not feed back into the quantitative analysis. The hermeneutic circle is only closed in a very indirect manner by re-running analyses with different algorithms or parameter settings, or by applying them to a different data set. This severely limits the effectiveness



of quantitative algorithms in understanding complex social media discourses.

### 3. What is Needed

#### 3.1. Corpus Annotation Tools

Large-scale corpus studies of social media discourses depend on the development of off-the-shelf annotation tools for CMC content, especially for reliable POS tagging, lemmatisation, and dependency parsing. Training and development data sets are readily available in various languages. For German, the EmpiriST 2.0 gold standard provides manually annotated POS tags, normalisation, lemmatisation and semantic tagging (Proisl et al., 2020).

Fine-tuning of LLMs should achieve good results even with small amounts of training data, exploiting their robustness against non-standard language and the large amount of Web and CMC data in their pre-training corpora. In my experience, simple HMM-based clustering (Brown et al., 1992) can be very effective for detecting and normalising spelling variants in large social media corpora (Owoputi et al., 2013).

#### 3.2. New Corpus Query Languages

I believe that new CQLs (and, of course, corresponding implementations) need to be developed, with four essential innovations:

1. Integrate the two main query paradigms of current CQLs, namely finite-state queries for matching lexico-grammatical surface patterns and anchored queries for following dependency links and other structural relations (Evert et al., 2025, Ch. 4+5).
2. Conceptualise corpus queries as consecutive approximations, starting from a relatively general initial query and adding heuristic filtering constraints until a sufficient precision is obtained. This mirrors the process followed by many corpus linguists and explicitly supported through subqueries and set operations in the CQP query language (Evert and The CWB Development Team, 2020).
3. Enable fuzzy matching at the level of token sequences (e.g. by skipping extra tokens between query elements), linguistic annotation (e.g. by allowing certain substitutions of POS tags), orthographic similarity (to account for spelling variation), and semantic similarity (ideally based on sophisticated LLM embeddings).
4. Extract frequency data tables directly via queries (rather than just lists of query matches), which can be much more efficient on large corpora and simplifies the integration of queries with quantitative analysis (whereas in current practice, corpus queries are mostly a starting point for concordance reading separate from quantitative methods).

A sensible first step in the development of a new CQL is to document its functionalities, syntax, and semantics in the CQLF Ontology (Evert et al., 2020).<sup>4</sup> This enables a

direct comparison with other CQLs and invites comments and suggestions from potential users. For the query implementation, the Ziggurat data model (which builds on and extends the well-established tabular data format) provides an excellent foundation (Evert and Hardie, 2015; Evert et al., 2023).

#### 3.3. Automatic Classification

CADS research would often benefit from automatic text classification according to ad-hoc categories relevant to a particular study. These might include metaphors, typical linguistic features of disinformation, fallacious argumentation patterns, hedging and indirection, etc.

Approaches based on pre-trained LLMs can often achieve satisfactory results with very small amounts of training data. For example, a zero-shot learning approach has been used successfully for the identification of conspiracy narratives (Heinrich et al., 2024a; Blombach et al., 2025a). In my research group, we are currently experimenting with few-shot training for the automatic annotation of linguistic and rhetorical characteristics of disinformation (Blombach et al., 2025b). An alternative strategy is the high-precision identification of argumentation patterns in social media with corpus queries, which can then be used as training data for automatic classifiers with a more balanced recall-precision trade-off (Dykes et al., 2024b).

#### 3.4. A Framework for Digital Hermeneutics

A genuine integration of quantitative and qualitative approaches must ensure a bidirectional workflow, in which human interpretation feeds back directly into the quantitative analysis. There is an urgent need for research on the necessary theoretical, methodological, and algorithmic foundations, which I refer to as “digital hermeneutics”.

A first step towards digital hermeneutics for corpus-assisted discourse studies is the recent MMDA approach (Heinrich et al., 2024b; Heinrich and Evert, 2024). It operationalises one part of the typical CADS interpretation process – the manual grouping of collocates or keywords – as the formation of “discourseemes”, defined as minimal units of lexical meaning in the context of a specific discourse. Constellations of such discourseemes then indicate framings and discursive positions (consider e.g. a combination of the discourseemes MIGRANT, FLOOD, and MENACE). Since discourseemes are represented by sets of lexical items, they can easily be identified in a corpus and used by quantitative algorithms, e.g. to show temporal trends, to track the spread of discursive positions across social media networks, or to highlight discourseemes in concordance displays.

A second approach focuses on the algorithms that corpus linguists use to organise concordance lines for interpretation. Off-the-shelf concordancing tools are often limited to a relatively small set of traditional approaches such as sorting alphabetically by left or right context, random shuffling or thinning, and filtering by manually specified keywords or typical collocates. The RC21 project<sup>5</sup> aims to integrate algorithms more flexibly and more tightly into the concordance reading process. Based on a mathematical taxonomy

<sup>4</sup><https://github.com/cqlf-ontology/>

<sup>5</sup><https://www.dhss.phil.fau.eu/research/reading-concordances/>

rooted in five general strategies of organising concordances (Selecting, Sorting, Ranking, Partitioning, and Clustering), a wide range of algorithms can be implemented in a common framework and their application is documented in the form of an analysis tree, ensuring reproducibility of the concordance analysis.<sup>6</sup>

### 3.5. An Integrated CADS Platform

The ultimate goal, though, is the creation of an integrated online platform for CADS research that enables researchers to develop collaborative analyses across multiple topics, corpora, and languages. This platform should combine the innovations I have suggested above with established concordancing and CADS tools. A useful starting point could be the Swiss-AL platform (Krasselt et al., 2021).<sup>7</sup> As a first step, the MORCDA project aims for an experimental integration of MMDA and innovative concordance reading algorithms into Swiss-AL.<sup>8</sup>

## 4. References

- Arnold, T. and Tilton, L. (2023). *Distant Viewing: Computational Exploration of Digital Images*. The MIT Press.
- Baker, P.; Gabrielatos, C.; Khosravini, M.; Krzyżanowski, M.; McEnery, T. and Wodak, R. (2008). A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), pp. 273–306.
- Beißwenger, M.; Bartsch, S.; Evert, S. and Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pp. 44–56, Berlin, Germany.
- Blombach, A.; Dykes, N.; Heinrich, P.; Kabashi, B. and Proisl, T. (2020). A corpus of German reddit exchanges (GeRedE). In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 6310–6316, Marseille, France.
- Blombach, A.; Doan Dang, B. M.; Evert, S.; Fuchs, T.; Heinrich, P.; Kalashnikova, O. and Unjum, N. (2025a). Narrlängen at SemEval-2025 task 10: Comparing (mostly) simple multilingual approaches to narrative classification. In Sara Rosenthal, et al., editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 2240–2248, Vienna, Austria.
- Blombach, A.; Evert, S.; Havenstein, L. and Heinrich, P. (2025b). Dimensions of drivel in German Telegram posts: Manual annotation and predictive power. In *Proceedings of the 12th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2025)*, Bayreuth, Germany.
- Brown, P. F.; Della Pietra, V. J.; de Souza, P. V.; Lai, J. C. and Mercer, R. L. (1992). Class-based  $n$ -gram models of natural language. *Computational Linguistics*, 18(4), pp. 467–479.
- Dykes, N.; Heinrich, P. and Evert, S. (2022). Retrieving Twitter argumentation with corpus queries and discourse analysis. In Susanne Flach et al., editors, *Broadening the Spectrum of Corpus Linguistics. New approaches to variability and change*, number 105 in Studies in Corpus Linguistics. John Benjamins.
- Dykes, N.; Evert, S.; Heinrich, P.; Humml, M. and Schröder, L. (2024a). Finding argument fragments on social media with corpus queries and LLMs. In Philipp Cimiano, et al., editors, *Robust Argumentation Machines*, pp. 163–181, Cham. Springer Nature Switzerland.
- Dykes, N.; Evert, S.; Heinrich, P.; Humml, M. and Schröder, L. (2024b). Leveraging high-precision corpus queries for text classification via large language models. In Annette Hautli-Janisz, et al., editors, *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pp. 52–57, Torino, Italia.
- Evert, S. and Hardie, A. (2015). Ziggurat: A new data model and indexing format for large annotated text corpora. In *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora (CMLC-3)*, pp. 21–27, Lancaster, UK.
- Evert, S. and The CWB Development Team, (2020). *The IMS Open Corpus Workbench (CWB) CQP Interface and Query Language Manual*. CWB Version 3.5, available at <http://cwb.sourceforge.net/documentation.php>.
- Evert, S.; Harlamov, O.; Heinrich, P. and Bański, P. (2020). Corpus Query Lingua Franca part II: Ontology. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 3346–3352, Marseille, France. Also see <https://github.com/cqlf-ontology/>.
- Evert, S.; Hardie, A. and Weber, T. (2023). The Ziggurat data model and file format (draft 1.5). Available from <https://osf.io/n75es/>.
- Evert, S.; Weber, T.; Bothe, S.; Heinrich, P. and Piperski, A. (2025). Data exploitation: Corpus queries. In Piotr Bański, et al., editors, *Standards for Language Data and Infrastructures*, Digital Linguistics. De Gruyter. To appear.
- Fairclough, N. (2013). *Critical Discourse Analysis: The Critical Study of Language*. Routledge, London.
- Giesbrecht, E. and Evert, S. (2009). Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus. In Iñaki Alegria, et al., editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, pp. 27–35, San Sebastian, Spain.
- Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), pp. 380–409.
- Heinrich, P. and Evert, S. (2024). Operationalising the hermeneutic grouping process in corpus-assisted discourse studies. In Christopher Klamm, et al., editors, *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and*

<sup>6</sup><https://pypi.org/project/FlexiConc/>

<sup>7</sup><https://swiss-al.linguistik.zhaw.ch/>

<sup>8</sup><https://oscars-project.eu/projects/morcda-making-open-research-data-suitable-comparative-discourse-analysis>

- short papers, pp. 33–44, Vienna, Austria.
- Heinrich, P.; Blombach, A.; Doan Dang, B. M.; Zilio, L.; Havenstein, L.; Dykes, N.; Evert, S. and Schäfer, F. (2024a). Automatic identification of COVID-19-related conspiracy narratives in German Telegram channels and chats. In Nicoletta Calzolari, et al., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1932–1943, Torino, Italia.
- Heinrich, P.; Blombach, A.; Dykes, N.; Evert, S.; Fuchs, T.; Havenstein, L. and Schäfer, F. (2024b). From linguistic to discursive patterns: Introducing discourseemes as a basic unit of discourse analysis. *CADAAD Journal*, 16(2), pp. 87–111.
- Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovář, V.; Michelfeit, J.; Rychlý, P. and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.
- Krasselt, J.; Fluor, M.; Rothenhäusler, K. and Dreesen, P. (2021). A workbench for corpus linguistic discourse analysis. In Dagmar Gromann, et al., editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASISs)*, pp. 26:1–26:9, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
- Martinez Pandiani, D. S.; Tjong Kim Sang, E. and Ceolin, D. (2025). ‘Toxic’ memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, 47, pp. 100317.
- Mautner, G. (2009). Corpora and critical discourse analysis. In Paul Baker, editor, *Contemporary Approaches in Corpus Linguistics*, pp. 32–46. Continuum Books, London.
- Yelena Mejova, et al., editors. (2015). *Twitter: A Digital Socioscope*. Cambridge University Press, New York.
- Owoputi, O.; O’Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N. and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pp. 380–390, Atlanta, GA.
- Pfeffer, J.; Matter, D.; Jaidka, K.; Varol, O.; Mashhadi, A.; Lasser, J.; Assenmacher, D.; Wu, S.; Yang, D.; Brantner, C.; Romero, D. M.; Otterbacher, J.; Schwemmer, C.; Joseph, K.; Garcia, D. and Morstatter, F. (2023). Just another day on Twitter: A complete 24 hours of Twitter data. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1), pp. 1073–1081.
- Primig, F.; Szabó, H. D. and Lacasa, P. (2023). Remixing war: An analysis of the reimagination of the Russian–Ukraine war on TikTok. *Frontiers in Political Science*, 5. <https://doi.org/10.3389/fpos.2023.1085149>.
- Proisl, T.; Dykes, N.; Heinrich, P.; Kabashi, B.; Blombach, A. and Evert, S. (2020). EmpiriST corpus 2.0: Adding manual normalization, lemmatization and semantic tagging to a German Web and CMC corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 6142–6148, Marseille, France.
- Segura-Bedmar, I. and Alonso-Bartolome, S. (2022). Multimodal fake news detection. *Information*, 13(6), pp. 284.
- Sharma, S.; Agarwal, S.; Suresh, T.; Nakov, P.; Akhtar, M. S. and Chakraborty, T. (2023). What do you MEME? generating explanations for visual semantic role labelling in memes. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’23/IAAI’23/EAAI’23*, pp. 9763–9771. AAAI Press.

# Studying language and identity in a corpus of computer-mediated communication with (and without) sociodemographic metadata

Gavin Brookes

Lancaster University

E-mail: g.brookes@lancaster.ac.uk

## Abstract

This paper explores the methodological and interpretive implications of analysing language and identity in large corpora of computer-mediated communication (CMC), both with and without sociodemographic metadata. Drawing on a 14-million-word corpus of online patient feedback about UK cancer care, I compare two approaches: one using metadata (e.g. patient-declared sex) and another relying on patients' in-text self-references. The metadata approach enables large-scale, statistically grounded comparisons, revealing broad patterns, such as male patients' focus on procedures and female patients' emphasis on emotional and interpersonal dimensions of care. The self-reference approach, while limited by smaller sample sizes, offers nuanced insights into how patients perceive and mobilise intersecting aspects of identity, including sex and age. The paper highlights the trade-offs between scale and contextual richness, advocating for a combined, bottom-up and top-down approach. It concludes that identity analysis in CMC benefits from attending to both declared demographic categories and emergent, textually embedded identity cues.

**Keywords:** corpus linguistics, identity, sociolinguistics

## 1. Introduction

This talk will reflect on the challenge of answering questions relating to language and identity in corpora of computer-mediated communication (CMC) when we, as analysts, do not have access to reliable sociodemographic metadata. The talk reflects on an experiment, reported in Baker and Brookes (2022), which compared the affordances of two approaches to studying identity in CMC: (i.) using sociodemographic metadata; and (ii.) using language users' in-text attestations of their identities. To do this, we performed two sets of analyses, each one adopting either of the approaches noted above, in particular comparing the language used by male and female patients in a corpus of online patient feedback about cancer care services in the UK (14,403,694 tokens).

## 2. Data and approach(es)

Our methodology, then, comprised two approaches. For the first approach, we used the sociodemographic metadata available to us. Focussing on sex identity, we tagged the corpus and divided it into two sub-corpora, stored and analysed on *CQPweb* (Hardie 2012). One of comments in which patients checked a box to indicate that they identify as male, and another of comments in which patients checked a box to indicate that they identify as female (note that a small number of patients contributing to this corpus identified as 'Other', including non-binary. However, there was not enough data of this kind to facilitate the kind of analysis being undertaken in this study). For the purposes of this experiment, we refer to this approach as the 'metadata approach', as it relied on the sociodemographic metadata that our healthcare provider partners made available to us. Within our corpus, there were 97,774 comments from male patients (5,720,898 tokens) and 116,564 comments from female patients (8,683,079 tokens).

For the second approach, we operated under the artificial assumption that we did *not* have access to any sociodemographic metadata. For this analysis, we adopted an approach resembling one we were forced to adopt in previous work with similar data (Baker et al. 2019), and

searched for cases where patients referenced their sex identity within the comments themselves. To exemplify, one patient prefaced their feedback with the phrase, 'As a 52 year-old man...'. On this basis, we determined the patient contributing this comment to identify as male. Again, we grouped the comments into two sub-corpora: one in which patients referred to themselves as male in their comments, and another in which patients referred to themselves as female. And as this approach relied on patients *referring* to their sex identity, we can refer to this approach as the 'self-reference approach'.

## 3. Findings

### 3.1. The metadata approach

We then compared the two sets of comments against each other using the keywords technique (statistic: log-likelihood with log ratio). This gave two sets of keywords – one for the male patients' comments compared against the female patients' comments, and one for the female patients' comments compared against the male patients' comments. We focused on the top 30 keywords from each set, ranked by log-likelihood score. This was an arbitrary cut-off but it did give a manageable number of keywords for analysis. These keywords are shown in Table 1 (for full table with statistical information, see Baker and Brookes 2022: 18-19).

<i>class, bladder, treatment, good, hospital, nhs, first, no, by, condition, test, carried, blood, thanks, kidney, gp, bowel, endoscopy, ), yes, quality, problem, attention, period, general, months, removal, myeloma, professionalism, successful</i>
--

Table 1: Keywords for male patients' comments versus female patients' comments.

We then analysed these keywords qualitatively with the broad aim of interpreting their uses in terms of recurrent rhetorical patterns and gendered discourses (Sunderland 2004). To do this, we went beyond concordance lines and examined the comments in their entirety.

Male patients were more likely to refer to their cancer and other diseases in their comments, evidenced through the keyness of words such as *bladder, bowel, kidney* and

*myeloma* in this data. The keywords also featured more general disease-related terms, such as *problem* and *condition*. Male patients also tended to focus on treatment processes, which evidenced in uses of keywords such as *removal*, *tests* and *endoscopy*. This tendency also accounts for the keyness of the constituents of the phrasal verb, *carried out*, as well as the word *by*, which tended to be used in passive constructions of medical processes.

Healthcare staff were also indexed by male patients through uses of keywords such as *NHS*, *General* and *Hospital*. These words could function metonymically, being used to denote all staff involved in a patient's care. Through such constructions, male patients could present their feedback as applying not just to a single staff member or team, but to an entire site of care or even the healthcare system as a whole. This could therefore represent a rhetorical strategy used by male patients in particular to generalise and present their complaints as being particularly pressing.

A characteristic theme of the male patients' comments is time, indicated in the keywords *months* and *period*. These tended to be used to quantify the amount of time that male patients had to *wait* for something, typically a diagnosis or an appointment for treatment. While the theme of waiting was frequent in both the male and female patients' comments, the male patients' comments provided more precise quantification of their waits.

The final group of keywords from male patients' comments are the words *no*, *yes* and *thanks*. And these keywords reflected the almost dialogic manner in which these patients in particular interacted with the voice of the feedback form, as in their comments they answered the prompt questions framing the feedback literally – with a *no* or a *yes* – and to express thanks for the quality of the service they received. This feature seems to be an effect of age as well as sex identity. Inspecting the frequencies of these keywords across the age groups, as well as between the sexes, we found that these words were all much more likely to be used by older patients, and by older *male* patients at every age group. Because these words are more common in men at all ages, this feature is likely an effect of the mixture of age and sex as factors.

*i, kind, felt, n't, amazing, feel, husband, she, so, lovely, oncologist, chemotherapy, me, they, had, radiotherapy, her, wonderful, did, you, nurse, unit, when, wait, supportive, lump, chemo, everyone, caring, busy*

Table 2: Keywords for female patients' comments versus male patients' comments.

Moving onto the female patients' keywords (Table 2; see also Baker and Brookes 2022: 27-28), and while the male patients' comments focused on procedural and transactional aspects of service, female patients tended to adopt a more personalised style, as reflected in the keyness of the pronouns *I* and *me*. This more gave rise to a more characteristic focus on how female patients' experiences made them *feel*. Staff were also evaluated using keywords such as *kind*, *lovely*, *supportive* and *caring*. They were also evaluated as *amazing* and *wonderful* and using the intensifier *so*. When we analysed 100 uses of each of these latter keywords, we again found that they tended to denote staff interpersonal skills.

Also key for female patients' comments were words indicating a stronger focus on individuals (e.g., *she*, *oncologist*, *her* and *nurse*), as well as words denoting relatives, units and smaller teams of staff. The keywords *chemotherapy*, *radiotherapy* and *chemo*, while ostensibly denoting types of treatment, tended instead to refer to teams of staff. Meanwhile, the keyword *everyone* could refer to staff working in teams or on wards, but at other points referred to other patients. In these cases, the female patients rendered their experiences as more generalisable, and this was also something we saw in uses of the general *you*.

A shared concern for male and female patients is the theme of waiting. When female patients described and evaluated waits, they did so in much less precise terms than male patients did. These patients specified the duration of waits in just 15 per cent of cases, which might be why words such as *months* and *period* are key for male patients' comments compared to female patients' ones.

### 3.2. The self-reference approach

The first step of this approach was to search for uses of the term 'man' and then the term 'woman'. We then extracted 100 comments in which patients self-identified as male and a hundred comments in which patients identified as female. We manually checked both samples to ensure that patients were indeed referencing their own sex identities, and not someone else's. For this analysis, we were forced to adopt a slightly different approach to obtaining keywords. We began by trying to compare the samples directly against each other, as we did in the metadata approach. However, this yielded a very small number of keywords, and these did not really tell us anything about gender-based patterns. This is likely a result of the small sample sizes that this approach forced us to work with (the maximum number of comments we could have analysed to have balance across male and female patients was 102). As a work-around, we generated keywords by comparing each of our samples against the rest of the comments in our corpus as a whole. And we might regard this reference corpus as a general corpus of cancer patient feedback. So these comparisons gave us two sets of keywords: one for the sample of male patients and one for our sample of female patients (show in in Tables 3 and 4, respectively; see also: Baker and Brookes 2022: 33-34).

*man, old, a, said, i, that, lucky, prostate, am, life, we, young, now, ", sick*

Table 3: Keywords for the sample of male patients' comments compared to the rest of the corpus.

*age, old, women, younger, hair, wig, intelligent, children, should, fertility, me, said, !, this, ovarian, be, that*

Table 4: Keywords for the sample of female patients' comments compared to the rest of the corpus.

Because we compared the samples against the same reference corpus, rather than against each other, we had some overlapping keywords, which could be viewed as indicating what is lexically characteristic of feedback in which patients declare their sex identities compared to

feedback more generally. A drawback of this approach is that the differences between the keywords here are not statistically significant between our two samples. However, an advantage of the approach is that it does at least let us look at *similarities* between the two samples, by looking at the overlapping keywords. We then undertook a close analysis of these keywords, proceeding in the same way as we did for the metadata approach.

A striking similarity between both samples is the keyness of the quotative *said*. The fact that this is key suggests that patients in both samples quote others in their comments more often than we might expect in feedback on cancer care in general. This also helps to explain the keyness of the word *that*, which tended to be used to frame quotations. The use of quotations seems to emerge as especially frequent in these samples because the patients' sex identity is often mentioned in the quoted speech. The use of quotes is linked to negative feedback in particular, as patients tended to use quotes when recounting cases in which they were given advice that they viewed as inconsistent or inaccurate, or cases in which they experienced staff rudeness.

Another overlapping feature across both the male and female samples was the use of keywords relating to age. For the male patients, this includes the words *old* and *young*, and for the women's comments we get *age*, *old* and *younger*. Both male and female patients frequently referenced their age in conjunction with their sex for evaluative purposes. For example, both male and female patients referenced their age in order to construct themselves as having particular healthcare requirements. Sometimes these requirements were met and sometimes they were not, and this could determine whether the feedback given was broadly positive or negative. In some cases, the negative evaluation targeted gendered stereotypes that patients attributed to healthcare staff. For example, one male patient complained about being treated like a 'grumpy old man', while a female patient complained about being treated like a 'silly old woman'. Both male and female patients drew on the intersection of age and sex, then, to frame descriptions of experiences in which they felt belittled by staff members.

As well as older age, both the male and female patients in our samples also referenced youth. Some of the male patients used the keyword *young* to construct themselves as socially and sexually active, with these aspects of their identity being linked to both their age and sex. And so this was again about constructing particular healthcare needs, and whether or not these were met could again motivate positive or negative feedback. Where the adjective *young* was key in the male comment sample, the comparative form *younger* was key for the female sample. Female patients tended to use the keyword *younger* to refer either to younger female patients in general, or to hypothetical others. Such comments typically described how particular aspects of service provision would not be suitable for younger female patients, and often made recommendations about how services could be improved for younger women in the future. This pattern, of the female patients issuing recommendations, also helps to account for the keyness of *should* in the female patient sample.

Male patients, on the other hand, frequently produced a

discourse of exceptionalism. This was realised, for example, in the keyword *lucky*, which male patients tended to use to describe themselves as being lucky for having been treated by a highly skilled practitioner or team. In other contexts, *lucky* is used by male patients when relaying interactions with staff in which they'd been told that they're *lucky* to be alive. In either case, male commenters imply that their experiences are somehow exceptional or even unique, either in terms of the high standards of care they received, or the severity of their illness. Cases of the latter also help to account for the keyness of *sick* in the male patient sample, as some of the men described how staff informed them that they were 'very *sick* men'.

Another keyword which indicates the male patients' focus on their own experiences is the temporal adverb *now*. While female patients frequently made recommendations as to how services could be improved for others in future, male patients tended to focus instead on the past, in addition to the present. These descriptions of the past took on an almost autobiographical tone, as the male patients often recounted their previous experiences with a provider, and described the different forms of treatment that had brought them to the present – i.e. to the *now*. Thus, male patients used *now* in order to draw comparisons between their current experiences and previous ones. A similar tendency is observable for uses of the keyword *life*, with male patients either thanking staff for 'saving' or 'improving' their *life*, or evaluating an experience as being the 'worst of [their] *life*'.

## 4. Conclusions

The metadata had the advantage of allowing us to base our findings on a much larger dataset. This not only meant that we could have greater confidence in the trends we identified, but it also allowed us to perform direct statistical comparisons of our sex-based subsets using the keywords technique. Another advantage of this approach was that we were able to draw on other metadata tags to interpret some of the patterns we found. For instance, our interpretation of the finding that male patients engaged with the feedback form in a more dialogic way, was enriched by our ability to look at age-related metadata too, where we could see that this was a feature of older male patients in particular. This supplementary perspective was only possible because we could draw on this extra sociodemographic information. Without it, we would not have been able to arrive at that interpretation.

Yet the metadata approach also had some shortcomings. While having a vast corpus tagged for sociodemographic information brings lots of clear advantages, *assembling* such a corpus – and with all of this metadata – remains a demanding (and resource-intensive) task. We were helped in this project by our collaboration with NHS England, as our contacts there collected the metadata from patients, in an ethically appropriate manner, at the point at which the feedback was given. They then provided that metadata to us in a format whereby it was relatively straightforward for us to convert it into a series of searchable tags. Without their support, this would have been a much more resource-intensive exercise.

A criticism of sociodemographic annotations is that they often depend on quite broad social categories. In this work, we were forced to work with the categories of ‘male’ and ‘female’. But these broad categories could result in us taking a top-down and overly simplistic view of identity. While these categories might be suitably broad to be operationalizable in a large-scale corpus analysis, they also risk obscuring more nuanced types of identity relations. In other words, what is gained from broad categories in terms of scalability and practicality, might be lost in terms of granularity and contextual nuance.

Relatedly, we should also reflect here on a more general criticism that is often made of studies which correlate social categories with language use. Statistically significant correlations between a social attribute and the use of a linguistic feature are often interpreted as relationships of causation. In other words, if we find that use of a particular word or feature correlates with language users being male, we might be tempted to conclude that this trend occurs *because* those language users are men. However, the marked use of a linguistic feature can be related not just to the particular variable under focus, but to some other aspect of identity, or even a combination of these. It is also important to bear in mind that no set of sociodemographic annotations is ever complete. In our case, there were numerous other aspects of their identities that patients *could have* been asked about but were not. With the kind of data we were working with, then, which was elicited through a survey that we, ourselves, did not design, we were restricted by what the survey creators decided to ask about, either because they thought it was important, or because it was easy to measure and categorise.

Turning to the self-reference approach, an advantage of this was that we could have greater confidence that patients’ sex identities were more directly relevant to their comments. We knew this because the patients explicitly oriented to these aspects of their identities in the comments themselves. In this way, this approach gave us something of an interpretive warrant, which meant that we could be more confident that the differences we were observing were indeed related to patients’ self-attested sex identities. Another advantage of this approach was that it gave us an arguably more organic route into looking at intersectionality, as patients’ orientations to one identity category frequently accompanied, or gave rise to, another. For example, we found that patients who referenced aspects of their sex identity were also particularly likely to reference their age too. These intersectional aspects of their identities were highlighted because patients perceived them as relevant to their experiences, and so to their feedback too.

This approach also had some limitations, too. The first concerns the size of the samples that the approach allowed us to work with. Because patients referred to their identities relatively *infrequently* in their comments, we were forced to work with very small samples. This posed several methodological challenges, and for example meant that we did not have sufficient data to perform a direct keyword comparison between the samples. Our small sample size also reduced the generalisability of our findings. Relatedly, comments in which patients went on-record about their identity in their feedback may not be considered to be representative of all the comments in our corpus as a whole.

That is, when patients went on-record about their identities in their comments, they often did so because they perceived these qualities to be central somehow to the type of feedback they were giving, and this was not the norm. As such, this approach might train our focus on certain types of comments which are not necessarily representative of the wider corpus (nor, indeed, the wider context of language that the corpus is intended to represent). If adopting this kind of approach, then, some caution is likely to be needed regarding making generalisations.

Finally, just like sociodemographic metadata can never capture *all* identity variables, we should also be mindful with our self-reference approach that, just because a language user doesn’t mention an aspect of their identity explicitly, that doesn’t mean that that aspect of their identity is not in fact relevant to their language use in a given context.

A pertinent question we might ask at this point, regards which approach might be best for researchers studying identity in a large corpus of texts. And of course, the answer is likely to depend on the levels of granularity and accuracy required, as well as, on a more practical note, the type of data we are working with (and what its limitations are).

Even if we do have access to reliable sociodemographic metadata, any approach to studying language use and identity (in computer-mediated communication or any other context) will nevertheless stand to benefit from our bringing in qualitative, bottom-up methods of analysis. In this vein, we could combine both of the approaches presented here. Such an analysis could start by looking closely at a sample of texts in the corpus and noting emergent identity categories. We could then use that analysis as means of narrowing our focus to those emergent categories when presented with (a potentially overwhelming range of) sociodemographic tags. This kind of bottom-up approach has the advantage of directing our analytical focus to those aspects of identity that the language users in our corpus *themselves* perceive to be contextually relevant. The reference-based approach that would precede any annotation-based analysis could also help us to account for more subtle or even implied forms of identity self-referencing. And this, in turn, could not only give our analysis focus, but also help us to guarding against an uncritical overreliance on correlational statistics.

## 5. References

- Baker, P., Brookes, G. and Evans, C. (2019). *The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication*. London: Routledge.
- Baker, P. and Brookes, G. (2022). *Analysing Language, Sex and Age in a Corpus of Patient Feedback: A Comparison of Approaches*. Cambridge: Cambridge University Press.
- Hardie, A. (2012). CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.
- Sunderland, J. (2004). *Gendered Discourses*. Basingstoke: Palgrave Macmillan.





## **IV. Talks**



# Towards a new Curation Workflow for the CMC Corpora Resource Family

Egon W. Stemle<sup>1</sup>, Lionel Nicolas<sup>1</sup>, Alexander König<sup>2</sup>

<sup>1</sup>Eurac Research, <sup>2</sup>CLARIN ERIC

E-mail: egon.stemle@eurac.edu, lionel.nicolas@eurac.edu, alex@clarin.eu

## Abstract

This paper aims at raising awareness regarding a recently started and ongoing effort of CLARIN ERIC and the CLARIN Knowledge Centre for CMC and Social Media Corpora (CKCMC) to enhance the visibility and accessibility of the CMC community's datasets through the CLARIN CMC Corpora Resource Family (CMC-RF), which, as of May 2025, the CKCMC officially *adopted*, that is, it took over responsibility.

We offer some possible scenarios regarding how curation (addition, change and deletion of entries) of the CMC-RF could be approached, with the objective to prepare a productive context for a roundtable at the forthcoming 2025 edition of the CMC-Corpora conference. We intend to use the outcomes of this roundtable to devise a grounded and informed first version of Curation Guidelines for the CMC-RF, together with a community-oriented procedure to update it.

**Keywords:** CLARIN, Knowledge Centre, CKCMC, Resource Family, Community, Curation Criteria

## 1. Introduction

CLARIN (Common Language Resources and Technology Infrastructure) is a pan-European research infrastructure for language data, organised as a European Research Infrastructure Consortium (ERIC) composed of national and regional nodes, each responsible for their respective countries or regions. The goal of CLARIN is to provide access to language data and tools, help in (1) finding suitable data to work with, (2) finding the right tools to analyse said data and (3) finding experts and knowledge on best research practices. (Hinrichs and Krauwer, 2014)

Concerning the aforementioned third aspect, finding experts and knowledge on best practices in research, CLARIN has set up a dedicated framework known as the *CLARIN Knowledge Infrastructure* to facilitate the creation, sharing, and reuse of knowledge and expertise among researchers, developers and other stakeholders in the language resource community.

At the heart of the Knowledge Infrastructure lie the CLARIN Knowledge Centres (K-centres), which serve as hubs of expertise on specific languages, topics, or domains. These centres are first certified by CLARIN, ensuring that the people involved do indeed possess the necessary expertise and are willing to provide their services to the CLARIN community as a whole, for example by setting up a helpdesk for questions. As of May 2025, CLARIN accounted for 35 K-centres spread across Europe and beyond, with a heterogeneous set of central foci including South Slavic languages (CLASSLA)<sup>1</sup>, Learner Corpora (CKL2CORPORA)<sup>2</sup>, atypical communication (ACE)<sup>3</sup> and CMC and Social Media Corpora (CKCMC)<sup>4</sup>.

The CKCMC is a distributed centre with partners from the Leibniz Institute for the German Language (IDS) in Germany, the Jožef Stefan Institute (IJS) in Slovenia and the

Laboratoire de Linguistique Formelle (LLF) in France. The Institute for Applied Linguistics (IAL) of Eurac Research in Italy coordinates the centre.

Another cornerstone of CLARIN's overall infrastructure are the CLARIN Resource Families (CRF)<sup>5</sup>, which are primarily aimed at helping researchers to find suitable data to work with. These curated lists feature high-quality corpora, tools and lexical resources for various domains identified as particularly relevant to the CLARIN community. The CRF initiative was launched by Darja Fišer and Jakob Lenardič, introduced to the wider public in 2018 (Fišer et al., 2018), and is now coordinated and curated by a small team of CLARIN ERIC employees led by Jakob Lenardič. The CRF currently comprise 27 families, many with numerous entries, and the list continues to grow. However, as the number of resources has increased, it has become clear that a small team can no longer effectively manage the curation process. To address this challenge, the Resource Families coordination team has started to invite thematically linked K-centres to *adopt* a family that aligns with their area of expertise. In particular, CLARIN ERIC invited the CKCMC to take over the curation of the CMC Corpora Resource Family (CMC-RF)<sup>6</sup>, which is a natural fit given the centre's objectives: adopting the CMC-RF was part of the CKCMC's agenda right from its inception, and the invitation was welcomed and accepted by the CKCMC in May 2025.

This paper discusses our ongoing effort to devise a grounded and informed first version of the Curation Guidelines for the CMC-RF and a community-oriented procedure to update it. This effort will be officially kickstarted with a roundtable at the CMC-Corpora conference.

The main objectives of this paper are as follows: (1) introduce our initiative and discuss the context, reasons and objectives motivating it; as well as how it situates itself within a larger set of community-focused past and ongoing efforts undertaken within the context of the CKCMC

<sup>1</sup><https://www.clarin.si/info/k-centre/>

<sup>2</sup><https://www.uclouvain.be/en/research-institutes/ilc/clarin-knowledge-centre-for-learner-corpora>

<sup>3</sup><https://ace.ruhosting.nl/>

<sup>4</sup><https://cmc-corpora.org/ckcmc/>

<sup>5</sup><https://www.clarin.eu/resource-families>

<sup>6</sup><https://www.clarin.eu/resource-families/cmc-corpora>

(Section 2.); (2) discuss the relevant aspects – as we currently see them – to be taken into account for devising the Curation Guidelines of the CMC-RF (Section 3.), together with a community-oriented procedure to update it.

## 2. CMC Corpora Resource Family (CMC-RF) Handover

The CKCMC collaborated on various initiatives to enhance corpus accessibility, interoperability and reusability. By promoting FAIR principles, standardised formats like TEI schemas and metadata standards, the CKCMC has demonstrated its commitment to advancing the field while ensuring alignment with broader linguistic resource initiatives.

### 2.1. CMC Community and the CKCMC

The CMC community comprises researchers, developers and practitioners interested in computer-mediated communication (CMC) and social media corpora. It actively organises conferences, workshops and other events to share research and ideas, and it facilitates collaboration and knowledge-sharing among its members.

The CKCMC is closely connected to the CMC community, and its members have been involved in the conference series for a long time. A testimony of this close connection is the fact that the official webpage of the CKCMC has been hosted on the official websites of the CMC community, as well as the fact that all partners of the CKCMC have hosted the CMC-Corpora conference in the past.

### 2.2. Community-focussing CKCMC Efforts

The CKCMC offers expertise on language resources and technologies for Computer-Mediated Communication and Social Media. Its basic activities are to (1) give researchers, students and other interested parties information about the available resources, technologies and community activities, (2) support interested parties in producing, modifying or publishing relevant resources and technologies, and (3) organise training activities. (Stemle et al., 2022)

Over the last few years, the CKCMC has promoted different aspects with contributions to the CMC-Corpora Conference Series<sup>7</sup>:

**Promoting FAIR Principles for CMC corpora:** Frey et al. (2019) evaluate the compliance of CMC corpora with FAIR principles, emphasising that while findability and accessibility are partially achieved through repositories like CLARIN B Centres, interoperability and reusability are lacking due to reliance on implicit community standards. The paper highlights the need for standardised practices to fully align with FAIR principles.

**Promoting a standardised TEI format for CMC corpora:** Beißwenger and Lungen (2020) introduce the "CMC-core" schema, a specialised TEI format designed for representing CMC corpora. The paper demonstrates how this schema facilitates interoperability and integration into established language resource infrastructures like CLARIN and ORTOLANG, promoting standardisation in corpus representation.

**Promoting Metadata schema creation for CMC corpora:** Stemle et al. (2024) focus on metadata collection for social media corpora, emphasising the importance of documenting platform-specific details to enhance reusability. The paper advocates for collaborative efforts among stakeholders to develop metadata standards and best practices that align with FAIR principles.

All these efforts tackle the same objective of facilitating access and reuse of corpora but from different perspectives.

### 2.3. Origin and Objectives of the CLARIN Resource Families – and the CMC-RF

The CRF represent one of the most visible successes of the CLARIN initiative. Researchers frequently express satisfaction with these curated resource lists, which often appear at the top of search results for specific queries, such as *corpora of computer-mediated communication*. While the CRF are currently composed of 27 families, the initiative originally started with four members, one of which was the CMC-RF<sup>8</sup>, thus demonstrating the particular value of this specific resource family.

Initially, the CRF were established in an ad-hoc manner, as the project's creators did not anticipate its eventual growth and success. For many years, information was maintained in several CSV files, typically ranging from 1 to 5 per family, depending on logical subdivisions (for example, monolingual vs. multilingual corpora). As the families expanded, so did these CSV files, eventually becoming cumbersome and difficult to manage.

In 2024, the CRF have transitioned to a more sustainable technical infrastructure. Each resource entry is stored in a single JSON file adhering to a strict schema. These files are hosted on GitHub<sup>9</sup>, enabling efficient tracking of changes over time. This new system allows external contributors, such as Knowledge Centres, to participate in curation by submitting pull requests. These contributions undergo automated validation against the JSON schema; if no issues are identified, the changes can be merged into the main branch by the central CRF curation team. Additionally, a pipeline has been implemented to import these JSON files from GitHub into the CLARIN website<sup>5</sup>, automatically generating family-specific overviews. While this final aspect is still under development, it is expected to be fully operational shortly.

This new technical infrastructure opened the possibility of decentralising and delegating work on single resource families to different stakeholder groups with greater expertise. To this end, CLARIN started to approach K-centres at the end of 2024, including the CKCMC, to inquire about their willingness to manage resources families, and this enquiry became the origin of our effort.

Shifting responsibility for curating the Resource Family from CLARIN ERIC to the CKCMC and the broader CMC community will likely solve specific problems, hiccups or inconveniences. Issues in the past have stemmed from the

<sup>7</sup><https://cmc-corpora.org/series>

<sup>8</sup>The others were *newspaper corpora*, *parliamentary corpora* and *parallel corpora*.

<sup>9</sup><https://github.com/clarin-eric/clarin-resource-families>

initially ad-hoc nature of the initiative, staffing constraints on the CLARIN ERIC side, and broader structural and technical challenges. Admittedly, some new challenges will arise for the CKCMC and the community to overcome. In general, we believe that this is an opportunity to strengthen the link between the CMC community, the CKCMC and CLARIN by bringing the tasks closer to the respective experts, those who have the knowledge, insight and ability to respond to challenges, and also have the interest to promote this specific resource family.

### 3. A Splash of CMC-RF Curation Criteria

In order to ensure the continued success of the CMC-RF and its capacity to serve the CMC community, the curated metadata will be openly licensed, and all procedures need to be properly documented and conducted transparently; additionally, community-driven criteria for what the CMC-RF should represent need to be identified. These include:

1. defining the *scope*, *goals* and *objectives* of the Resource Family, as well as establishing clear guidelines for the resources that should be included,
2. devising procedures to ADD new resources, UPDATE existing resources and DELETE resources that no longer adhere to the established criteria.

As mentioned, we intend to flesh out these criteria by holding a roundtable at the 2025 edition of the CMC-Corpora conference (CMC2025)<sup>10</sup>. Our reason for choosing such a first step is simple: this effort should be community-driven. Right from the start, our effort needs to be open to the community and should rely on input from the community. The annual edition of the CMC-Corpora conference, as the main scientific event of the CMC community, is naturally the most adequate context to do so.

In preparation for this roundtable, we present the preliminary criteria that we deem relevant – and the community-driven approach we would suggest to define the new curation workflow for the CMC-RF and refine it over the years. Nonetheless, it is worth noting that we only do so to get “the ball rolling” with the hope of fostering greater input from the roundtable.

CLARIN has established some broad curation criteria for the Resource Families (Lenardič and König, 2025). However, they are primarily concerned with ensuring consistency among the various families and do not cover the specificities of individual families. Regarding the inclusion criteria, CLARIN gives these guidelines:

- Resources should be hosted at a CLARIN repository<sup>11</sup>.
- Regarding granularity, the aim should be to include the entire resource rather than listing all possible subsets.
- Entirely inaccessible resources should not be included.

A very coarse characterisation of the aforementioned criteria could look like:

- Scope: curate a list of CMC and Social Media corpora that showcase interesting aspects of the data, for example, data representation, sociolinguistic features or research questions.

Of course, *interesting* would need to be defined.

- Goals: promote information about, and open access to diverse CMC resources, encourage reusability and support research in computer-mediated communication.
- Objectives: include varied and/or multiple modalities (text, audio, video), multiple languages, ensure FAIR compliance and maintain up-to-date resources.

The actual procedures agreed upon by the community could then be devised along these suggestions:

- ADD new resources: conduct a compliance check and take into account adherence to FAIR principles; desire citations from academic literature; consider search metrics from platforms like VLO; consider modalities and languages; consider corpus size details; strongly suggest an accompanying published paper.
- UPDATE existing resources: incorporate community feedback for improvements; align updates with evolving research interests and technical standards.
- DELETE resources: no longer accessible; non-compliant; outdated.

### 4. Recap and Next Steps

By delegating the curation of the CMC-RF, CLARIN ERIC has handed the CKCMC and, by extension, the CMC community a truly useful tool but also a somewhat delicate one. We believe that the CKCMC is the proper entity to perform such curation, while we also believe the CMC research community as a whole should be deciding on how such curation should be performed.

Accordingly, we suggest having a dedicated roundtable meeting, in addition to the presentation of this paper, at the 2025 edition of the CMC-Corpora conference to discuss and, hopefully, take collective decisions on the questions introduced in this paper.

Given that the roundtable will be of limited duration and will likely raise additional questions, we plan to define a preliminary roadmap and kickstart a taskforce of CMC community members willing to further refine it over recurrent online meetings after the conference.

While we cannot predict what will come out of the roundtable and the likely taskforce that will follow up on it, the CMC community has shown its capacity to undertake such an endeavour successfully over the years. As such, we believe it will welcome this opportunity.

### 5. References

Beißwenger, M. and Lungen, H. (2020). CMC-core: a schema for the representation of CMC corpora in TEI. *Corpus*, (20). <http://journals.openedition.org/corpus/4553>.

<sup>10</sup><https://www.cmc2025.uni-bayreuth.de/>

<sup>11</sup><https://www.clarin.eu/content/certified-b-centres>

- Fišer, D., Lenardič, J., and Erjavec, T. (2018). CLARIN's Key Resource Families. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1210/>.
- Frey, J.-C., König, A., and Stemle, E. W. (2019). How FAIR are CMC Corpora? In *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)*, pp. 25–30, Cergy-Pontoise University, France. <https://hdl.handle.net/10863/11294>.
- Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 1525–1531, Reykjavik, Iceland. European Languages Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2014/pdf/415\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf).
- Lenardič, J. and König, A. (2025). Guidelines for preparing CLARIN resource families. <https://office.clarin.eu/v/CE-2024-2451-guidelines-for-preparing-clarin-resource-families.pdf>.
- Lenardič, J., König, A., and Van Uytvanck, D. (2024). A Collaborative Approach to CLARIN Resource Families (poster). CLARIN Annual Conference 2024, Barcelona, Spain.
- Stemle, E. W., Frey, J.-C., König, A., Falaise, A., Erjavec, T., and Lungen, H. (2022). Introducing the CLARIN K(nowledge)-Centre for CMC and Social Media Corpora (CKCMC). In *Book of Abstracts of the 9th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC2022)*, pp. 46–47, Santiago de Compostela, Spain. <https://hdl.handle.net/10863/37043>.
- Stemle, E. W., König, A., and Nicolas, L. (2024). Collecting Metadata for Social Media Corpora in the Face of Ever-changing Social Media Landscapes. In Céline Poudat et al., editors, *Proceedings of the 11th Conference on CMC and Social Media Corpora for the Humanities*, pp. 75–78, Nice, France. Université Côte d'Azur. <https://shs.hal.science/halshs-04673776>.



# HopeEmo: A Bilingual Social Media Corpus for Emotion and Hope Speech Analysis

Wajdi Zaghouani<sup>1</sup>, Md. Rafiul Biswas<sup>2</sup>

<sup>1</sup>Northwestern University in Qatar, Education City, Doha, Qatar

<sup>2</sup>Hamad Bin Khalifa University, Doha, Qatar

Email: wajdi.zaghouani@northwestern.edu, mbiswas@hbku.edu.qa

## Abstract

This paper presents HopeEmo, a bilingual dataset of English and Arabic social media texts annotated for emotions and hope speech. With ~28,000 Arabic and ~10,500 English entries, it extends the existing corpora with annotations for emotion intensity, complexity, cause, and hope speech categories. Designed for social media research, HopeEmo supports the analysis of positive, inclusive content across diverse linguistic contexts. Baseline models, including AraBERT and BERT, demonstrate their potential for detecting hope speech, fostering healthier online communities.

**Keywords:** Emotion, Hope Speech, Dialectal Arabic, Social media, Sentiment

## 1. Introduction

Social media platforms are critical for online interactions, shaping public sentiment and emotional expression. While negative content like hate speech is widely studied, positive communication, such as hope speech, remains underexplored, particularly in multilingual contexts (Chakravarthi, 2020). The HopeEmo dataset addresses this gap by providing a bilingual (English and Arabic) social media corpus annotated for emotions and hope speech. The author extends the emotion work to hope speech in English and Arabic to observe how emotional content represents positivity in these languages. It supports natural language processing (NLP) applications that amplify positive content, promote inclusion, and healthier digital environments. Key contributions include:

- A large-scale dataset for emotion and hope speech analysis in social media.
- Insights into cross-cultural patterns of positive communication.
- Baseline models for enhancing inclusive online interactions.

## 2. Related Work

Analyzing hope speech from emotional context is not new and have been explored in multilingual setup (Arif et al., 2024; Malik et al., 2023; Subramanian et al., 2022). Emotion and hope speech are closely interconnected, playing significant roles in psychological disorders (Arif et al., 2024). Emotion analysis in social media has focused on English datasets, such as SemEval-2019 (38,424 texts) and TSEAR (7,665 texts) (Chatterjee et al., 2019; Wallbott and Scherer, 1986). Hope speech datasets exist for English, Tamil, Malayalam, and Spanish (Chakravarthi, 2020; Divakaran et al., 2024; Balouchzahi et al., 2023), but Arabic is underrepresented. Advances in Arabic NLP, like AraBERT (Antoun et al., 2020), highlight the need for culturally sensitive corpora. Our work integrates emotion and

hope speech annotations, tailored for multilingual social media analysis.

## 3. Dataset Creation and Annotation

### 3.1. Data Collection and Overview

HopeEmo was constructed from social media data (x formerly known as Twitter) to capture diverse emotional and hope speech expressions. For Arabic, we sourced data from three publicly available datasets: Arabic Poetry Emotions (9,500 rows from poetic texts shared on Twitter), Emotional Tone (10,000 rows from Arabic tweets), and Multi-label Hate Speech (~30,000 rows from various Arabic social media platforms) (Shahriar et al., 2023; Al-Khatib and El-Beltagy, 2017; Zaghouani et al., 2024). The keywords for searching in these platforms belong to the basic emotion terms: happiness, happiness, fear, satisfied, optimistic, confidence, trust, and other synonyms of emotion. These datasets were filtered to include texts with 5 to 80 words to ensure meaningful content while excluding overly brief or excessively long posts, resulting in ~28,000 Arabic entries. The word limit balanced the need for context (e.g., complete sentences) with the brevity typical of social media. Challenges included handling dialectal variations (e.g., Gulf vs. Levantine Arabic) and ensuring data privacy by excluding personally identifiable information, adhering to GDPR principles.

For English, we merged datasets from (Anjali, 2024) (5,000 rows from Reddit posts) and (Saravia et al., 2018) (7,000 rows from Twitter), yielding ~10,500 entries after applying the same word-length filter. English data collection faced challenges in harmonizing informal language (e.g., slang, emojis) and ensuring representativeness across platforms. Source URLs were retained for transparency, allowing traceability to original posts while respecting ethical data-sharing practices. Table 1 summarizes the dataset’s characteristics:

### 3.2. Dataset Creation

Creating HopeEmo involved merging and harmonizing disparate datasets to form a cohesive corpus suitable for social

Feature	Description
Languages	English, Arabic
Size	~28,000 Arabic entries, ~10,500 English entries
Modality	Text (social media: tweets, Reddit posts)
Annotation Labels	Emotion (intensity, complexity, cause); Hope Speech (Hope, Counter, Neutral, Hate/Negativity, with subcategories: e.g., Encouragement, Solidarity)
Sources	Arabic: Poetry Emotions, Emotional Tone, Hate Speech; English: Anjali (2024), Saravia et al. (2018)

Table 1: HopeEmo Dataset Characteristics

media analysis. For Arabic, the three source datasets had varying formats (e.g., CSV, JSON) and annotation schemas (e.g., different emotion labels). We standardized the data by converting all entries to a unified CSV format, normalizing text encoding (UTF-8), and mapping basic emotion labels (e.g., happiness, sadness) to a consistent set (e.g., joy, fear). To address dialectal diversity, we retained regional variations (e.g., Egyptian colloquialisms) but filtered out non-Arabic scripts. Ethical considerations included anonymizing user handles and removing sensitive content, ensuring compliance with data protection standards. We removed personally identifiable information such as address, location, phone number, email. We compiled a set of sensitive words dictionary (e.g., religion, ethnicity, gender, politics) and removed data if any of them fall into it. For English, merging Reddit and Twitter data required aligning timestamp formats and removing duplicate posts. We applied text cleaning to handle platform-specific features (e.g., hashtags, retweet markers), preserving informal elements like emojis that convey emotional nuance. A key challenge was balancing the dataset to avoid overrepresentation of specific platforms or topics (e.g., political tweets). We used stratified sampling to ensure diversity in content (e.g., personal narratives, public discussions), enhancing the dataset’s applicability to varied social media contexts.

### 3.3. Annotation Process

Annotation was conducted by five native speakers per language, selected for their linguistic expertise and familiarity with social media communication. Arabic annotators represented diverse dialects (Qatar, Tunisia, Jordan, Egypt), ensuring sensitivity to regional variations. English annotators included speakers from the US and UK to capture linguistic diversity. Recruitment prioritized gender balance and cultural diversity to minimize bias in interpreting emotional and hope speech content. Annotators extended basic emotion labels to include:

- **Emotion Intensity:** Low, Medium, High, or Not Applicable, assessing the strength of expressed emotions.
- **Emotion Complexity:** Simple, Medium, Complex, or Not Applicable, evaluating the layering of emotions.

- **Emotion Cause:** Identifying the trigger of the emotion (e.g., a life event, social issue).
- **Hope Speech:** Categorized as Hope Speech, Counter Speech, Neutral, or Hate Speech/Negativity, with subcategories (e.g., Encouragement, Solidarity).

Training spanned two weeks, with sessions focusing on social media-specific challenges, such as interpreting informal language, sarcasm, and emojis. Guidelines were developed iteratively, incorporating annotator feedback to clarify ambiguous cases (e.g., distinguishing Counter Speech from Neutral). Examples in both languages were provided, such as Arabic tweets expressing solidarity during crises and English Reddit posts offering encouragement. Conflicts were resolved through weekly discussions moderated by a manager, achieving high inter-annotator agreement (Fleiss’ Kappa: 0.65-0.81 for Arabic, 0.61-0.79 for English).

Bias mitigation strategies included stratified sampling to capture underrepresented groups, transparent filtering to preserve linguistic diversity, and regular audits of annotations to detect systematic errors. Social media-specific challenges, like interpreting context in short texts, were addressed by cross-referencing posts with their threads where possible. These efforts ensured a robust, inclusive dataset (Babanne et al., 2020).

## 4. Dataset Evaluation

We evaluated HopeEmo using transformer-based models (AraBERT for Arabic, BERT for English) and traditional methods (Logistic Regression, Naive Bayes). The preprocessing involved tokenization, stopword removal (Alrefaie, 2019) for Arabic; NLTK for English) and normalization (Devlin et al., 2018). Table 2 shows emotion and hope speech prediction performance. The evaluation metrics of prediction labels contains the categories of emotion of hope speech (multitasking work). We didn’t compute the subcategories of hope speech here to keep the result simple. BERT outperformed others for English (accuracy: 0.69),

Model	Language	Precision	Recall	F1-Score	Accuracy
AraBERT	Arabic	0.53	0.55	0.52	0.55
LR	Arabic	0.53	0.55	0.52	0.55
BERT	English	0.64	0.69	0.64	0.69
NB	English	0.49	0.53	0.52	0.53

Table 2: Model Performance Metrics

while Arabic models showed moderate performance, reflecting linguistic complexity. These results confirm the dataset’s utility for social media analysis.

## 5. Social Impact and Applications

HopeEmo supports research on positive communication in multilingual social media, promoting inclusion by identifying hope-inspiring content. Applications include social media moderation to amplify positive posts, language learning using positive examples, and cross-cultural analysis of emotional patterns, aligning with the conference’s focus on diversity and social benefits (Balouchzahi et al.,

2023; Babanne et al., 2020). The dataset is publicly available through an online repository <https://zenodo.org/records/15511565> and accessible via a consent form (<https://tinyurl.com/3wfvshjh>) ensuring GDPR compliance(Chintala, 2024). Due to privacy requirements under GDPR compliance, users accessing and analyzing social media data—even anonymized must acknowledge terms ensuring ethical usage. Therefore, a consent or terms-of-use form clarifies that the dataset should only be used for research purposes and prohibits attempts to identify or re-identify individuals.

## 6. Limitations

The construction of the HopeEmo dataset posed certain limitations. Since the raw data was collected without preserving user demographic details (e.g., gender, country, religion, race), we could not control or sample data based on these attributes. Consequently, the annotators received randomly sampled tweets without any demographic information. Currently, there are no available benchmark values from similar studies for comparison. Thus, we couldn't directly compare or reference our results. Future studies should aim to provide such reference points.

## 7. Conclusion and Future Work

HopeEmo is a novel corpus for studying emotion and hope speech in bilingual social media. Its focus on positivity and multilingualism supports inclusive online environments. Future work will incorporate multimodal data (e.g., images) and explore cross- platform patterns, fostering healthier digital communities.

## Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

## 8. Reference

- Al-Khatib, A. and El-Beltagy, S. R. (2017). Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 105–114. Springer.
- Alrefaie, M. T. (2019). Arabic stop words.
- Anjali, S. (2024). Emotion analysis based on text, March.
- Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- Arif, M., Shahiki Tash, M., Jamshidi, A., Ullah, F., Ameer, I., Kalita, J., Gelbukh, A., and Balouchzahi, F. (2024). Analyzing hope speech from psycholinguistic and emotional perspectives. *Scientific reports*, 14(1):23548.
- Babanne, V., Borgaonkar, M., Katta, M., Kudale, P., and Deshpande, V. (2020). Emotion based personalized recommendation system. *Int. Res. J. Eng. Technol.(IRJET)*, 7:701–705.
- Balouchzahi, F., Sidorov, G., and Gelbukh, A. (2023). Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.
- Chakravarthi, B. R. (2020). Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.
- Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.
- Chintala, S. (2024). Emotion ai in business intelligence: Under-standing customer sentiments and behaviors. *Central Asian Journal of Mathematical Theory and Computer Sciences*, 5(3):205–212.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Divakaran, S., Girish, K., and Shashirekha, H. L. (2024). Hope on the horizon: Experiments with learning models for hope speech detection in spanish and english. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, CEURWS. org.
- Malik, M. S. I., Nazarova, A., Jamjoom, M. M., and Ignatov, D. I. (2023). Multilingual hope speech detection: A robust framework using transfer learning of fine-tuning roberta model. *Journal of King Saud University-Computer and Information Sciences*, 35(8):101736.
- Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.
- Shahriar, S., Al Roken, N., and Zuolkernan, I. (2023). Classification of arabic poetry emotions using deep learning. *Computers*, 12(5):89.
- Subramanian, M., Chinnasamy, R., Kumaresan, P. K., Palanikumar, V., Mohan, M., and Shanmugavadivel, K. (2022). Development of multi-lingual models for detecting hope speech texts from social media comments. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 209–219. Springer.
- Wallbott, H. G. and Scherer, K. R. (1986). How universal and specific is emotional experience? evidence from 27 countries on five continents. *Social science information*, 25(4):763–795.
- Zaghouni, W., Mubarak, H., and Biswas, M. R. (2024). So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055.

# Tracking Ephemerality in YouTube Comments: Towards Methods for Building Dynamic Social Media Corpora

Yining Wang, Katrin Weller

GESIS – Leibniz Institute for the Social Sciences  
Unter Sachsenhausen 6-8, 50667 Köln, Germany  
Yining.Wang@gesis.org, Katrin.Weller@gesis.org

## Abstract

The ephemeral nature of social media content poses challenges for the development of social media corpora. This paper introduces a systematic methodology for collecting and tracking temporal changes in YouTube comments, highlighting the dynamic nature of online discourse. As a trial case study, we apply this approach to a dataset of 500 YouTube videos to propose a potential setting for building temporally-aware social media corpora. We collected 407,463 comments and monitored content changes over five consecutive days. During this period, 169,296 comments were added, 9,738 were removed, and 263 were edited. These findings demonstrate the extent of content volatility and underscore the limitations of static snapshot-based approaches. The proposed methodology offers a scalable framework for constructing corpora that better reflect the evolving character of social media communication.

**Keywords:** Social media discourse, Temporal corpus analysis, YouTube comments, Content volatility

## 1. Introduction

The ephemeral nature of social media content poses significant challenges for corpus linguistics and discourse analysis. Unlike traditional corpora that remain static after collection, online platforms host content that can be edited, deleted, or hidden at any time—even after being collected—creating fundamental difficulties for researchers.

This paper addresses a methodological gap in social media research by proposing a systematic approach to track temporal changes in social media discourse. Using YouTube comment sections as a trial case study, we demonstrate that conventional static data collection methods miss critical aspects of discourse activity. To address this, we introduce a dynamic monitoring framework designed to capture the full lifecycle of online commentary.

Our approach builds on YouTube’s hierarchical discussion structure and post-publication editing capabilities to develop a time-sliced corpus architecture. This enables the observation of discourse dynamics that remain invisible to snapshot-based collection methods, offering new insights into the temporality of online interaction.

## 2. Background and Related Work

Social media platforms have become valuable sources for linguistic research by offering naturally occurring language data at unprecedented scales. Current research practices often rely on large-scale API downloads, with most studies adopting a ‘snapshot’ approach that freezes data at the point of initial collection. This static perspective, however, overlooks important phenomena such as post-publication edits, platform-driven moderation, and user deletions, which imply that posts, tweets, and comments collected by researchers could vanish or transform posterior to data collection. This ephemeral nature of such contents thus presents unique challenges compared to traditional text corpora, where it is typically assumed that texts remain fixed, accessible, and unchanged over time. Social media content, by contrast, can be edited or deleted, undermining the stability

needed for reproducible and consistent linguistic analysis.

Androutsopoulos (2013) was among the early contributions that critically examined how the mutable and unstable nature of social media content challenges conventional understandings of what constitutes a linguistic corpus, while Zappavigna (2012) introduced the concept of “ambient affiliation” in Twitter studies, demonstrating that meaning emerges from continuously evolving discourse streams rather than fixed textual artifacts. Bolander and Locher (2014) identify four key methodological challenges in sociolinguistic research on computer-mediated communication. They advocate for research designs that consider the full lifecycle of online texts. Similarly, Hakimi et al. (2021) emphasize that digital trace data are inherently dynamic, as they are continually generated, altered, and repurposed across time and platforms. This ongoing mutability creates challenges for researchers seeking to capture stable datasets and calls for methodological approaches that reflect the evolving nature of digital environments.

In the case of YouTube, its extensive comment sections generate rich discourse data, but the platform’s architecture allows for multiple forms of content modification: users can edit or delete their comments after posting, content creators can moderate discussions through filtering and removal, and automated systems can flag content for potential violations. However, these important features of site were largely neglected by previous studies. While Burgess and Green (2018) explore the participatory nature of YouTube and its cultural significance, they do not address the technical mechanisms behind comment moderation in detail. Similarly, Thelwall (2018) developed a systematic method for analyzing YouTube comments using one-time data collection and sentiment analysis, but his approach does not account for how comments and interactions may change over time.

This study fills this gap by developing a novel corpus architecture specifically designed to capture the temporal dimensions of online discourse. While each video provides

contextual background for the discussion, our focus is on the evolving comment sections, which reflect dynamic user interaction over time. Through daily re-crawling of 500 YouTube videos and their associated comment threads over five consecutive days, combined with comment identifier tracking, we detect and log all additions, deletions, and modifications occurring between collection intervals. This approach enables researchers to observe language change patterns, measure moderation impacts, and analyze the social dynamics of online discourse communities in ways that conventional static corpora cannot support.

### 3. Methods

#### 3.1. Data Collection Architecture

Figure 1 provides an overview of our data collection pipeline, which consists of six main stages: sampling, filtering, crawling, storing, re-crawling, and differencing. The pipeline is implemented through three sequential Python programs that systematically capture and analyze YouTube comment dynamics over time.

The first program performs the initial crawl by connecting to the YouTube Data API v3, collecting video metadata, channel information, and complete comment threads, including nested replies. Use of the API is subject to YouTube’s Terms of Service and the YouTube API Services Terms of Service. Each result is stored as a JSON file to preserve the full data structure. The second program conducts temporal monitoring by recrawling the same videos at 24-hour intervals over five consecutive days, creating snapshots that capture comment evolution. The third program performs change detection by comparing snapshots to identify added, removed, and edited comments using precise comment identifier matching.

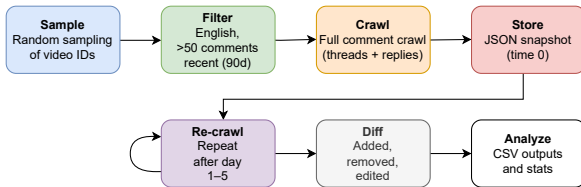


Figure 1: Overview of the data collection and analysis pipeline.

#### 3.2. Sampling Strategy

We employed a systematic sampling approach that cycles through YouTube’s four sort parameters (date, rating, relevance, viewCount) to ensure diverse content representation. This setup served as a case study to test our data collection approach, rather than being driven by a specific research question. Videos were included only if they were in English, contained at least 50 public comments, and were published within 90 days of collection. This process continued until 500 unique video IDs were gathered. For each video, we extracted comprehensive metadata, including video statistics, content details, and channel information, while preserving comment reply hierarchies to enable discourse analysis.

#### 3.3. Temporal Monitoring

Beginning 24 hours after the initial crawl, we performed five additional data collection cycles at fixed daily intervals (time 1 through time 5). Each snapshot was stored separately to enable comparison between any two time points without data loss, creating a longitudinal dataset spanning six days of comment activity.

#### 3.4. Change Detection

Our comparison algorithm identifies three types of comment changes between snapshots: additions (comments appearing in later but not earlier snapshots), removals (comments present initially but missing later), and edits (existing comments with modified content). Comments are matched using YouTube’s unique identifiers, with edits detected through API flags and content comparison. The system categorizes each video’s status and provides comprehensive statistics on all change types across the temporal sequence.

### 4. Corpus Composition and Dynamics

#### 4.1. Initial Corpus Characteristics

The baseline corpus collected at *time 0* contains 301,308 top-level threads and 106,155 replies, totaling 407,463 comments across 500 videos. This results in an average of 814.9 comments per video, indicating active user engagement across the dataset. The sample spans a broad range of YouTube content across fifteen categories.

*News & Politics* forms the largest segment with 126 videos (25.2%), followed by *Entertainment* with 118 videos (23.6%). Other notable categories include *People & Blogs* (73 videos, 14.6%), *Education* (56 videos, 11.2%), *Sports* (38 videos, 7.6%), and *Gaming* (37 videos, 7.4%). Smaller categories include *Comedy* (11 videos), *Music* (10), *Film & Animation* (9), *Science & Technology* (7), *Autos & Vehicles* (6), and *Howto & Style* (4). *Travel & Events* and *Pets & Animals* each contribute 2 videos, while *Nonprofits & Activism* is represented by a single video. This distribution reflects the diversity of YouTube discourse while maintaining a focus on videos with frequent user engagement in the comments.

#### 4.2. Temporal Dynamics and Change Patterns

Over the five-day monitoring period, we observed a large number of changes in the comment sections of the sampled videos. These changes fall into three categories: additions, removals, and edits, each following a distinct temporal pattern (Table 1).

Day	Added	Removed	Edited
1	72,318	1,494	186
2	108,283	2,520	229
3	135,343	4,720	240
4	153,395	5,565	251
5	169,296	9,738	263

Table 1: Cumulative number of comments added, removed, and edited by the end of each day, compared to the initial comment state at time 0.

Comment additions increased steadily throughout the monitoring period, rising from 72,318 on day one to 169,296 by day five. This trend indicates that YouTube videos continue to attract ongoing new engagement well beyond their initial publication. These added comments make up a large portion of the overall discourse and would be entirely missed by a static or one-time data collection.

Comment removals, although smaller in number, also increased over time, rising from 1,494 on day one to 9,738 by day five. This growth likely reflects both platform moderation and user-driven deletions. Since many removed comments were initially visible, their disappearance raises concerns about the temporal stability of online discourse and the replicability of observational research.

In contrast, comment edits remained rare and relatively stable, ranging from 186 to 263 edits per day. This suggests that most users do not revisit or revise their posts after publication, supporting the view that YouTube comments are generally treated as spontaneous and final rather than subject to revision.

These trends underscore the importance of repeated collection for capturing the full lifecycle of social media interactions. Additions and removals, in particular, demonstrate that online discourse evolves rapidly and is shaped by both user activity and platform interventions.

### 4.3. Corpus Volatility and Methodological Implications

These results highlight the limitations of static approaches to social media data collection. A snapshot collected only at *time 0* would have omitted over 169,000 comments that appeared later, accounting for more than 40% of the final dataset, and included nearly 10,000 comments that were later removed. Such differences pose serious risks of distortion in content analysis, topic modeling, or sentiment evaluation.

The evolving characteristics of YouTube comment threads shows the importance of temporally aware corpus construction. Our findings show that social media corpora are not stable archives but dynamic, socially shaped records. Ignoring temporal dynamics risks misrepresenting both the scale and content of public discourse.

By capturing a time-series of comment changes, our method provides a more accurate and comprehensive view of online conversational activity. It allows researchers to track how discussions develop, how moderation impacts visibility, and how public opinion shifts in near real time. This framework offers a practical path toward building corpora that reflect not only what users say, but also how and when those expressions are changed or withdrawn.

## 5. Discussion

### 5.1. Implications for Social Media Corpus Development

Our findings highlight the limitations of traditional static data collection methods in capturing the full scope of online discourse. The high rate of change observed within a five-day period shows that snapshots taken at only one point in time risk producing incomplete or unrepresentative corpora. Researchers relying on static datasets may overlook a

large portion of the discourse, particularly newly added or later deleted content.

Notably, the low rate of comment editing compared to much higher rates of additions and removals suggests that users typically engage with comment sections by posting new messages rather than modifying previous ones. This behavioral pattern has implications for studies of discourse development and language change in online communities, indicating that conversational progress often occurs through the addition of new messages and the removal of existing ones rather than through iterative revision.

### 5.2. Methodological Contributions

The architecture presented in this study enables detailed tracking of YouTube comment activity while considering limitations such as API quotas. Approaches like this may not scale easily beyond certain sample sizes, but the design helps mitigate such issues through efficient API key management and caching mechanisms. The system allows precise identification of content changes over time by consistently using unique comment identifiers, enabling exact matching across time points. It also enhances reproducibility by generating standardized outputs for added, removed, and edited comments. Its modular structure supports flexibility and adaptability to different observation periods, datasets, or platforms.

### 5.3. Technical Challenges and Solutions

Several technical challenges were addressed in building this system. The hierarchical structure of comment threads, including nested replies and @-mentions, was preserved through careful reconstruction of reply trees. Detecting edits required combining the API's `edited` flag with direct text comparison to reliably capture actual modifications. These solutions together form a robust foundation for future dynamic corpus projects that aim to track ephemeral user-generated content at scale.

### 5.4. Limitations and Future Work

This study was limited to a five-day observation period and focused on English-language YouTube content. Future research could extend the temporal scope to capture longer-term discourse changes, scaling the analysis to weeks or even months, to better understand persistent or delayed patterns of comment dynamics. It could also investigate whether patterns of volatility vary by content category, topic, or popularity. Cross-linguistic studies could explore whether similar dynamics occur in non-English comment sections, while cross-platform comparisons could reveal whether ephemerality functions differently on platforms with different user interfaces and moderation policies. Further investigation into the relationship between comment dynamics and video characteristics, such as age or engagement level, would also deepen our understanding of how content evolves in digital communication environments.

## 6. Ethical Considerations

All data collected in this study comes from publicly available YouTube content. We did not access private information, and no attempts were made to identify or contact users.

To further protect user privacy, we store only anonymized comment identifiers and avoid including usernames or channel names in any published material. The purpose of this research is methodological, focusing on patterns of content change rather than individual behavior. Future applications of this approach should continue to uphold ethical standards by respecting user anonymity and platform guidelines.

## 7. Conclusion

This paper introduced a comprehensive methodology for tracking ephemerality in YouTube comments and demonstrated that social media discourse is highly dynamic. We showed that static data collection methods fall short in capturing the full development of user contributions, as they miss later additions and preserve content that is eventually removed from the platform.

Our method supports the construction of temporally-aware social media corpora by preserving comment identifiers and systematically comparing snapshots over time. This approach reflects the evolving nature of comment threads, revealing not only what users express, but also how those expressions emerge, change, or disappear. As a result, it provides a more complete and accurate view of digital communication.

The five-day study showed that over 40% of comments were added after the initial crawl, while nearly 10,000 were removed. These patterns highlight the importance of incorporating temporal monitoring into corpus design. As social media platforms continue to change and user interaction patterns shift, accounting for ephemerality will become essential for robust and reliable research in corpus linguistics and digital discourse analysis.

This trial case study offers a clear direction for further development. The methodology presented here can be adapted to other platforms, extended across longer observation periods, or applied in cross-platform studies. Accounting for the temporal dimension of online discourse will enable researchers to better understand how digital conversations develop, evolve, and disappear over time.

## 8. References

- Androutsopoulos, J. (2013). Computer-mediated communication and linguistic landscapes. *Research methods in sociolinguistics: A practical guide*, pages 74–90.
- Bolander, B. and Locher, M. A. (2014). Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media*, 3:14–26.
- Burgess, J. and Green, J. (2018). *YouTube: Online video and participatory culture*. John Wiley & Sons.
- Hakimi, L., Eynon, R., and Murphy, V. A. (2021). The ethics of using digital trace data in education: A thematic review of the research landscape. *Review of educational research*, 91(5):671–717.
- Thelwall, M. (2018). Social media analytics for youtube comments: Potential and limitations. *International Journal of Social Research Methodology*, 21(3):303–316.

Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. Bloomsbury Academic.

# Deepfakes in Criminal Investigations: Interdisciplinary Research Directions for CMC Research

Lorenz Meinen<sup>1\*</sup>, Astrid Schomäcker<sup>1\*</sup>, Stefanie Wiedemann<sup>1,2\*</sup>, Markus Hartmann<sup>3</sup>,  
Timo Speith<sup>1</sup>, Lena Kästner<sup>1</sup>, Niklas Kühl<sup>1,2</sup>, Christian Rückert<sup>1</sup>

<sup>1</sup> University of Bayreuth, Bayreuth, Germany

<sup>2</sup> Fraunhofer FIT, Bayreuth, Germany,

<sup>3</sup> ZAC NRW, Cologne, Germany

E-mail: {firstname.lastname, astrid.schomaecker, lena.kaestner, kuehl, christian.rueckert}@uni-bayreuth.de

## Abstract

The emergence of deepfake technologies offers both opportunities and significant challenges. While commonly associated with deception, misinformation, and fraud, deepfakes may also enable novel applications in high-stakes contexts such as criminal investigations. However, these applications raise complex technological, ethical, and legal questions. We adopt an interdisciplinary approach, drawing on computer science, philosophy, and law, to examine what it takes to responsibly use deepfakes in criminal investigations and argue that computer-mediated communication (CMC) research, especially based on social media corpora, can provide crucial insights for understanding the potential harms and benefits of deepfakes. Our analysis outlines key research directions for the CMC community and underscores the need for interdisciplinary collaboration in this evolving domain.

**Keywords:** Deepfakes, law enforcement, interdisciplinarity, CMC research, social media corpus

## 1. Motivation

The development and widespread availability of generative AI bears the potential to fundamentally transform human communication. It opens up numerous possibilities, including creating synthetic media using convenient and ready-to-use tools. While some applications of this technology appear unproblematic (e.g., creating animations for school lessons), others pose significant risks. A particularly concerning example are deepfakes: AI-generated images, videos, or audio files, that convincingly simulate real individuals saying or doing things they never actually did.

As such, deepfakes are readily associated with deception and fraud, misinformation, or opinion manipulation (Verdoliva, 2020). While this sounds like they are primarily posing threats to society, there is a flip side: Deepfakes may open up possibilities to create useful, realistic media for various contexts. One such context, which has not yet received proper attention, is criminal investigations. In criminal investigations, deepfakes could be utilized to infiltrate criminal networks from afar and gain access to privileged information. To that end, investigators would rely on a subtype of deepfakes commonly referred to as (voice or video) *clones*. A clone is the digital impersonation of the voice and/or visage of a particular person, with capabilities of simulating said person in real time, effectively allowing for digital puppeteering. Deploying such clones potentially enables new strategies for infiltration and evidence gathering by impersonating members of a criminal organization in certain digital environments. This significantly lowers the risks and costs (in terms of both money and time) usually required to create undercover identities and gain access to relevant criminal circles. That way, deepfakes might prove vital in the fight against organized crime, which so far has been notoriously challenging to infiltrate.

However, the use of deepfakes in criminal investigation is far from straightforward. Apart from questions of technological feasibility there are significant ethical and legal concerns, particularly regarding the level of deception involved and the safety of the people “cloned.” These questions are not only morally but also legally highly relevant, especially since the publication of the AI Act (Regulation (EU) 2024/1689). Yet, scholarly discussions on the potentials and risks of utilizing deepfakes “for the greater good” have been sparse.

We believe that the question of how to (legally and morally) utilize deepfakes in criminal investigations needs to be tackled by an interdisciplinary approach. Additionally, in this paper, we suggest several avenues for further improvement with the help of CMC research. We believe that in doing so, our project underscores the potential of analyzing CMC corpora to understand and mitigate the risks associated with deepfakes—a topic which the CMC community has started to actively discuss (Russo, 2024). In the following, we first discuss questions arising from computer science (Section 2), philosophy (Section 3), and legal studies (Section 4), respectively; afterwards, we highlight some important questions that can only be addressed by taking an interdisciplinary perspective (Section 5). For all these questions, we suggest avenues of how their treatment could be supplemented with CMC research.

## 2. Technological Possibilities

In the context of deepfakes, the primary focus in computer science is on developing and improving increasingly powerful techniques for generating synthetic media and detecting and preventing such fabrications. The ability to create realistic deepfakes has advanced rapidly in recent years (Dragar et al., 2023). While previously manipulations were carried out manually or with simple tools and often required significant effort, resulting in so called “cheapfakes,” to-

---

\* Authors contributed equally.



day's deepfakes are based on complex deep learning architectures, such as Generative Adversarial Networks (GANs), Encoder-Decoder Networks (EDs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Verdoliva, 2020; Mirsky and Lee, 2022). In the context of clones used for criminal investigations, techniques based on these architectures, such as face swapping, face reenactment, and voice conversion, are considered promising but also have considerable potential for misuse.

The technical potential and limitations of these methods warrant careful analysis. Several factors play a role here, including the quality and quantity of training data required to train models that generate realistic deepfakes, data protection issues when using personal data, the availability of suitable hardware and computing resources, and the required expertise in machine learning (Hwang, 2020). These technical requirements and barriers vary considerably across user groups. While laypeople or criminal actors often benefit from access to user-friendly software tools, government agencies, such as law enforcement authorities, are subject to strict legal and ethical constraints, for example, regarding data protection and transparency, which can limit their scope of action.

As generative technologies continue to advance, so does research in the field of deepfake detection. Current detection methods typically rely on the analysis of spatial and temporal artifacts, such as blurred object boundaries, inconsistent contextual elements, missing or tampered watermarks, unnatural behavioral patterns, or asynchrony between speech and lip movements (Le et al., 2023). A thorough understanding of existing detection strategies, along with anticipation of future technical measures to prevent and counteract deepfakes, is crucial for two reasons. First, it allows for reliable identification of forgeries. Second, it ensures that deepfakes used in covert digital operations for criminal investigation purposes cannot easily be unmasked.

**Directions for CMC Research.** To better understand the technical possibilities and limitations of deepfakes, it is important to investigate their current use and creation in real communication environments. CMC researchers could contribute to this effort by answering the following questions: What types of deepfakes are used, and what technologies are used to create them? In this context, it would be useful to empirically investigate which software tools or platforms were used to create these deepfakes, as they may be identifiable based on metadata or file traces. Additionally, CMC research could address the important questions of how easy it is for non-experts to produce convincing deepfakes with publicly available tools and how easily laypeople are convinced by them. Additionally, CMC corpora containing deepfakes could serve as a resource for improving detection systems by evaluating existing methods and providing training material for machine learning-based detection programs.

### 3. Moral & Epistemological Implications

From the philosophical perspective, a central question concerns the ethical permissibility of deepfakes in general, and of clones within criminal investigations specifically. Ad-

ressing this question requires weighing potential harms against the potential benefits.

The philosophical literature highlights several negative consequences associated with deepfakes. Chief among them are deepfakes' capacity to deceive, their potential to erode epistemic trust in media (Fallis, 2021; Rini, 2020; Matthews, 2023), and the violation of the rights of depicted individuals (De Ruiter, 2021; Rini and Cohen, 2022). The relationship between deepfakes and our knowledge, i.e. the epistemological effect of deepfakes, is philosophically especially interesting. Concerning the (simplified) definition of knowledge as justified true belief (Ichikawa and Steup, 2024), deepfakes can affect all aspects of knowledge.

First, deepfakes inherently convey false information, which can lead to the formation of false beliefs with potentially serious consequences—for example, individuals may fall victim to scams by transferring money to impostors they mistakenly believe to be trusted business partners or loved ones, or voters may be misled into acting against their own interests due to deceptive political content. Second, with an increase in deepfakes, viewers might become more skeptical about the contents of media in general and hence less likely to believe the contents of any recording. Third, and philosophically most complex, the increase in deepfakes might undermine our justification to believe the contents of any audio-visual media.

The relationship between deepfakes and the rights of the depicted individuals presents a further but no less pressing issue. De Ruiter argues that the distinct moral wrong of deepfakes lies in portraying people in a way to which they would object (De Ruiter, 2021). Deepfakes can thus be seen to hurt several of a person's moral rights, including privacy, dignity and autonomy. To weigh the severity of the concerns, it can be useful to understand their relationship to the epistemological issues. Worries about the depicted person's rights might be considered less pressing if viewers are unlikely to attribute the contents to the person, either because the fake is easily detectable or because users have generally become skeptical.

These concerns must be balanced against potential benefits. It has been noted that deepfakes can be used beneficially within creative processes (Kerner and Risse, 2021), deepfakes of deceased individuals, so-called deathbots, could have therapeutic use for mourning relatives (Lindemann, 2022) and it has even been hypothesized that deepfakes can *increase* online trust (Etienne, 2021). Similarly, the use of clones by criminal investigators may have clear positive effects for law enforcement (and thus society at large): It can increase the effectiveness of criminal investigations while simultaneously reducing the risks investigators have to take and the resources that need to be invested. However, if we assume that there are different categories of deepfakes with different moral evaluations then the deepfakes that could be used for criminal investigations are probably among the morally most problematic: They need to resemble a real person convincingly and likely against that person's will, they are meant to deceive the recipient and can put the faked person into harms way. It thus requires careful consideration whether the positive outweighs the negative regarding the use of deepfakes by criminal investigators.

**Directions for CMC Research.** For several of these philosophical questions, it can be useful to investigate how deepfakes are used in CMC and how they are received. One approach can be to investigate the effects of deepfakes on the recipient's beliefs. How likely are internet users to perceive deepfakes as factual images? And how often do they question the veridicality of media shown to them? Can we identify conditions under which recipients are more or less likely to believe in the veridicality of an image? And have these behaviors changed since generative AI has become popularized? Answering these questions could help understanding the severity of the effects of deepfakes on our communication and knowledge. Regarding the rights of depicted persons, corpus analysis could be useful to investigate how closely distributed deepfakes resemble the target person and whether comments on such posts indicate that they can change a viewer's opinion of the person. Regarding the positive usage, a core question is whether we can identify positive uses of deepfakes in social media or elsewhere, what their numerical relationship is to problematic uses and whether in those cases the positive effects outweigh potential negative ones.

#### 4. Regulation & Legal Use of Deepfakes

Taking a look at the EU law, one finds a vivid discussion on the regulation concerning risks to privacy, issues arising from data protection legislation, such as GDPR and, most recently, the AI Act. A key question of these research fields is whether there is a need for more deepfake specific regulation, or if current legislation sufficiently addresses new issues. Even if there was sufficient regulation in theory, one key issue remains the lack of enforceability in the online environment due to anonymity and jurisdiction, rendering existing regulation ineffective (Lantwin, 2019).

A key tool in regulating people's behavior is criminal law. Since criminal law is a national affair, we take (following our expertise) a closer look at the German criminal law and criminal proceedings. Currently, there are efforts to criminalize the publication of deepfakes depicting picture-based sexual violence (CDU et al., 2025). This will most likely be done by passing the draft of §201b StGB (German Penal Code), which aims to criminalize the publication of deepfakes violating privacy and intimacy of the depicted (for details see: Bundesrat, 2024). This draft addresses concerning trends in online communication; however, it is debated whether there is a need for such a (deepfake-specific) law (Woerlein, 2024).

Deepfakes can not only be used in harmful ways but can also potentially be utilized by criminal investigators. The AI Act assumes the potential legality of clone use by investigators under EU law in Art. 50 (4), by exempting law enforcement from the requirement to disclose the synthetic nature of content for purposes of criminal investigations, if there is a legal basis for said use (for details: Pehli-van et al., 2024, p. 806). Said legal basis is also necessary under national law as every interference with rights requires a legal basis, outlining the extent of interference and the conditions clearly and in a comprehensible manner (in depth: BVerfG, 6.7.1999 - 2 BvF 3-90; see Jarass, 2024, Rn. 78 f.). Because of the severity of interference with pri-

vacy and informational self-determination the investigative powers of §§161, 163 Strafprozessordnung (StPO) do not extend to the creation and deployment of deepfake clones (cf. BVerfG 27.02.2008 - 1 BvR 370/07). Although some parts of creation might have a legal basis within the StPO, under the jurisprudence of the Bundesgerichtshof (BGH), the entire procedure needs to be grounded on a single legal basis (BGH 31.07.2007 - StB 18/06; for in-depth analysis: Rückert, 2023, p. 469 ff.). With German law lacking the latter, the creation and deployment of clones in law enforcement is not legal in Germany, as of now (Margerie and Hartmann, 2025). However, this is not the only issue concerning the legality of such practices. Both the rights of the person cloned and the person who is misled need to be considered to determine whether a legal basis for the deployment of deepfake clones by law enforcement could even be constitutional. As interference with privacy and informational self-determination weighs in considerably against the legality of said practices, it remains to be seen whether the German legislator undergoes the process of designing a legal basis for the deployment of deepfakes by law enforcement. For creation of a legal basis legislators will not only need to address national law, but also EU law, as law enforcement gathering personal data and creating a clone falls under Art. 10 Directive (EU) 2016/680. This means the legislator needs to consider the basic rights granted by the German constitution and the Charter of Fundamental Rights of the European Union. In summary, legislation regarding the use of deepfakes in criminal investigation needs to (i) answer the question for which crimes the deployment of clones shall be lawful, (ii) ensure that each individual act of interfering with the rights of the affected (gathering personal data as samples for clone creation, the creation itself, and the deployment of the clone) is addressed, (iii) clarify what safety measures are required when a person is "cloned," and (iv) be in line with the requirements of data protection legislation.

**Directions for CMC Research.** The necessity of deepfake specific criminal regulation in part depends on whether deepfakes are used in a harmful way by the public. Without a good understanding of the status quo of deepfake use and the extent of the harm caused by deepfakes one cannot adequately judge the need for more regulation. Here, CMC research can provide valuable insights by analyzing the use of deepfakes on social media. This analysis would also allow to draw conclusions relevant to understanding the feasibility of deepfake use by investigators, by examining how easily people are deceived by state of the art deepfake technologies. This also gives insights into the requirements for creating convincing deepfakes.

#### 5. Interdisciplinary Issues

Apart from the discipline-specific issues just outlined, there are a number of further inherently interdisciplinary concerns surrounding the use, legitimacy, and implications of the prevalence of deepfakes. These concerns show why our interdisciplinary approach is essential to adequately address the risks and possibilities deepfakes might imply in criminal investigations and beyond. Additionally, answer-

ing these research questions could also benefit from insights derived by CMC research.

**Trust in Testimony.** Linking legal and epistemological considerations, we must wonder how investigators' use of deepfakes within their investigations would square with the use of media as evidence in court. Do we run the risk of jeopardizing the trustworthiness of the legal system if investigators, on the one hand, use deepfakes to deceive suspects and gather evidence, and, on the other hand, provide different recordings as evidence in court? Apart from the general risk of deepfakes undermining trust in media, we need to ask whether the use of deepfakes by specific individuals or outlets undermines their individual perceived trustworthiness.

CMC research can provide guidance on such questions by analyzing how recipients' behavior changes toward individuals or media after they have been proven to (accidentally or intentionally) spread deepfakes or other fake media. It would be interesting to see whether information about the previous use of fakes leads recipients to be more skeptical in general and question the veridicality of content more than for other outlets. Towards this end, comment sections in social media could be analyzed to look for a change of sentiment.

**The Impact of Regulation.** At the intersection of law and computer science lies the impact of regulation or more precisely the question: How do we design regulation to be impactful and how do we need to implement technological tools to do so? As outlined above, the Achilles Heel of regulation in the online environment is often the enforceability, or rather the lack thereof. As good regulation is, among various other factors, characterized by actually impacting peoples behavior, this poses an issue, which needs to be addressed. Looking at deepfakes specifically, there are two angles new regulation could consider: Regulation could address either the individual or the platform (both are, of course, not mutually exclusive).

The individual can be addressed by penalizing the publication of certain kinds of deepfakes causing harm, while the platform might be obliged to deny the upload of those deepfakes and/or take down deepfakes under certain conditions. Both angles of regulation need to be supported by technological tools to maximize impact. If we decide to address the individual, we need better ways of overcoming online anonymity. If we decide to address the platforms, we need well-functioning upload filters capable of detecting deepfakes. This means deepfakes need to be disclosed as such in a machine-readable format (as required by Art. 50 (2) AI Act). However, how this obligation is actually fulfilled is a purely technological question. Possible approaches include metadata tagging, digital watermarks or cryptographic provenance systems, whereby the approaches differ in terms of the degree of reliability, manipulability and implementation effort.

Creating meaningful regulation can substantially benefit from empirical insights. CMC research could, for instance, address the following questions: Do people and/or platforms change their behavior when there is new regulation in place which can only be enforced to a limited extend? In

which formats are deepfakes actually labeled or disclosed in practice? Following on from this, how do people perceive the different methods of deepfake disclosure and how do they react to them?

**Corpora for Criminal Investigations.** The issue of how to devise corpora appropriate for criminal investigations lies at the intersection of computer science, philosophy and law. For one thing, as outlined in Section 2, sufficient quantity and quality of data must be available to train deepfake generation models in order for investigators to successfully use deepfakes—especially clones—to deceive recipients. At the same time, as outlined in Section 3, there are significant moral and ethical concerns regarding the potential misuse of deepfake technologies and the broader implications of deploying synthetic media in sensitive legal contexts. Finally, as outlined in Section 4, investigators must navigate a complex legal landscape that strictly regulates which types of data may be lawfully collected and used.

Addressing these challenges requires close interdisciplinary collaboration. Computer scientists must identify what kinds of data are required to meet specific technological objectives; notably, the type of data needed is dictated by the requirements for successful deception, which are only fulfilled when a synthetic artifact achieves perceived authenticity. Ethicists must evaluate the normative implications of using such data and technology in law enforcement. Finally, legal experts must determine whether the collection, processing and storage of the relevant data complies with existing laws and, if not, examine the possibilities of legalizing the collection, processing, and storage. Against this background, the crucial question becomes how and where relevant data for the creation of clones utilized in criminal investigations can be obtained, processed, and stored in a way that is both legally permissible and ethically sound. It may be interesting to examine whether or to what extent existing (e.g., social media-based) corpora can be legally used in this context. If so, how can we ensure that the data they contain meet the relevant standards?

Answering these questions requires not only technical insight into data adequacy and model performance but also a careful, interdisciplinary examination of the legal and ethical boundaries governing the use of data in criminal investigations as well as expertise in CMC research.

## 6. Conclusions & Outlook

AI-generated media—deepfakes in particular—bear the potential to change communication significantly. While there are clear risks associated with the increasing dissemination of synthetic media, their widespread availability also offers potentials, e.g., when deepfakes might be utilized “for the greater good” to support the rule of law. Yet, even such benevolent uses of deepfakes raise significant ethical and legal questions. To address these, an interdisciplinary perspective is essential. Adding to technological, philosophical, and legal expertise, we believe that systematic corpus-based investigations into how communication is affected by the presence of synthetic media can offer important insights contributing to an effective regulation of deepfakes in criminal investigation and beyond.

## 7. Acknowledgments

Work on this paper has been supported by the project “For the Greater Good? Deepfakes in Law Enforcement (FoGG)” funded by the Bavarian Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities, under the code KON-024-008.

## 8. References

- Bundesrat (2024). Entwurf eines Gesetzes zum strafrechtlichen Schutz von Persönlichkeitsrechten vor Deepfakes. Bundesratsdrucksache 222/24.
- CDU, CSU, and SPD (2025). Verantwortung für Deutschland – Koalitionsvertrag zwischen CDU, CSU und SPD. Published May 5, 2025.
- De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, 34(4), pp. 1311–1332.
- Dragar, L., Peer, P., Štruc, V., and Batagelj, B. (2023). Beyond detection: Visual realism assessment of deepfakes. *arXiv preprint arXiv:2306.05985*.
- Etienne, H. (2021). The future of online trust (and why deepfake is advancing it). *AI and Ethics*, 1(4), pp. 553–562.
- Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology*, 34(4), pp. 623–643.
- Hwang, T. (2020). Deepfakes: A grounded threat assessment. Technical report, Center for Security and Emerging Technology, Washington, D.C.
- Ichikawa, J. J. and Steup, M. (2024). The analysis of knowledge. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Fall 2024 edition.
- Jarass, H. (2024). Art. 20 GG. In Jarass, H., Pieroth, B., and Martin, K., editors, *Grundgesetz für die Bundesrepublik Deutschland*.
- Kerner, C. and Risse, M. (2021). Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, 8(1), pp. 81–108.
- Lantwin, T. (2019). Deep Fakes – Düstere Zeiten für den Persönlichkeitsschutz? – Rechtliche Herausforderungen und Lösungsansätze. *MMR Zeitschrift für IT-Recht und Recht der Digitalisierung*, pp. 574–578.
- Le, B., Tariq, S., Abuadbba, A., Moore, K., and Woo, S. (2023). Why do facial deepfake detectors fail? In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, pp. 24–28.
- Lindemann, N. F. (2022). The ethics of ‘deathbots’. *Science and Engineering Ethics*, 28(6), p. 60.
- Margerie, M. A. and Hartmann, M. (2025). Strafverfolgung mit Hilfe von Deepfake-Technologien – Realität oder Wunschdenken? *EuDIR – Zeitschrift für Europäisches Daten- und Informationsrecht*, pp. 84–89.
- Matthews, T. (2023). Deepfakes, fake barns, and knowledge from videos. *Synthese*, 201(2), p. 41.
- Mirsky, Y. and Lee, W. (2022). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), pp. 1–41.
- Pehlivan, C. N., Forgó, N., and Valck, P., editors (2024). *The EU Artificial Intelligence (AI) Act: A Commentary*. Wolters Kluwer, Alphen aan den Rijn, the Netherlands, 1 edition.
- Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers’ Imprint*, 201(4), pp. 1–16.
- Rini, R. and Cohen, L. (2022). Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy*, 22(2), pp. 143–161.
- Rückert, C. (2023). *Digitale Daten als Beweismittel im Strafverfahren*. Mohr Siebeck, Heidelberg, Germany.
- Russo, A. (2024). AI device for deradicalization process. In *Proceedings of the 11th Conference on computer-mediated Communication and Social Media Corpora*, p. 65.
- Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), pp. 910–932.
- Woerlein, A. H. (2024). Die Komplexität der sanktionsrechtlichen Gesetzgebung im Zusammenhang mit Deepfakes am Beispiel des geplanten §201b StGB. *MMR-Aktuell*.

# CRIME: The Corpus of Recorded Investigative, Media, and Evidence-based Proceedings

Steven Coats<sup>1</sup>, Dana Roemling<sup>2</sup>

<sup>1</sup>University of Oulu, Finland <sup>2</sup>University of Birmingham, UK

E-mail: steven.coats@oulu.fi, danaroemling@gmail.com

## Abstract

This paper presents CRIME: The Corpus of Recorded Investigative, Media, and Evidence-based proceedings, a structured, searchable resource comprising audio and ASR-generated transcripts from investigative interviews, courtroom interactions, and related media. Collected from publicly available YouTube sources under the EU Data Mining Act, the corpus addresses a critical gap in current research: the lack of large-scale, real-world datasets that integrate reliable transcripts with corresponding audio. Previous studies often rely on limited data, constraining generalizability and hindering methodological innovation. By enabling detailed analysis of linguistic, phonetic, pragmatic, and discourse-level features, CRIME supports interdisciplinary research in linguistics, law, psychology, and computational modeling. Future applications include the identification of language patterns associated with interviewing strategies and outcomes, as well as leveraging large language models to explore affective and interactional dynamics. This resource offers substantial potential to inform both academic inquiry and evidence-based practices in investigative interviewing and broader criminal justice contexts.

**Keywords:** corpus linguistics, YouTube, forensic linguistics, investigative interviewing, large-scale discourse analysis

## 1. Introduction

This paper introduces CRIME – the *Corpus of Recorded Investigative, Media, and Evidence-based proceedings* – a new structured and searchable language resource designed to support research at the intersection of language, crime, and justice. CRIME brings together high-quality Automatic Speech Recognition (ASR) transcripts and audio from three distinct but related domains: police investigative interviews, courtroom proceedings, and criminal-justice-related media content. The corpus is intended to facilitate linguistic, forensic, and interdisciplinary analysis by providing access to naturally occurring spoken data across a range of legal and quasi-legal contexts. In addition to its value for forensic and legal linguistic inquiry, CRIME also contributes to the study of Computer-Mediated Communication (CMC) by enabling analysis of spoken interactions captured, processed, and transmitted through digital technologies. In what follows, we review related work, describe the design and construction of CRIME, outline the processes used to collect and curate the data, and present a sample analysis to demonstrate the corpus’s potential for exploring discourse features in criminal justice settings.

## 2. Related Work

CRIME provides data for corpus-based research into language and discourse content in forensic linguistics, a specialized branch of applied linguistics which applies linguistic methods, approaches and knowledge to legal, investigative, and criminal contexts (see Coulthard et al., 2017). Forensic linguistics encompasses, for example, the analysis of language evidence such as ransom notes (Roemling & Grieve, 2024), but also the analysis of interview settings in legal, criminal or investigative contexts. Investigative interviewing refers to a non-coercive, evidence-based approach to interviewing suspects, witnesses, and victims, designed to gather accurate and reliable information while respecting the rights of the interviewee (see Meissner et al., 2023). Over

the past decade, interest in investigative interviewing has grown significantly, marking a clear departure from more confrontational or accusatory interrogation styles (e.g., Yuan, 2010), which have been shown to increase the risk of false confessions and unreliable testimony. The field has become increasingly interdisciplinary, drawing on insights from, for example, psychology, linguistics, and policing research (Denault & Talwar, 2023), and encompasses a broad range of topics. For instance, some studies have explored how different question types shape the course and outcome of interviews (see Oxburgh et al., 2010), while others have investigated how authority and interactional asymmetries are constructed within interview discourse (e.g., Madrunio & Lintao, 2024).

Researchers have also considered how institutional roles and hierarchies are enacted in broader legal contexts. For example, Rañosa-Madrunio (2014) draws on a small corpus of five interviews from the Philippines, while Tkačuková (2010) conducts a detailed single-case study of courtroom discourse in the U.S. Others have focused on how honesty, deception, or denial are constructed and negotiated in discourse: Stokoe (2010), working with a corpus of 120 UK police interviews, analyses how men deny accusations of violence, while Benneworth-Gray (2015) explores obligations of honesty using a smaller sample of three UK interviews. Carter (2014), in a single-case analysis, questions how deception is linguistically framed. Studies have also explored the role of language mediation in investigative interviews. Filipović (2008), for instance, draws from a corpus of 10,000 pages of U.S. police transcripts involving Spanish-English interactions. Additionally, researchers have turned their attention to how age (Heini, 2023; Jol & Van der Houwen, 2014) or disability (Pereira, 2024) affect communication in legal settings.

While research into investigative interviewing has progressed significantly, a key challenge remains: the scarcity of large, systematically compiled corpora that integrate both transcripts and corresponding audio. Much

of the existing work is based on limited datasets or individual case studies, which can fall short of capturing the full complexity of real-world interactions and limit their generalizability to other contexts. Furthermore, the reliability of transcriptions is often compromised, raising concerns about the accuracy and validity of subsequent analyses (Richardson et al., 2022). To address this gap, the following section introduces the *Corpus of Recorded Investigative, Media, and Evidence-based proceedings*.

### 3. Corpus

The corpus was created from content hosted on the YouTube platform, using a Python-based data-collection pipeline, a method increasingly common in dialectology, sociolinguistics, and other linguistic subfields (Coats, 2023). The approach relies on the stability and widespread use of common streaming protocols such as DASH (Dynamic Adaptive Streaming over HTTP; Sodagar 2011) or HLS (HTTP Live Streaming; Pantos & May 2017), which enable content including video, audio, and transcripts to be harvested.

Two YouTube channels devoted to criminal justice content comprise the majority of the corpus: *Court TV*, a U.S. television channel founded in 1991, and *Law & Crime Network*, an internet-based content provider founded in 2017. In addition to these channels, content from four YouTube playlists was harvested: *Full length criminal interrogations*, *Law and Crime interrogations* (a subset of the content from *Law & Crime*), *Interrogation raw*, and *Trial archives*. For each video in these channels and playlists, scripts collected YouTube’s own ASR transcripts, any other available transcripts uploaded for the video, and the full audio file for the episode, in .wav format. Data collection was undertaken with *yt-dlp*, an open-source Python library for harvesting content from YouTube and other platforms. After removal of duplicated content, ASR transcripts were tagged for part of speech using SpaCy’s *en\_core\_web\_sm* model (Honnibal et al. 2020); word timing tags from YouTube were retained. An overview of corpus size in terms of transcripts, word tokens, and audio duration is provided in Table 1.

Channel/Playlist	# trans.	# words (auto)	# words (other)	Length (hrs.)
Court TV	3,901	16,780,779	21,911,434	1,929.43
Law and Crime	19,111	114,723,611	6,820	1,7212.55
Law and Crime interrogations	21	113,973	0	12.35
Full length criminal interrogations	72	729,283	0	103.52
Interrogation raw	80	126,190	15,478	21.54
Trial archives	85	836,533	76,260	97.45
<b>Total</b>	<b>23,270</b>	<b>133,310,369</b>	<b>22,009,992</b>	<b>19,376.84</b>

Table 1: Corpus summary

The corpus provides access to transcript material from YouTube for purposes of research and education according to provisions of US and EU copyright law.<sup>1</sup> Two versions of the corpus exist: a static, downloadable table containing links to the source audio and transcript files, and an interactive searchable online version.

#### 3.1 Static Corpus

The static version of CRIME<sup>2</sup> contains, in tabular form, parsed transcripts, metadata fields automatically retrieved for the corresponding YouTube video by *yt-dlp*, as well as links to downloadable audio. Metadata fields are recorded in the columns *Playlist*, *Channel*, *ID*, *Title*, *URL*, *Description* (if any), *View Count*, *Duration (seconds)*, *Uploader*, *Uploader ID*, *Uploader URL*, *Thumbnails*, *Timestamp*, *Release Timestamp*, *Availability*, *Live Status*, *Channel Verified*, *auto\_transcript*, *other\_transcript*, *wav*, *timed\_auto*, *timed\_other*, *timed\_auto\_words*, and *timed\_other\_words*. The parsed, part-of-speech-tagged ASR transcripts in the *timed\_auto* column are suitable for linguistic analysis; the *timed\_other* column contains non-YouTube transcripts that have been uploaded for the video.

#### 3.2 Searchable online corpus

The online version of the corpus<sup>3</sup> contains the parsed, PoS-tagged YouTube ASR transcripts, the audio content, as well as most of the metadata information. The web interface provides search functionality in which transcript-linked 20-second audio segments are playable in the browser as mp3 files and downloadable. The corresponding video, as provided by the YouTube platform, can be viewed in an embedded window in the search interface. The preliminary version of the online corpus is hosted on infrastructure at Finland’s Centre for Scientific Computing; it comprises a customized version of BlackLab (De Does et al., 2017), implemented using OpenShift/Kubernetes container orchestration.

The online corpus permits targeted searches for transcribed utterances from specified content types, as indicated in the metadata fields. For example, Figure 1, a screenshot from

<sup>1</sup> The “Fair Use” provision of US copyright law (17 U.S.C. § 107) and EU Directive 2017/790 permit reproduction and use of copyrighted materials for purposes of research and education.

<sup>2</sup> <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MLMB6E>

<sup>3</sup> <https://forensic.corpora.li>



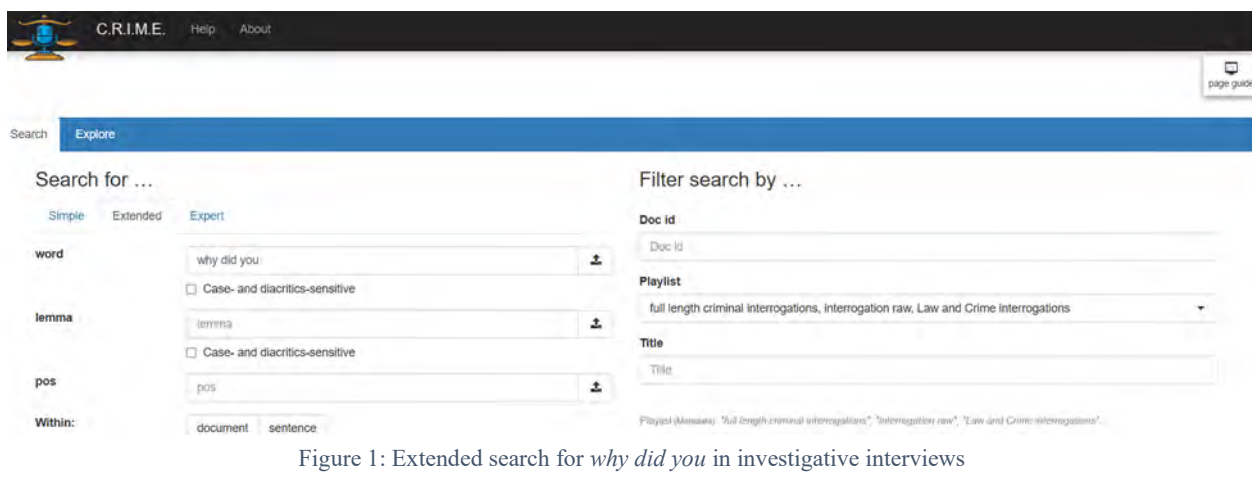


Figure 1: Extended search for *why did you* in investigative interviews

the online search interface, shows a search for the transcribed sequence *why did you* in which the “Playlist” field has been limited to the three playlists that contain transcripts of investigative interrogations.

#### 4. Potential Analyses

To illustrate the research possibilities enabled by CRIME, this section highlights potential analytical approaches. One such area involves deontic modal and semi-modal verb forms, which express necessity or obligation and are thus essential for the effective functioning of forensic, courtroom and legal proceedings. The expression of deontic modality has undergone a shift in 20<sup>th</sup>-century English, away from the standard *must* and towards the semi-modals *have to* and *need to* (Leech, 2003; Leech et al., 2009; Mair and Leech, 2020), but the use of deontic modals and semi-modals in legal contexts has mostly been restricted to analyses of legal documents and contracts, rather than speech in investigative interviews or courtroom proceedings. CRIME offers the opportunity to investigate the use of these items in forensic speech.

Another possible application for the corpus would be to investigate the use of epistemic stance adverbials, or expressions that express and delimit a proposition’s truth value in terms of semantic categories such as reality, certainty, or precision (Biber and Finegan, 1988; Hunston and Thompson, 2000). Stance markers such as *actual/actually* or *real/really* can serve to strengthen the

truth value of evidential claims, while markers such as *supposedly* or *allegedly* can be used to diminish them. Attitudinal stance markers such as *honest/honestly* or *truthful/truthfully* can be used in investigative interviews to elicit specific responses as well as to strengthen claims in courtroom proceedings. While some previous studies have analyzed use of some of these items in forensic contexts (e.g. Glougie, 2016), their systematic patterning in courtroom or investigative interview discourse has mostly not been considered on the basis of evidence from larger corpora.

Figure 2 shows the relative frequencies per million transcript words of the deontic modal and semi-modals *must*, *have to*, and *need to*; the epistemic stance adverbials *actual(ly)*, *real(ly)*, *supposedly*, and *allegedly*; and the attitudinal stance markers *honest(ly)* and *truthful(ly)*, according to corpus subsection. The more formal *must* is most frequent in the *Law and Crime* channel and in the *trial archives* playlist; the former contains scripted content, while the latter comprises transcripts of court proceedings. *Have to* and *need to* are more common in investigative interviews. For the stance markers, the most striking pattern is that the attitudinal markers *honest(ly)* and *truthful(ly)* are much more common in investigative interviews, presumably as admonitions of the interviewer to the interviewee. A more in-depth study could focus on these and other frequency patterns in the corpus.

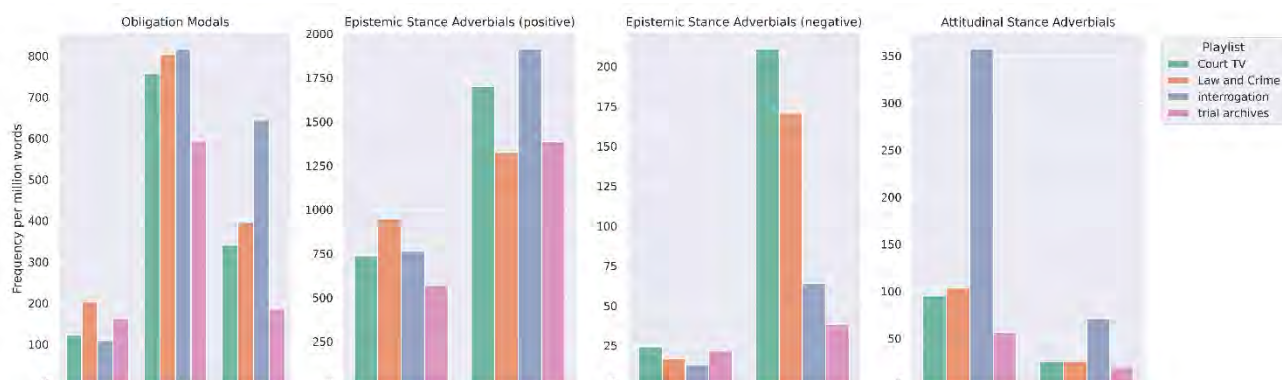


Figure 2: Relative frequencies of selected (semi-)modals and stance adverbials by corpus subsection

## 5. Caveats

While CRIME provides a large-scale resource for research, several caveats should be acknowledged. First, although ASR offers an efficient method for transcript generation, it is subject to transcription errors, particularly in cases of overlapping speech, strong regional accents, low audio quality, or use of out-of-vocabulary items such as some legal terminology. These errors may affect the precision of fine-grained linguistic analysis and should be taken into account when interpreting results. Second, YouTube ASR transcripts are undiarized: there are no indications of speaker turns, and therefore associating transcript segments with individual speakers requires manual annotation. Finally, as the corpus draws on publicly available YouTube channels or playlists, there is some inconsistency in the structure and content of the source material, particularly for content retrieved from the much larger *Court TV* and *Law and Crime* channels. These channels contain a mixture of content types, including “true crime” entertainment, crime and criminal justice news, commentary on court cases, and excerpts from investigative interviews and trial recordings. Researchers interested in, for example, the speech content of investigative interviews, will need to filter metadata fields such as “playlist” in order to exclude transcripts such as expert commentary or courtroom proceedings. Although some of the corpus content can be disambiguated for interaction type or for genre/register with targeted searches based on video title substrings, the variability in content type, as well as format, speaker roles, recording quality, and recording context may introduce noise or limit comparability across files.

## 6. Conclusion

This paper has introduced the *Corpus of Recorded Investigative, Media, and Evidence-based proceedings*. By addressing the scarcity of large-scale, real-world corpora of speech from legal contexts with aligned audio and transcript data, CRIME enables new forms of empirical analysis across disciplines like linguistics, law, psychology, and computational modeling. The corpus is already suited for a range of fine-grained linguistic investigations. For example, we highlighted how it can support the analysis of epistemic stance adverbials and deontic modal and semi-modal verb forms - features central to meaning-making in legal and investigative discourse.

Looking ahead, further development of CRIME will focus on expanding the dataset and improving transcription accuracy. New data can be added to the corpus by retrieving and parsing content that has been more recently uploaded to the targeted YouTube playlists and channels; in addition, the corpus could be expanded through the inclusion of material from other online sources. Another planned upgrade for CRIME is the implementation of larger and potentially more accurate ASR models, as well as transcript diarization, via a pipeline that incorporates Whisper, WhisperX, pyannote, or similar tools (Radford et al. 2022, Bain et al. 2023, Bredin et al. 2023). These improvements aim to enhance the corpus’s value as a resource for

interdisciplinary research and applied work in criminal justice communication.

## 7. References

- Benneworth-Gray, K. (2015). ‘Are you going to tell me the truth today?’: Invoking obligations of honesty in police-suspect interviews. *International Journal of Speech Language and the Law*, 21(2), Article 2. <https://doi.org/10.1558/ijssl.v21i2.251>
- Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of Interspeech 2023*, pp. 4489–4493. <https://doi.org/10.21437/Interspeech.2023-78>
- Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse Processes*, 11, 1–34. <https://doi.org/10.1080/01638538809544689>
- Bredin, H. (2023). Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark and recipe. In *Proceedings of Interspeech 2023*, pp. 1983–1987. <https://doi.org/10.21437/Interspeech.2023-105>
- Carter, E. (2014). When is a lie not a lie? When it’s divergent: Examining lies and deceptive responses in a police interview. *Language and Law*, 1, 19.
- Coats, S. (2023). Dialect corpora from YouTube. In B. Busse, N. Dumrukcic, & I. Kleiber (Eds.), *Language and Linguistics in a Complex World* (pp. 79–102). Berlin: De Gruyter. <https://doi.org/10.1515/978311017433-005>
- Coulthard, M., Johnson, A., & Wright, D. (2017). *An Introduction to Forensic Linguistics: Language in Evidence* (2nd edition). Routledge, Taylor & Francis Group.
- De Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries*. London: Ubiquity Press, pp. 245–257. <https://doi.org/10.5334/bbi.20>
- Denault, V., & Talwar, V. (2023). From criminal interrogations to investigative interviews: A bibliometric study. *Frontiers in Psychology*, 14, 1175856. <https://doi.org/10.3389/fpsyg.2023.1175856>
- Filipovic, L. (2008). Language as a witness: Insights from cognitive linguistics. *International Journal of Speech Language and the Law*, 14(2), Article 2. <https://doi.org/10.1558/ijssl.2007.14.2.245>
- Glougie, J.R.S. (2016). *The semantics and pragmatics of English evidential expressions: The expression of evidentiality in police interviews*. Ph. D. thesis, University of British Columbia. <https://doi.org/10.14288/1.0319268>
- Heini, A. (2023). ‘Basically, I’m gonna ask you a load of questions’ Cautioning exchanges in police interviews with adolescent suspects. *Language and Law* 9(2), 11–31. [https://doi.org/10.21747/21833745/lanlaw/9\\_2a3](https://doi.org/10.21747/21833745/lanlaw/9_2a3)
- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *SpaCy: Industrial-strength Natural Language Processing in Python*. <https://doi.org/10.5281/zenodo.1212303>
- Hunston, S., & Thompson, G. (2000). *Evaluation in Text:*



- Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press.
- Jol, G. A. H., & Van der Houwen, F. (2014). Police interviews with child witnesses: Pursuing a response with maar (= Dutch but )- prefaced questions. *International Journal of Speech, Language and the Law*, 21(1), Article 1. <https://doi.org/10.1558/ijssl.v21i1.113>
- Leech, G. (2003). Modality on the move: The English modal auxiliaries 1961–1992. In R. Facchinetti, F. Palmer, & M. Krug (Eds.), *Modality in Contemporary English*. Berlin: De Gruyter, pp. 223–240. <https://doi.org/10.1515/9783110895339.223>
- Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge University Press.
- Madrugno, Ma. K. J. R., & Lintao, R. B. (2024). Power, control, and resistance in Philippine and American police interview discourse. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, 37(2), 449–484. <https://doi.org/10.1007/s11196-023-10045-8>
- Mair, C., & Leech, G. N. (2020). Current changes in English syntax. In B. Aarts, A. McMahon, & L. Hinrichs, (Eds.), *The Handbook of English Linguistics*. London: Wiley, pp. 249–276. <https://doi.org/10.1002/9781119540618.ch14>
- Meissner, C. A., Kleinman, S. M., Mindthoff, A., Phillips, E. P., & Rothweiler, J. N. (2023). Investigative interviewing: A review of the literature and a model of science-based practice. In D. DeMatteo & K. C. Scherr (Eds.), *The Oxford Handbook of Psychology and Law* (1st ed., pp. 582–603). Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780197649138.013.34>
- Oxburgh, G. E., Myklebust, T., & Grant, T. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *International Journal of Speech Language and the Law*, 17(1), Article 1. <https://doi.org/10.1558/ijssl.v17i1.45>
- Pantos, R., & May, W. (2017). HTTP Live Streaming (RFC 8216). Internet Engineering Task Force (IETF). <https://doi.org/10.17487/RFC8216>
- Pereira, T. (2024). Establishing common ground using low technology communication aids in intermediary mediated police investigative interviews of witnesses with an intellectual disability. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, 37(2), 517–546. <https://doi.org/10.1007/s11196-023-10035-w>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356 [eess.AS]*. <https://doi.org/10.48550/arXiv.2212.04356>
- Rañosa-Madrugno, M. (2014). Power and control in Philippine courtroom discourse. *International Journal of Legal English*, 2(1), 4–30.
- Richardson, E., Haworth, K., & Deamer, F. (2022). For the record: Questioning transcription processes in legal contexts. *Applied Linguistics*, 43(4), 677–697. <https://doi.org/10.1093/applin/amac005>
- Roemling, D., & Grieve, J. (2024). Forensic Authorship Analysis. *CREST Security Review*, #18: Communication
- Sodagar, I. (2011). The mpeg-dash standard for multimedia streaming over the internet. *IEEE Multimedia*, 18(4), 62–67.
- Stokoe, E. (2010). ‘I’m not gonna hit a lady’: Conversation analysis, membership categorization and men’s denials of violence towards women. *Discourse & Society*, 21(1), Article 1.
- Tkačuková, T. (2010). The power of questioning: A case study of courtroom interaction. *Discourse and Interaction*, 3(2), 49–61.
- Yuan, C. (2010). Avoiding revictimization: Shifting from police interrogations to police interviewing in China. *International Journal of Speech Language and the Law*, 16(2), 293–297. <https://doi.org/10.1558/ijssl.v16i2.293>

# Dimensions of Drivel in German Telegram Posts

## Manual Annotation and Predictive Power

Andreas Blombach<sup>1</sup>, Stephanie Evert<sup>1</sup>, Linda Havenstein<sup>2</sup>, Philipp Heinrich<sup>1</sup>

<sup>1</sup>Lehrstuhl für Korpus- und Computerlinguistik      <sup>2</sup>Lehrstuhl für Japanologie

Friedrich-Alexander-Universität Erlangen-Nürnberg

<sup>1</sup>Bismarckstr. 6, 91054 Erlangen      <sup>2</sup>Artilleriestr. 70, 91052 Erlangen

{first.name}. {last.name} @fau.de

### Abstract

Social media platforms abound with unsubstantiated claims and conspiratorial or conspiracy-adjacent discourse. In this paper, we aim to quantify the overall drivel-like quality of posts along six dimensions, each assessing a different aspect. Six student assistants annotated a random sample of 1,000 Telegram posts from well-known German conspiracy theorists, based on carefully formulated guidelines to ensure consistency. Each dimension was rated on a scale from 1 to 5, and inter-annotator agreement was evaluated using Krippendorff's  $\alpha$  for ordinal scales, revealing moderate to substantial agreement across dimensions. We also report experiments on predicting the overall drivel-like quality of posts from these dimensions using a simple linear regression model. It shows that posts are considered drivel overall in particular when their contents appear distant from reality and when authors strongly assert their views.

**Keywords:** Telegram, conspiracy theories, drivel, annotation

## 1. Introduction

Conspiratorial, conspiracy-adjacent, esoteric, or pseudo-scientific discourse – often referred to in German as *Geschwurbel* – presents a persistent challenge in contemporary digital spaces, particularly on social media (Douglas et al., 2019). Such discourse is not always explicit or fully formed; rather, it frequently consists of vague insinuations, rhetorical devices, and partial references that can easily evade traditional classification systems.

Recent research has largely focused on identifying and categorising conspiracy narratives (Heinrich et al., 2024; Piskorski et al., 2025). However, with the exception of a few recent schemes (Piskorski et al., 2023), classification frameworks tend to ignore subtle cues such as rhetorical strategies, emotive language, or ambiguous references, which contribute to the conspiratorial tone without adhering to a concrete topic or narrative. Instead, much of this work has focused explicitly on narrative detection and binary classification of drivel. Many posts lacking a clear narrative – containing only hints or allusions – often appear vaguely conspiratorial but ultimately slip through the classification framework. Furthermore, since (some) narratives evolve over time (some fade while new ones emerge), relying solely on narrative structures may miss important cross-narrative patterns.

These observations motivated the development of a multi-stage scale to better capture the varying degrees (or dimensions) of drivel. Specifically, we aim to identify cross-narrative features of drivel as a step toward a broader investigation into its underlying properties beyond rigid narrative boundaries. In the present contribution, we propose six key dimensions for annotating drivel within a given text:

1. its distance from reality,
2. its linguistic and argumentative peculiarities,
3. claims to absoluteness (and overall handling of sources),
4. its suggestiveness,
5. its tendency to oversimplify complicated matters, and

6. the (apparent) emotionality of its author.

We present the development of a gold standard of 1,000 manually annotated German Telegram posts, including detailed guidelines and an analysis of inter-annotator agreement. Furthermore, we try to predict the overall drivel-like quality of a text from the annotated dimensions. In this context, we also analyse the correlations between these dimensions.<sup>1</sup>

## 2. Data

In 2020, as platforms like YouTube and Facebook intensified efforts to curb disinformation during the COVID-19 pandemic, skeptics, lockdown critics, and conspiracy theorists began migrating to alternative networks. Much of the text- and image-based discussion moved to Telegram – a minimally moderated messaging and microblogging platform – making its channels and groups key data sources for studying conspiracy theories (Lamberty et al., 2022; Holnburger et al., 2022).

### 2.1. Schwurpus: a corpus of conspiratorial talk

We use channels of prominent German COVID-19 conspiracy figures scraped via Telegram's export function (Heinrich et al., 2024). Since channels often interact through message forwarding, the corpus was expanded by iteratively including frequently mentioned channels with large follower counts, supplementing this with publicly available channel statistics. The final corpus – called “Schwurpus” – includes over 200 channels (followers ranging from a few thousands to over 300,000) and more than 100 public group chats from January 2020 to July 2022, totaling over 13 million posts and nearly 400 million tokens.

### 2.2. Sample

We drew a random sample from the Schwurpus for manual annotation. Only posts with 400 or more characters were

<sup>1</sup>The sample, guidelines, and adjudicated annotations can be found at <https://github.com/fau-klue/infodemic>.

considered. To ensure balanced representation and avoid bias towards highly active channels, the data were stratified by month and by channel frequency category. Two samples were initially drawn: the first consisted of approximately 1,000 posts and was originally used for the automatic detection of narratives related to the pandemic (Heinrich et al., 2024)<sup>2</sup>, while the second comprised roughly 2,000 posts and included data from the entire Schwurpus. For the final dataset, both samples were merged and posts were sorted chronologically. To introduce narrative variation for initial testing, the first 100 posts were randomly sampled from the entire dataset.

In the present contribution, we present the results based on the first 1,000 posts of this combined dataset. The dataset contains posts dated between January 1, 2020, and July 29, 2022. The majority of posts were made during early to mid-2020, specifically between January and September. Only 83 posts were created after October 1, 2020. The posts originate from a total of 143 distinct channels. The channels with the highest number of posts are *evahermanof-fiziell* (50 posts), *qglobalchange* (42 posts), *alternativemedien* (40 posts), *kulturstudio* (39 posts), and *oliverjanich* (35 posts). Texts are rather short, ranging from 56 to 1,024 tokens per post, with a median of 144 tokens and a mean of 208 tokens.

### 3. Annotation: dimensions of drivell

Since we assume that texts can be more or less drivell-like, it makes sense to visualise the opposite poles: at one end of the scale are incoherent texts full of far-fetched assertions without evidence, which are nonetheless presented in a tone of conviction; at the other end are fact-based, scientific texts with clean argumentation.

To characterise drivell, we defined six different dimensions. Student assistants rate texts on every dimension with values between 1 and 5. In addition, the overall drivell-like quality of a given text is also rated on the same scale (intuitively, without further instructions).

#### 3.1. Guidelines

The guidelines to annotate posts include detailed descriptions of the six proposed dimensions of drivell, typical features, as well as fully annotated examples. Annotators were instructed to avoid middle values for inconspicuous texts, and to rate dimensions independently of each other.

**Dimension 1: distance from reality.** This category is concerned with how far a text departs from widely accepted reality, especially scientific consensus, by assessing the plausibility and number of assumptions required to believe its claims. The lowest rating indicates fact-based or experience-based content requiring no assumptions, while the highest indicates completely fabricated, fantastic or conspiratorial content requiring numerous implausible assumptions. Intermediate levels reflect increasing reliance on questionable premises, half-truths, unverifiable personal beliefs, or spiritual/religious claims with real-world implications.

**Dimension 2: linguistic and argumentative peculiarities.** This category assesses the linguistic and argumentative clarity of a text, focussing on whether conclusions logically follow from the stated premises. The lowest rating is reserved for clear, logical and coherent argumentation as well as for non-argumentative texts (e.g. purely social interactions). The highest rating indicates completely incoherent ramblings and texts where the argumentation is incomprehensible or outright missing (despite making claims). Key features include logical gaps, semantic incoherence, accumulations of grammatical or spelling mistakes, associative rather than logical reasoning, informal fallacies, personal attacks, and clickbait style. As these features become more frequent and disruptive, the rating increases.

**Dimension 3: claim to absoluteness and handling of sources.** This category evaluates how strongly authors assert their views – especially bold or controversial ones – and, to a lesser extent, how they handle sources. The lowest score reflects cautious, balanced language, hedging expressions, and a thoughtful, open engagement with evidence and alternative or opposing views. As scores increase, authors appear more convinced of their own perspective, use fewer qualifiers, and rely on anecdotal or selectively interpreted evidence. Higher scores indicate ideological rigidity, a lack of self-reflection, disregard or disparagement of differing views, manipulative use of sources and/or reliance on dubious ones. The highest score denotes an absolutist stance with unquestioned beliefs and missionary zeal.

**Dimension 4: suggestiveness.** This category assesses how much a text subtly tempts readers to draw unjustified conclusions without explicitly stating them. Key features include subtext, dogwhistling, implications, rhetorical questions, framing, loaded language, manipulative contrasts (splitting/black-and-white thinking), intentional use of fallacies, and emotionalisation. The lowest score is intended for texts free of suggestive elements. As the score rises, so does the presence of suggestive features, with the highest score indicating a clearly recognisable suggestive intention. Note that some of the features used here are also used in the annotation of framing and persuasion techniques (Piskorski et al., 2023).

**Dimension 5: oversimplification.** This category evaluates how much a text oversimplifies complex issues, particularly by presenting single causes for multifaceted problems or simply omitting key aspects. The lowest score indicates a nuanced, well-rounded presentation that respects a topic’s inherent complexity and includes (nearly) all important factors. Increasing scores reflect texts beginning to generalise, ignore counterarguments and -evidence, and reduce explanations to fewer causes. At the highest level, the portrayal is extremely reductive – presenting only one factor as the sole explanation of one or even multiple issues, often in a way that distorts understanding and disregards complexity entirely.

**Dimension 6: emotionality.** This category assesses how emotional the author appears to be in their writing, be it angry, enthusiastic, despairing, or anxious. Key features of emotionality include words that explicitly refer to the author’s emotional state, use of expressive emojis, dramatic

<sup>2</sup>As this sample was drawn earlier, it does not include data from 2022.

punctuation, and fully capitalised words or sentences. The lowest score indicates a neutral, objective tone with little to no emotional expression, whereas the highest is assigned to texts that are dominated by a very emotional or agitated tone.

### 3.2. Manual annotation

We employed a total of six annotators, with three participating in the first (finished) phase of the annotation process and another three in the second phase. In total, annotators 1–3 contributed 995 annotated posts (i.e. they annotated the complete sample). Annotators 4 and 5 provided an additional 300 annotated posts, and annotator 6 contributed a further 180 annotated posts – these annotations were predominantly executed for training the second batch of annotators, who are now annotating the next batch of the overall sample.

Figure 1 illustrates the score distributions for each annotator (note the varying  $y$ -axis scales, as the overall number of annotations differs across annotators). We observe that annotators 2 and 5 display a clear tendency to assign low scores, with annotator 5 in particular frequently opting for score 1. Annotator 6, by contrast, tends to assign scores 3 and 4 most of the time. In comparison, annotators 1, 3, and 4 exhibit a more uniform distribution across scores, though annotators 1 and 4 notably avoid the highest score (score 5). It is important to note that it is generally easier to reach high agreement when annotators consistently choose the middle of the scale (see below for inter-annotator agreement) but annotators were instructed to make clear-cut decisions and to avoid systematically choosing the middle scores.

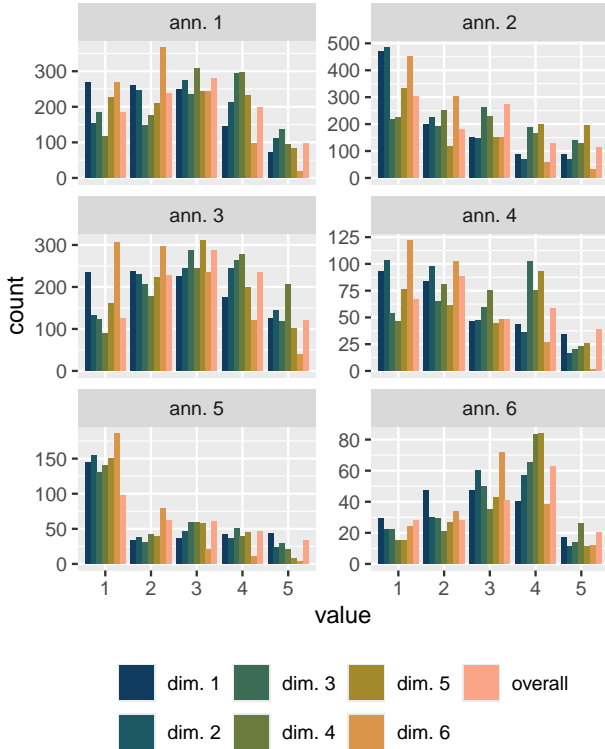


Figure 1: Distribution of annotation scores per annotator.

Additionally, manual adjudication was performed by two of

the authors in collaboration with the annotators, focussing on the most challenging cases, i.e. those exhibiting the highest disagreement. This process was very slow, concentrating on difficult cases to ensure quality and improve understanding of each dimension. In total, 198 annotations were manually discussed and curated across all dimensions, plus 13 annotations regarding the overall driveline quality.

### 3.3. Inter-annotator agreement

**Measuring agreement** Inter-annotator agreement (IAA) scores were computed using Krippendorff’s  $\alpha$  for ordinal data (Landis and Koch, 1977), where the scores are treated as categories with inherent ranks. This measure can be calculated for pairs of annotators or for groups. Importantly, it accounts for the order of categories by employing a difference function  $d_{ij}$  that reflects the distance between categories  $i$  and  $j$ . In our analysis, we use the squared difference of ranks as the difference function:

$$d_{ij} = (r_i - r_j)^2$$

where  $r_i$  and  $r_j$  denote the ranks of categories  $i$  and  $j$ , respectively. The agreement coefficient is computed as

$$\alpha = 1 - \frac{D_o}{D_e}$$

where

$$D_o = \frac{\sum_{i,j} n_{ij} d_{ij}}{N}, \quad D_e = \frac{\sum_{i,j} n_i n_j d_{ij}}{N(N-1)}$$

represents the observed agreement and expected disagreement by chance, respectively. Here,  $n_i$  denotes the total number of times category  $i$  was assigned,  $N$  is the total number of assignments, and  $n_{ij}$  denotes the number of pairs of assignments where one annotation was given category  $i$  and the other category  $j$ .

Krippendorff’s  $\alpha$  ranges from  $-1.0$ , indicating perfectly discordant annotations, to  $+1.0$ , indicating perfect agreement. Negative values indicate agreement worse than chance. The interpretation of  $\alpha$  values follows commonly accepted guidelines (Landis and Koch, 1977, 165): values between  $-1.0$  and  $0.0$  indicate poor agreement; values greater than  $0.0$  up to  $0.2$  are considered slight; from  $0.2$  to  $0.4$  fair; from  $0.4$  to  $0.6$  moderate; from  $0.6$  to  $0.8$  substantial; and values above  $0.8$  up to  $1.0$  are interpreted as near-perfect agreement.

**Overall agreement** Agreement with the manually adjudicated data is generally very low, with most scores falling below zero. This is not unexpected, as we only adjudicated the most difficult cases, which naturally show higher disagreement. The highest individual agreement scores with the adjudicated data set were observed for annotation of the overall driveline quality and for dimension 1 (distance from reality), both reaching values above  $0.6$ , which indicates substantial agreement.

Figure 2 visualises the distribution of IAA scores across all dimensions and for all subsets of annotators. Dimensions 1 (distance from reality) and overall driveline quality are clearly the most straightforward categories as measured by

the median of IAA scores – showing moderate to substantial median agreement levels and low variability. We observe moderate agreement for dimension 3 (claim to absoluteness and handling of sources). Moderate agreement, albeit with substantial variability, is also observed for dimension 6 (emotionality), dimension 2 (linguistic and argumentative peculiarities), and dimension 5 (oversimplification). In contrast, dimension 4 (suggestiveness) exhibits only fair agreement.

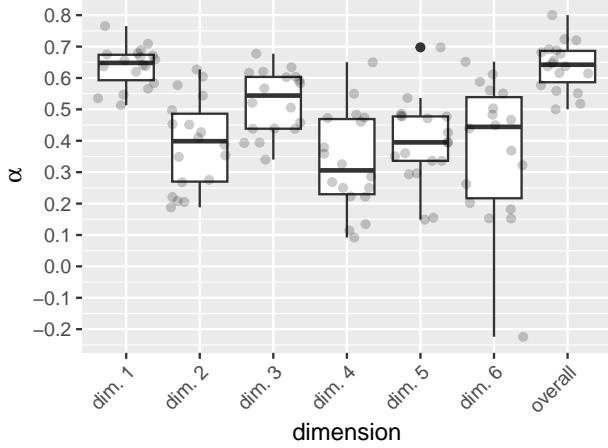


Figure 2: Distribution of agreement scores (Krippendorff's  $\alpha$ ) for each dimension.

**Pairwise agreement** Figure 3 presents average pairwise agreement scores for all dimensions. Overall, most annotator pairs demonstrate moderate agreement with one another, with a few notable exceptions. Annotators 2 and 4 exhibit substantial agreement on average, while annotators 3 and 5 as well as annotators 4 and 6 only achieve fair agreement. Agreement scores for the worst dimension (suggestiveness) range from .1 to .65, and from .51 to .77 for the best dimension (distance from reality). Interestingly, the agreement between annotators 2 and 4 remains exceptionally high – even substantial agreement for the most difficult dimension of suggestiveness – whereas annotators 3 and 5 drop down to slight agreement in this case.

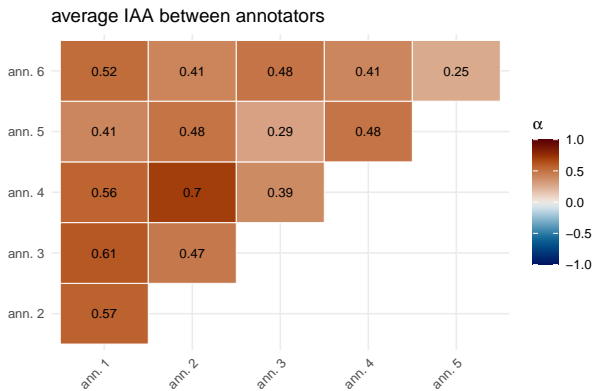


Figure 3: Average pairwise agreement scores across dimensions.

## 4. Prediction

How well can the dimensions predict the overall drive scores of the annotated posts? To get a first impression using our annotated sample, we took the adjudicated values, added mean values of all annotators for non-adjudicated dimensions, and used the resulting dataset for simple multiple linear regression analyses.

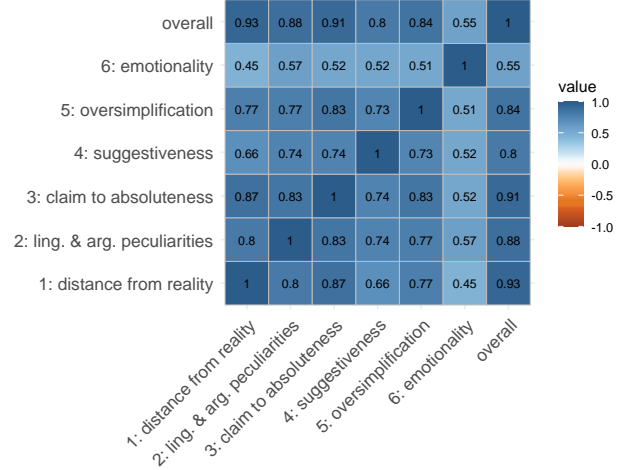


Figure 4: Pearson correlation coefficients between dimensions and overall scores.

Figure 4 shows the correlation coefficients between the individual variables, while Table 1 shows results from two different regression models. As can be clearly seen from the full model, *distance from reality* appears to be the single most important predictor (whereas *emotionality* does not contribute much). Even the second model, which includes only predictors that can be approximated by linguistic features alone (rather than, e.g., world knowledge), retains much explanatory power.

Table 1: Full regression model (left) and model with reduced number of predictors (right).

	Dependent variable:	
	‘overall’	
	model 1	model 2
‘1: distance from reality’	0.469*** (0.016)	
‘2: ling. & arg. peculiarities’	0.163*** (0.017)	0.406*** (0.023)
‘3: claim to absoluteness’	0.156*** (0.019)	0.608*** (0.020)
‘4: suggestiveness’	0.201*** (0.014)	
‘5: oversimplification’	0.087*** (0.015)	
‘6: emotionality’	0.042*** (0.012)	0.043** (0.018)
Constant	−0.174*** (0.028)	−0.151*** (0.039)
Observations	995	995
R <sup>2</sup>	0.945	0.879
Adjusted R <sup>2</sup>	0.945	0.879
Residual Std. Error	0.267 (df = 988)	0.396 (df = 991)
F Statistic	2,843.781*** (df = 6; 988)	2,398.010*** (df = 3; 991)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01



## 5. Conclusion

We presented an analysis of drivel found in German Telegram in terms of six distinct dimensions, based on a sample of 1,000 posts. Consistent manual annotation is a difficult endeavour, yet we can show moderate to substantial agreement between annotators. We will continue our effort to provide high-quality annotation for a larger sample. We hope to eventually provide a rich resource both for computational linguistics (e.g. automatic prediction of individual dimensions and overall drivel-like quality from text) and for corpus linguistics (potentially providing insight into how different dimensions manifest linguistically).

We also showed that the prediction of overall drivel-like quality from the six dimensions is straightforward. A simple linear regression model shows that posts are especially likely to be considered drivel when they appear distant from reality and when authors strongly assert their views. Future work will comprise further analysis of the associations between different dimensions.

A key limitation of the present study is the ambiguity and overlap among the six annotation dimensions. In particular, dimension 2 (linguistic and argumentative peculiarities) conflates several distinct elements – including spelling errors, coherence violations, and argumentative inconsistencies – into a single category. This lack of clear separation between formal, semantic, and logical features makes the development of precise annotation guidelines difficult, which in turn contributes to relatively low inter-annotator agreement and ultimately makes the dimension difficult to interpret. Future iterations of the study will aim to refine the dimensional structure to address these issues.

**Acknowledgements** This research has been partially funded by the German Research Foundation (DFG), project no. 466328567.

## 6. References

- Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., and Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, 40(S1):7.
- Heinrich, P., Blombach, A., Doan Dang, B. M., Zilio, L., Havenstein, L., Dykes, N., Evert, S., and Schäfer, F. (2024). Automatic identification of COVID-19-related conspiracy narratives in German telegram channels and chats. In Nicoletta Calzolari, et al., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1932–1943, Torino, Italia, May. ELRA and ICCL.
- Holnbürger, J., Goedeke Tort, M., and Lamberty, P. (2022). Q vadis? Zur Verbreitung von QAnon im deutschsprachigen Raum.
- Lamberty, P., Holnbürger, J., and Goedeke Tort, M. (2022). Das Protestpotential während der COVID-19-Pandemie.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Piskorski, J., Stefanovitch, N., Nikolaidis, N., Da San Martino, G., and Nakov, P. (2023). Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In Anna Rogers, et al., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.
- Piskorski, J., Mahmoud, T., Nikolaidis, N., Campos, R., Jorge, A., Dimitrov, D., Silvano, P., Yangarber, R., Sharma, S., Chakraborty, T., Guimarães, N., Sartori, E., Stefanovitch, N., Xie, Z., Nakov, P., and Da San Martino, G. (2025). SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation, SemEval 2025*, Vienna, Austria, July.

# A Case Study on Annotating and Analysing Situation Entity Types in Reddit Discussions on Democracy

Hanna Schmück\*, Michael Reder†, Katrin Paula‡, Annemarie Friedrich\*

\*University of Augsburg

{firstname.lastname}@uni-a.de

†Munich School of Philosophy

{firstname.lastname}@hfph.de

‡Technical University Munich

{firstname.lastname}@hfp.tum.de

## Abstract

Since anti-democratic movements increasingly use social media for political communication, studies examining democracy discourses in these spaces are critically needed. This paper introduces situation entity (SE) type (Smith, 2005) annotation as a promising framework for analysing political discourse in computer-mediated communication, focusing on rhetorical strategies used by writers with different political orientations. Our case study comprises 824 manually annotated situation segments (roughly clauses) from Reddit’s `r/PoliticalDebate` with six SE types: STATES, EVENTS, GENERIC SENTENCES, GENERALIZING SENTENCES, QUESTIONS, and IMPERATIVES. Our analysis reveals systematic differences across self-identified political orientations. The findings suggest SE type analysis effectively distinguishes argumentation patterns through specific versus generic content distinctions. Overall, the demonstrated framework offers promising applications for large-scale analysis of how members of different political movements construct their worldviews in digital environments. We emphasise that this case study merely attempts to propose a new method for analysing political discourse. Due to the small sample size, we cannot make any statements about political orientations and all of our analyses are intended to be exemplary.

**Keywords:** linguistic annotation, discourse mode, situation entity types, reddit discussion, democracy

## 1. Introduction

New social movements have emerged that are developing a self-image that is sometimes explicitly anti-democratic (Schedler, 2016), which in some cases implies an overthrow or at least a radical change of the political order with authoritarian tendencies. Direct communication within the movements and with the outside world in real time, made possible by digital transformation, follows the logic of social media algorithms and is often a central element in the self-image of anti-democratic movements (Karell et al., 2023). Their current growing importance poses an enormous challenge for democracies. However, the anti-democratic orientation of many populist or nationalist movements is not always easy to recognise in political rhetoric on social media platforms.

As an initial step in exploring how people argue about various perspectives and beliefs regarding democracy, we perform a case study on annotating and analysing *situation entity (SE) types* (Smith, 2005) as exemplified in Table 1 in Reddit discussions on democracy. SE types are a crucial component for distinguishing different *modes of discourse* (Smith, 2003) such as Narrative, Information, or Argumentative. Discourse modes differ in their distributions of *situation entity types* (Palmer and Friedrich, 2014). Framing information in one of these modes clearly has an impact on the reader’s perception, but SE types and discourse modes have to date not been studied in the context of computer-mediated communication.

The data for this study has been collected from Reddit via Communalytic (Gruzd and Mai, 2025), manually split into SE segments, i.e., roughly clauses, and annotated by

four expert and trained human annotators. Our findings show that STATES dominate overall discourse (52.4%), followed by GENERIC SENTENCES (22.2%). In our non-generalisable case study, Marxists stand out since they employ more EVENT-based reporting styles, Libertarians demonstrate more balanced distributions with higher QUESTION and IMPERATIVE usage, and Minarchists show a greater tendency to use GENERALIZING SENTENCES than the other groups.

## 2. Linguistic Background

SE types characterise the aspectual eventuality types of the situations invoked by the clauses of the text (Smith, 2003). In this case study, we follow the annotation scheme developed by Friedrich and Palmer (2014) and Friedrich et al. (2016). Besides the original types proposed by (2003) (including EVENTS, STATES, GENERIC SENTENCES, and GENERALIZING SENTENCES), the inventory was expanded by Palmer et al. (2007) to include the additional types QUESTION and IMPERATIVE to enable exhaustive text annotation.

Two key elements of a clause help determine its SE type: the *main verb* and the *main referent*. The main referent, loosely defined as the entity the segment is primarily about, is typically the subject in English. For instance, a GENERIC SENTENCE usually refers to general kinds or classes (e.g., “Rights only exist in three ways”). In the context of this annotation study, references to political parties (“AfD,” “Democrats”) and references to countries (“Germany”) were annotated as specific individuals.

By contrast, EVENTS, STATES and GENERALIZING SENTENCES focus on specific individuals (e.g., “The party I

SE Type	Examples
EVENT	<p><b>Minarchist</b> The NSDAP, won with 1rd of the vote in Germany back in 1933</p> <p><b>Libertarian</b> since obesity killed over 300,000 people in the US last year.</p> <p><b>Marxist</b> and the Democrats failed to turn out the same numbers in the places they needed.</p> <p><b>Marxist</b> Trump in 2016 and Biden both used it to do whatever</p>
STATE	<p><b>Libertarian</b> They would never do the same for us.</p> <p><b>Libertarian</b> My own ideology is leaving people alone</p> <p><b>Federalist</b> Kind of like how Germany has banned the Nazi party, and holocaust denial.</p> <p><b>Federalist</b> That should be an illegal position to have.</p> <p><b>Conservative</b> But it is a significant move against AfD.</p>
GENERALIZING SENT.	<p><b>Socialist</b> I've always felt [...]</p> <p><b>NONE</b> I also don't take example of bad behavior</p> <p><b>Minarchist</b> Also, Israel is fighting a defensive war against a terrorist organization</p> <p><b>Minarchist</b> that uses its own people as meat shields,</p> <p><b>Minarchist</b> and violates the laws of war.</p>
GENERIC SENT.	<p><b>Libertarian</b> Everyone seems to have a different idea of what democracy is.</p> <p><b>Libertarian</b> Children will always be a problem in this context</p> <p><b>NONE</b> Rights don't f*** exist outside of plots of land</p> <p><b>Federalist</b> A gay child has no choice in the community they wish to live in.</p> <p><b>Conservative</b> Democracy is quite paradoxical.</p>
QUEST.	<p><b>Libertarian</b> Who decides what kind of democracy we have?</p> <p><b>Marxist</b> Don't you want the people to be able to keep their leader. . . ?</p> <p><b>Conservative</b> How do you feel about Germany labeling AfD as Extremist?</p>
IMP.	<p><b>Voluntarist</b> Define human rights and how they would be enshrined.</p> <p><b>Libertarian</b> Just don't use my money for that!</p> <p><b>Federalist</b> ACT LIKE IT.</p>

Table 1: Examples of situation entity type annotation in Reddit discussions on democracy

voted for”). The main verb is the highest-ranked non-auxiliary verb in the dependency parse, e.g., “be” in “We shouldn’t be afraid.” STATES and EVENTS are distinguished by the lexical aspectual class of their main verbs (Siegel and McKeown, 2001): dynamic verbs indicate EVENTS (e.g., “reply”), while stative verbs signal STATES (e.g., “I own land”). Aspectual class is a property of the verb’s word sense. Moreover, habituality is a clause-level feature that also informs SE type classification. For example, EVENTS are episodic (“another libertarian replied”), whereas GENERALIZING SENTENCES are habitual (“I am always suspicious”). The annotation scheme also features the explicit annotation of the lexical aspectual class and the habituality of the main verb, and the genericity of the main referent. Operators like the perfect tense, negation, or modal verbs coerce EVENTS to STATES (this is not true for GENERIC SENTENCES and GENERALIZING SENTENCES).

### 3. Method

In this section, we explain the data collection and annotation process of our case study.

**Data collection and preprocessing.** The data was collected from Reddit using Communitary (Gruzd and Mai, 2025) which made it possible to download a batch of 2022 user entries from [r/PoliticalDebate](#) created between September 2024 and July 2025. These consist of two batches of top 50 most recent submissions containing the term ‘democracy’ that were filtered by the criterion ‘Hot’ via Reddit’s API client - one collected in May 2025 and one in July 2025 - as well as the associated comments and replies. A subsample

of 824 situation segments was used for SE type classification. A situation segment is the foundational unit of SE annotation and contains a coherent span of text that describes a single, unified situational context or event; situation segments often coincide with clauses. The benefit of using an online space such as [r/PoliticalDebate](#) is that it, in contrast to other CMC spaces, contains self-labels, so called *user flairs* which contributors use to self-ascribe a political label. As part of the preprocessing, we normalised user flairs such as [Minarchism - The Texan Minarchist \(Texanism\)](#) to [Minarchist](#) for all annotated examples.

**Annotation.** The entire sample for our case study has been annotated by two of the authors with experience in SE type annotation as well as two additional trained annotators who are undergraduate students of linguistics. We did not measure inter-annotator agreement (IAA) on the Reddit data, but Cohen’s  $\kappa$  scores for SE type annotation typically range around 0.66-0.69, with higher agreement ( $>0.9$ ) for IMPERATIVE and QUESTION, and somewhat lower agreement for identifying GENERALIZING SENTENCE (0.43) as reported by Friedrich et al. (2016). Becker et al. (2016) find  $\kappa$  to be around 0.52 when annotating argumentative microtexts, yet with a slightly larger set of SE types, including the types FACT, PROPOSITION, and RESEMBLANCE. They reflect embedded information (“I think SOME would, probably not all.”) and are generally hard to identify. The underlined PROPOSITION additionally receives the label STATE, so in this work, we focus on the more easily distinguishable basic set of SE types.



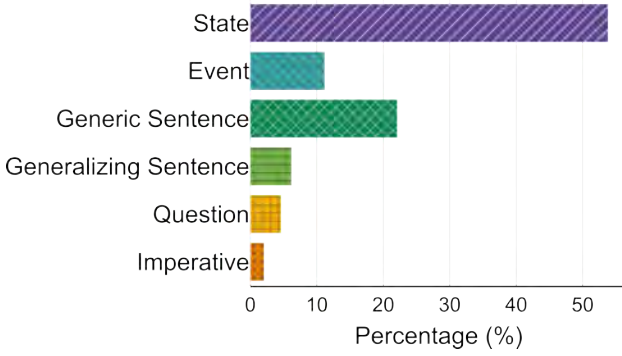


Figure 1: Overall distribution of SE types in sample of Reddit discussions on democracy (824 situation segments).

#### 4. Analysis

The overall distribution of SE types (see Figure 1) in the full annotated dataset reflects the predominantly argumentative nature of the discussions. It shows that STATES are most common, comprising half of all instances, followed by GENERIC SENTENCES at around 22%. EVENTS, GENERALIZING SENTENCES, QUESTIONS, and IMPERATIVES are significantly less frequent at 11%, 7%, 5%, and 2% respectively. These general findings match those of Becker et al. (2016), who also found a high percentage of generics in argumentative text.

Our case study further demonstrates that at the level of SE types, for the purpose of illustrating the method, interesting differences can be found in the texts written by contributors that self-assign to different political opinions. The five most frequent political user orientations present in our sample SE segments are Conservatives, Federalists, Libertarians, Marxists, and Minarchists. Figure 2 provides the SE type distributions by self-assigned political orientation. The SE type distributions in the texts written by Conservative and Federalist users follow the overall distribution in the dataset, with these contributors using predominantly STATES and GENERIC SENTENCES, which indicates that they generally use stative descriptions of their world view.

The data from the remaining political flavors follow notably distinguishable distributions. The Marxists contributing to our dataset use a distinct more reporting-like style which still predominantly relies on STATES but EVENTS notably comprise about 27% of their SE types - over twice the mean EVENT use in the overall dataset. They predominantly use EVENTS to back their arguments with specific examples, especially regarding statistics of past elections (see Table 1).

The widest variance in their use of SE types is exhibited by the Libertarians contributing to the Reddit excerpt. They also pose more QUESTIONS and utter more IMPERATIVES compared to the other political flavors. As illustrated by the examples in Table 1, their argumentation strategy seems to be more into the direction of influencing their readers by making them re-think their own positions.

The distribution of SE types for Minarchists shows that they use roughly twice the average percentage of GENERALIZING SENTENCE compared to the other flairs. As shown in Table 1, they contribute several sentences reporting on patterns of individual agents such as the state of Israel.

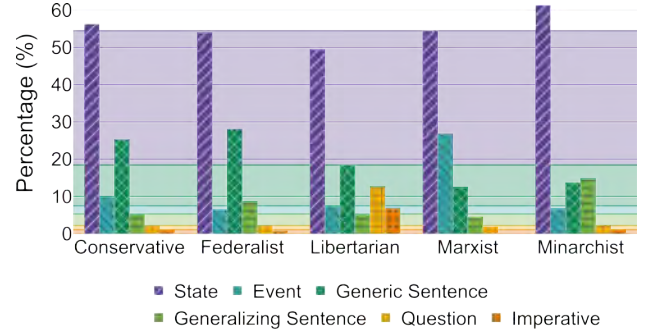


Figure 2: SE type distributions by political self-assigned user orientation (computed from 726 situation segments for top-5 user-assigned political orientation “flairs”. The background shows an area plot displaying the average values for the respective SE type amongst these five groups.

Flair	# of Users	# of SEs
Conservative	12	192
Federalist	6	138
Libertarian	10	183
Marxist	11	119
Minarchist	10	94
<b>Total</b>	<b>49</b>	<b>726</b>

Table 2: Distribution of users and SEs represented in the dataset by self-assigned political orientation “flair”.

#### 5. Discussion

It is important to acknowledge that our case study only draws from a limited sample of 49 unique users across five political orientations (see Table 2), which may not adequately represent the broader population or capture the full spectrum of political perspectives. Large-scale studies are necessary to study this development on a more comprehensive scale, as well as across time and following the development of individual users. Nevertheless, we argue that we have demonstrated that the methodology of analysing argumentative text in the computer-mediated communication domain can benefit from the linguistically motivated analyses of SE types.

Our case study illustrates the value of aspectual linguistic analysis for understanding the political discourse on democracy in computer-mediated communication. At the interface of computational linguistics, linguistics, and sociology, our proposed method facilitates the comparative examination of argumentation patterns of differently oriented social movements in large corpora. In particular by distinguishing specific from generic content, the digital linguistic analysis is closely linked to philosophical questions.

In future work, we will scale our method by enabling larger-scale text annotation supported by computational methods. A particular focus of the analysis is on generalising and generic statements (Friedrich and Pinkal, 2015; Friedrich et al., 2015) such as “The attack on free speech is, in fact, a problem in almost all EU countries [...]”<sup>1</sup> and their function in the performative constitution of the political self-

<sup>1</sup>Telegram channel “Freie Sachsen”, April 22, 2025

image of social movements. By studying both official documents and websites of the social movements as well as their publicly accessible chat channels, both the official self-image of the movements and the communication of the members themselves can be analysed. This allows for the investigation of different levels of the movements and different digital forms of communication. Our case study has demonstrated that the linguistic level of SE types, despite being motivated purely by linguistic aspectual distinctions, can provide valuable insights into argumentation structure.

## 6. Related Work

Similarly to our work, working towards the long-term goal of understanding what makes a message persuasive, Wei et al. (2016) study discussions on Reddits. They take a different approach, though, by training a supervised classifier and analysing the importance of linguistically motivated features. On the same data, Hidey et al. (2017) conduct an annotation study on argumentative text, though with more content-focused categories. They mark premises with Aristotle’s three types of persuasive modes: *ethos* (appealing to credibility), *logos* (appealing to reason), *pathos* (appealing to emotions), while claims are labeled as *interpretation*, *evaluation*, *agreement*, or *disagreement*.

Becker et al. (2016) annotate the argumentative microtext corpus (Peldszus and Stede, 2015), 112 German texts comprising a total of 668 situation segments, with SE types following the annotation scheme of Mavridou et al. (2015). They identify tendencies in the correlations between argument components (such as premises and conclusions) and SE types, as well as between argumentative functions (such as support and rebuttal) and SE types.

We are also aware of work studying the aspectual forms of clauses, in particular genericity, in other genres, e.g., in literary text (Dönicke et al., 2021), encyclopedic text (Friedrich et al., 2015; Friedrich and Pinkal, 2015) (Govindarajan et al., 2019), or English web text (Govindarajan et al., 2019).

## 7. Conclusion

This case study proposes situation entity (SE) type annotation as a novel framework for analysing political discourse in computer-mediated communication. Our analysis of 824 situation segments extracted from [r/PoliticalDebate](#) posts mentioning “democracy” shows that SE type distributions vary systematically across different self-identified political orientations, revealing distinct argumentation patterns.

Our study represents a first step towards identifying differences in argumentation patterns about democracy across political groups, which is essential for understanding how these discussions function and where potential threats to democratic discourse may emerge. The framework’s ability to distinguish specific from generic content proves valuable for examining how political movements construct ideological worldviews online. Generic and generalising statements appear particularly significant in performative political identity construction within CMC contexts.

## 8. References

- Becker, M., Palmer, A., and Frank, A. (2016). Argumentative texts and clause types. In C. Reed (Ed.), *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, Berlin, Germany: Association for Computational Linguistics. pp. 21--30.
- Dönicke, T., Gödeke, L., and Varachkina, H. (2021). Annotating quantified phenomena in complex sentence structures using the example of generalising statements in literary texts. In H. Bunt (Ed.), *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Groningen, The Netherlands (online): Association for Computational Linguistics. pp. 20--32.
- Friedrich, A. and Palmer, A. (2014). Situation entity annotation. In L. Levin et al. (Eds.), *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, Dublin, Ireland: Association for Computational Linguistics and Dublin City University. pp. 149--158.
- Friedrich, A. and Pinkal, M. (2015). Discourse-sensitive automatic identification of generic expressions. In C. Zong et al. (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics. pp. 1272--1281.
- Friedrich, A., Palmer, A., Peate Sørensen, M., and Pinkal, M. (2015). Annotating genericity: a survey, a scheme, and a corpus. In A. Meyers, et al. (Eds.), *Proceedings of the 9th Linguistic Annotation Workshop*, Denver, Colorado, USA: Association for Computational Linguistics. pp. 21--30.
- Friedrich, A., Palmer, A., and Pinkal, M. (2016). Situation entity types: automatic classification of clause-level aspect. In K. Erk et al. (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics. pp. 1757--1768.
- Govindarajan, V., Durme, B. V., and White, A. S. (2019). Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7, pp. 501--517.
- Gruzd, A. and Mai, P. (2025). Communalytic: A no-code computational social science research tool for studying online communities and public discourse on social media, 2025. Available at <https://Communalytic.org>.
- Hidey, C., Musi, E., Hwang, A., Muresan, S., and McKelown, K. (2017). Analyzing the semantic types of claims and premises in an online persuasive forum. In I. Habernal, et al. (Eds.), *Proceedings of the 4th Workshop on Argument Mining*, Copenhagen, Denmark: Association for Computational Linguistics. pp. 11--21.
- Karell, D., Linke, A., Holland, E., and Hendrickson, E. (2023). “Born for a Storm”: Hard-Right Social Media and Civil Unrest. *American Sociological Review*, 88(2), pp. 322--349.
- Mavridou, K.-I., Friedrich, A., Sorensen, M. P., Palmer, A., and Pinkal, M. (2015). Linking discourse modes and

- situation entity types in a cross-linguistic corpus study. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*. p. 12.
- Palmer, A. and Friedrich, A. (2014). Genre distinctions and discourse modes: Text types differ in their situation type distributions. In *Proceedings of the Symposium on Frontiers and Connections between Argumentation Mining and Natural Language Processing*,
- Peldszus, A. and Stede, M. (2015). An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2. pp. 801--815.
- Schedler, J. (2016). Die extreme Rechte als soziale Bewegung. In F. Virchow, et al. (Eds.), *Handbuch Rechtsextremismus*, Wiesbaden: Springer VS. pp. 285--326.
- Siegel, E. V. and McKeown, K. R. (2001). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4), pp. 595--628.
- Smith, C. S. (2003). *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press, 2003.
- Smith, C. S. (2005). Aspectual entities and tense in discourse. In *Aspectual inquiries*. Springer, 2005. pp. 223--237.
- Wei, Z., Liu, Y., and Li, Y. (2016). Is this post persuasive? ranking argumentative comments in online forum. In K. Erk et al. (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany: Association for Computational Linguistics. pp. 195--200.

# Annotating and Extracting Suggestive Language in English CMC: A Linguistically-Grounded Corpus and NLP Approach

Omnia Zayed, Sampritha Manjunath, Paul Buitelaar

Insight Research Ireland Centre for Data Analytics

Data Science Institute

University of Galway

IDA Business Park, Lower Dangan, Galway, Ireland

E-mail: firstname.lastname@universityofgalway.ie

## Abstract

In online discourse, suggestive language shapes interpersonal dynamics, collaborative problem-solving, and opinion formation. We present a new annotated corpus of English-language suggestive speech acts drawn from diverse computer-mediated communication (CMC) platforms. Grounded in linguistic theory, our annotation framework identifies concise textual spans where suggestions occur, capturing explicit and implicit forms. We demonstrate the utility of this corpus through experiments with state-of-the-art neural models for span extraction, accompanied by a qualitative discourse analysis that reflects on the pragmatic variability of suggestions across contexts. Our work contributes a linguistically-informed annotation methodology and empirical insights into modelling suggestive language in CMC data.

**Keywords:** suggestion extraction, linguistically-driven annotation, natural language processing

## 1. Introduction

Suggestions are an important speech act in online communication, shaping how people express expectations and propose actions. In computer-mediated communication (CMC), such as forums, social media, and surveys, suggestions surface in diverse forms, from explicit directives to subtle prompts (Negi, 2019). Automatically identifying and extracting these suggestive spans can support applications in public opinion analysis, civic engagement, and policy development.

While previous work in suggestion mining has mainly focused on sentence-level classification, less attention has been given to extracting the exact span of text where the suggestion occurs. This subtask is crucial for downstream applications such as suggestive language clustering, categorisation, and summarisation. For example, in the sentence "... if you do meet indoors, you should clean your hands regularly", the core suggestion is "*clean your hands regularly*". Extracted spans can be semantically clustered to reveal recurring patterns. Each cluster can be represented by a paraphrased category label that distils the shared intent of the suggestions to capture core intent and reduce redundancy. This makes accurate span extraction a foundational step in the structured analysis of suggestive language.

From a linguistic point of view, a suggestion is a directive speech act representing an interaction between the speaker and the hearer for the hearer to commit a future action that could benefit the hearer or the speaker or both in case of public suggestions (Searle, 1969; Martinez Flor, 2005). It is challenging to computationally distinguish suggestions from other types of directive speech acts, such as advice, commands, requests, and recommendations. Unlike other directive speech acts, which may carry a stronger sense of obligation, suggestions are typically softer and more open-ended, leaving the receiver free to accept or reject the proposed action. Explicit suggestions clearly state the speaker's intention using direct language to guide the hearer

towards a specific action or outcome. It can be recognised using particular keywords or expressions, such as imperatives, modal verbs, conditional statements, and politeness markers. On the other hand, implicit suggestions are indirect and subtle, relying on the context and the shared understanding between the speaker and the hearer rather than the choice of words. These differences should be clearly understood to develop a computational model that can identify and extract suggestions. Capturing these nuances is particularly challenging in CMC, where informal language, implicit cues, and varied genres prevail.

This paper presents a linguistically grounded annotation framework for suggestion extraction and applies it to create a gold-standard corpus of 700 CMC English texts from multiple domains. The corpus is used to evaluate transformer-based and large language models under both fine-tuned and zero-shot settings, compared to a pattern-based baseline. Our approach combines corpus creation, theoretical linguistic insight, and computational modelling to advance the study of suggestions in CMC.

## 2. Dataset Construction

### 2.1. Data Collection

To develop a gold-standard corpus for suggestive span extraction in English, we curated data from diverse CMC sources where suggestions naturally occur. Rather than restricting the dataset to a single domain, we selected three representative types of user-generated content: review forums, social media, and community surveys. Where possible, we leveraged existing suggestion classification datasets or applied a pre-trained classifier to filter candidate texts. We then manually annotated suggestive spans following our linguistically driven guidelines. Our gold-standard dataset draws from three types of CMC content to ensure domain diversity and relevance, as follows:

**Software Reviews:** We sample 300 suggestion-class instances from the SemEval 2019 dataset of developer feed-

back on UserVoice<sup>1</sup>, originally annotated for sentence-level suggestion classification (Negi et al., 2019).

**COVID-19 Tweets:** From a large public Twitter corpus on COVID-19 (Lamsal, 2021)<sup>2</sup>, we extract a random sample of 300 suggestions written in English using a deep learning model fine-tuned for suggestion classification (Zhou et al., 2019).

**Community Survey Responses:** We use 100 suggestions from open-ended responses to the Austin Community Survey (2015-2019)<sup>3</sup>, randomly sampled from over 2,900 answers to the open-ended question "*Q25 - If there was one thing you could share with the Mayor regarding the City of Austin (any comment, suggestion, etc.), what would it be?*" using a suggestion classifier. The question explicitly invites suggestions for city improvement.

To comply with ethical and privacy standards, all usernames were anonymised using a placeholder format (e.g., USER). The resulting dataset provides a cross-domain resource for studying suggestive speech acts in CMC.

## 2.2. Challenges

The process of annotating and extracting suggestive spans poses several significant challenges.

**Task Definition.** One primary issue is the inherent **subjectivity** in defining and identifying suggestive content, as interpretations can vary widely among annotators based on cultural, social, or personal biases. This subjectivity complicates the creation of a consistent and reliable annotation schema. The annotation guidelines should be detailed enough to cover various cases yet general enough to apply to diverse data. Another challenge lies in managing the **ambiguity** of natural language, where subtle linguistic nuances, such as sarcasm, idioms, or indirect expressions, may obscure the intended meaning. This can be exaggerated by the subtle differences between suggestions and other directive speech acts, e.g., commands, requests, and advice. Additionally, the **varied forms of suggesting**, in terms of the mode of delivery (direct/indirect) and the clarity of intent (implicit/explicit), introduce additional challenges to annotating and extracting suggestive spans. Direct suggestions are typically explicit and easier to identify, often containing clear linguistic markers such as imperative softening, hedging or modal verbs (e.g., "*You should try this*"). In contrast, indirect suggestions rely on subtler cues, such as implications, rhetorical questions, or contextual inferences (e.g., "*It might be worth considering*"). This variation complicates the development of annotation guidelines, as it requires annotators to distinguish between explicit and implicit suggestion strategies, which may not always be linguistically transparent.

**Data Validation and Quality.** Ensuring the reliability of annotated data requires rigorous **validation** processes, such

as measuring inter-annotator agreement (IAA) and verifying consistency across the dataset. However, the subjectivity of identifying suggestive spans, especially for indirect or implicit suggestions and other directive speech acts such as requests and commands, can lead to disagreement among annotators and hence, lower scores. On the other hand, establishing robust validation metrics that account for nuanced and context-dependent suggestions is required yet understudied. The **quality** of the data itself poses a serious problem due to the variability in text sources, e.g., some samples might lack sufficient context to interpret suggestions accurately.

**Re-annotation.** The aforementioned challenges might call for iterative re-annotation to refine the dataset and to learn from the disagreements, especially when new patterns of suggestions are discovered. Re-annotating large datasets is resource-intensive and time-consuming, requiring careful management to ensure that revisions enhance consistency without introducing additional biases.

These challenges highlight the need for clear annotation guidelines, annotator training, and iterative processes to ensure high-quality, validated data for suggestion extraction.

## 2.3. Data Annotation

### 2.3.1. Linguistically-Driven Guidelines

We developed annotation guidelines through an iterative process informed by linguistic strategies for the speech act of suggesting (Martinez Flor, 2005). The proposed guidelines build upon the previous recommendations on suggestion identification (Brun and Hagege, 2013; Negi and Buitelaar, 2017; Negi et al., 2019) by addressing their limitations and considering the challenges discussed in Section 2.2.. While earlier efforts focused on surface cues, including modal verbs (e.g., "*You should try this.*"), imperative softening (e.g., "*Why don't you consider this option?*"), hedging (e.g., "*Maybe you could look into this.*"), interrogative forms (e.g., "*Have you thought about doing it this way?*"), and politeness markers (e.g., "*perhaps, please, etc.*"), they often neglected contextual inference and failed to specify what should be excluded. To overcome this, we clearly defined both in-scope and out-of-scope cases to guide consistent annotation.

We introduced detailed criteria differentiating suggestions and similar statements, such as requests or commands. Although we kept using linguistic markers and patterns commonly associated with explicit suggestions, we incorporated context-sensitive rules for implicit suggestions. These rules consider the surrounding discourse to determine whether a statement qualifies as a suggestion, ensuring our framework adapts to varying contexts without losing accuracy. Indirect/implicit suggestions were explicitly defined as spans requiring contextual inference, supported by guiding examples and questions for annotators, e.g., "*Does the extracted span imply a course of action without directly stating it?*". To minimise annotator bias, we developed comprehensive examples and counterexamples for diverse linguistic phenomena, including sarcasm, rhetorical questions, and idiomatic expressions. Additionally, we included negative examples that should not be considered suggestions to further solidify what falls outside the scope,

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

<sup>2</sup>The data was collected using the Twitter Stream API before discontinuing its free access.

<sup>3</sup><https://data.austintexas.gov/City-Government/Community-Survey/s2py-ceb7> data accessed and downloaded on the 12th of Feb 2025.

reducing subjective interpretation. To achieve consistent annotations, we conducted multiple rounds of training and validation exercises, leveraging inter-annotator agreement (IAA) as an evaluation metric and discussing annotation disagreements to refine the guidelines iteratively.

### 2.3.2. Annotation Methodology

Since this is a highly subjective task, we relied on in-house expert annotators<sup>4</sup> instead of crowdsourcing. Four annotators were trained to annotate the compiled dataset, which comprises around 700 instances originally classified as containing suggestions from three sources, as detailed in Section 2.1.. The task defined what constitutes a suggestive span under the proposed framework. Comprehensive guidelines and training were provided to highlight the inclusion of direct/indirect and explicit/implicit suggestions and clarify edge cases. The annotators followed a multi-step process by 1) reading the entire text to establish context, 2) identifying spans explicitly or implicitly conveying suggestions, 3) taking note of indirect and implicit suggestive spans, and 4) documenting ambiguities or edge cases for review.

Annotation proceeded in multiple rounds to refine the guidelines. A pilot phase had annotators independently label a small data subset to identify ambiguities. Through review sessions, the annotations were compared, uncertainties were clarified, and interpretations were aligned. For challenging cases, cross-checking and collaborative reassessment were conducted. Disagreements were resolved through discussion to achieve consensus across all annotators.

## 2.4. Dataset Quality and Analysis

**Dataset Evaluation.** To assess the consistency and reliability of the annotations, IAA was carried out in terms of Fleiss’ kappa (Fleiss, 1971) between all annotators, achieving an average of 0.47. Based on the Landis and Koch (1977) scale, this is a moderate agreement despite the subjectivity of the task. While this may appear low in more objective annotation tasks, it is considered reasonable in similar inherently complex subjective linguistic tasks such as metaphor identification (Shutova et al., 2017), hate speech detection (de Gibert et al., 2018), offensive span identification (Ravikiran and Chakravarthi, 2022), and tropes detection (Flaccavento et al., 2025). It is worth noting that the IAA was 0.29 before providing the annotators with comprehensive guidelines and training exercises. In our analysis, rather than treating disagreement as an annotation error, we consider it an opportunity to explore the variability in human interpretation, improve annotation protocols, and better understand the subjectivity of the task. Table 1 shows the IAA for each subset. The results suggest that the software reviews and the Austin Survey subsets are more challenging to annotate, as will be discussed in the next section.

**Points of (Dis-)agreement.** A crucial part of this work is devoted to analysing annotation disagreements. This allowed for a better understanding of the subjectivity of the annotation task and refining the annotation guidelines. This

Domain	IAA	Agreement Strength
Software Reviews	0.41	moderate
COVID Tweets	0.52	moderate
Community Survey	0.48	moderate
average	0.47	moderate

Table 1: Breakdown of the Inter-annotator agreement among four annotators for each subset.

qualitative analysis is conducted through the ongoing analysis and discussion of the disagreements among the annotators after annotating each subset. The majority of disagreements centred around edge cases, such as ambiguous instances that lack sufficient context, sarcastic, rhetorical and figurative cases, and implicit suggestions. The annotators find the Software Reviews subset to be more challenging to annotate for suggestive spans due to the nature of the language and context typically used in such sources. One of the confusions came from the blurred boundaries between suggestions and other speech acts such as commands and requests. Suggestions typically imply optional proposition (e.g., *"It would be nice to add a dark mode"*), while commands are more direct (e.g., *"Add a dark mode now"*) and requests express a need or desire (e.g., *"Can you please add a dark mode?"*). Furthermore, the lack of context and use of domain-specific jargon can obscure the reviewer’s intent. For instance, technical statements such as *"Please add wiznote support as you support evernote."* may seem like requests, but could be intended as suggestions. The Austin Community Survey Responses subset emerged as the second most challenging dataset for annotation. The extended context of the responses often includes multiple suggestions spanning diverse topics, increasing annotation complexity. Similar to the Software Reviews subset, annotators frequently struggled to distinguish between suggestive expressions and other directive speech acts, which contributed significantly to inter-annotator disagreement (e.g., *"Fix this traffic mess!"*). Furthermore, the phrasing of the open-ended question, explicitly inviting *"any comment, suggestion, etc."*, introduced ambiguity that influenced annotators’ judgment (e.g., *"HOMELESS POPULATION NEEDS TO BE ADDRESSED."*). On the other hand, COVID Tweets contained more figurative language, such as sarcastic, rhetorical, and metaphoric suggestions. The annotators highlighted that some of the instances contain discussions with implicit suggestions, e.g., *"we need to think about all options on sites to get access for vaccines to large groups ..."*.

**Dataset Statistics and Examples.** The final gold-standard annotated dataset comprises 700 instances, out of which 68 are found to be not suggestive. The data sources varied between short and long contexts, resulting in an average word count per instance ranging from 31 to 67 words, and an overall average of 58.5 words per suggestive span. Notably, the minimum word count for a suggestive span is just 2 words, e.g., *"repair roads"*. Table 2 lists examples from each subset and the gold annotated suggestion span.

<sup>4</sup>NLP researchers of near-native English proficiency.



Subset	Instance with the Annotated Suggestive Span
SW	It would be nice it <b>there was an API Discovery endpoint.</b>
	Of course it would be even better to <b>make xaml use TypeConverter</b> additionally for compatibilitys sake with WPF.
Cvd	rt USER elon musk is not a genius he s an attention seeking man child and you should <b>pay as much attention to him on issues of...</b>
	rt USER covid is real please <b>stay safe wear your nose mask and sanitiser your hands regularly</b>
Aust	I would suggest <b>keeping up with the population growth and help reduce traffic.</b>
	Either <b>develop great initiatives to include and engage people of color</b> , or if the programs exist learn how to <b>promote them to the people they are meant to reach.</b>

Table 2: Examples of instances in each subset of the newly created gold-standard dataset of suggestive spans. The annotated gold suggestive span is highlighted in bold red. Domains are SW: Software, Cvd: COVID, and Aust: Austin Community Survey.

### 3. Modelling Suggestion Extraction

To validate our gold-standard dataset and explore effective modelling approaches for suggestive span extraction, we formulated the task as identifying the precise span(s) within a suggestion-classified text that convey suggestive language. We benchmarked three modelling strategies: 1) a linguistically grounded pattern-based baseline using modal verbs, imperatives, and conditional phrases to capture explicit suggestions which are based on previous work (Martinez Flor, 2005; Dong et al., 2013; Negi, 2019); 2) fine-tuned transformer-based models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and SpanBERT (Joshi et al., 2020)) leveraging contextual embeddings for span-level prediction; and 3) zero-shot large language models (LLaMA (Touvron et al., 2023) and Mistral (Jiang et al., 2023)) prompted to extract suggestions without additional fine-tuning.

We evaluated the models using our gold-standard dataset of suggestive spans. As no prior benchmark exists for this task, our pattern-based method serves as a baseline. We compared the performance of fine-tuned transformer models and zero-shot LLMs using word-level precision, recall, and F1-score (Rijsbergen, 1979). To assess semantic similarity beyond exact matches, we also employed BERTScore (Zhang et al., 2020), which captures contextual alignment between predicted and gold spans. Together, these metrics offer a robust evaluation of each model’s ability to extract meaningful and accurate suggestions.

As shown in Table 3, SpanBERT outperformed all other models, benefiting from its span-oriented design. While RoBERTa performed reasonably well, fine-tuned BERT lagged behind. Among the LLMs, Mistral showed moderate success, occasionally producing concise spans, whereas LLaMA struggled in the zero-shot setting, often failing to identify relevant suggestions. This is likely due to the absence of task-specific fine-tuning and the general-purpose nature of these models.

Our qualitative analysis further highlights these trends. BERT missed over 70% of gold spans, and RoBERTa missed around 31%. Mistral occasionally returned empty

outputs, ignoring prompt constraints, while LLaMA often rewrote or over-explained the spans. SpanBERT showed strong alignment, but sometimes truncated spans where contextual boundaries were complex.

	BERTScore					
	Prec.	Recall	F1	Prec.	Recall	F1
Pattern-based	0.09	0.09	0.08	-0.11	0.07	-0.02
BERT	0.21	0.24	0.22	0.32	0.09	0.2
RoBERTa	0.55	0.59	0.55	0.58	0.54	0.56
SpanBERT	<b>0.74</b>	<b>0.82</b>	<b>0.75</b>	<b>0.69</b>	<b>0.73</b>	<b>0.71</b>
Llama	0.09	0.05	0.05	0.05	0.01	0.03
Mistral	0.47	0.41	0.4	0.28	0.26	0.26

Table 3: The performance of the evaluated models against the pattern-based baseline on the gold-standard dataset.

## 4. Related Work

Early work on suggestion mining focused on sentence-level classification, particularly in product reviews and customer feedback (Ramanand et al., 2010; Viswanathan et al., 2011; Dong et al., 2013; Moghaddam, 2015). Brun and Hagege (2013) attempted to extract suggestions using linguistic patterns, assuming full sentences as suggestive, but their dataset remains unavailable. Negi and Buitelaar (2017) redefined suggestion identification through linguistic insights, leading to multiple annotated datasets and a SemEval Shared Task (Negi et al., 2019) on binary classification.

However, existing work largely focuses on suggestion classification at the sentence level. To address this gap, we broaden suggestion mining to include extraction, clustering, and paraphrasing. This work focuses on suggestive span extraction, a subtask related to span extraction (Papay et al., 2020) and sharing traits with extractive summarisation (Padmakumar and He, 2021; Jie et al., 2024), but with a unique focus on capturing subjective, actionable content.

## 5. Conclusion and Future Work

This paper addresses the underexplored task of suggestion extraction in English computer-mediated communication (CMC) by introducing a linguistically driven annotation approach and creating the first gold-standard dataset of suggestive spans. We benchmarked the task using fine-tuned transformer models and zero-shot LLMs, with SpanBERT achieving the best performance. A pattern-based baseline and qualitative analysis provided further insights into model behaviour.

Our dataset spans three domains covering various media of digital communication, including review forums, social media, and surveys, with plans to expand into other areas such as hospitality, politics, and education. Future work will explore few-shot prompting, weak supervision using SpanBERT, and additional models such as T5 and Gemma to enhance performance and generalisability.

## 6. Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland for the Postdoctoral Fellowship award

GOIPD/2023/1556 (Glór). This work has also received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 883285 (PANDEM-2). Additionally, this work has been partially funded by the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289\_P2 (Insight\_2).

## 7. References

- Brun, C. and Hagege, C. (2013). Suggestion mining: Detecting suggestions for improvement in users’ comments. *Research in Computer Science*, 70(79):pp. 171–181.
- de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In Darja Fišer, et al., editors, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *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 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Dong, L., Wei, F., Duan, Y., Liu, X., Zhou, M., and Xu, K. (2013). The automated acquisition of suggestions from tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):pp. 239–245, Jun.
- Flaccavento, A., Peskine, Y., Papotti, P., Torlone, R., and Troncy, R. (2025). Automated detection of tropes in short texts. In Owen Rambow, et al., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5936–5951, Abu Dhabi, UAE, January. Association for Computational Linguistics.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):pp. 378–382.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.
- Jie, R., Meng, X., Jiang, X., and Liu, Q. (2024). Unsupervised extractive summarization with learnable length control strategies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):pp. 18372–18380, Mar.
- Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Lamsal, R. (2021). Design and analysis of a large-scale covid-19 tweets dataset. *applied intelligence*, 51:pp. 2790–2804.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):pp. 159–174.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.
- Martinez Flor, A. (2005). A theoretical review of the speech act of suggesting: towards a taxonomy for its use in flt. *Revista Alicantina de Estudios Ingleses*, (18):pp. 167–187, 11.
- Moghaddam, S. (2015). Beyond sentiment analysis: Mining defects and improvements from customer feedback. In Allan Hanbury, et al., editors, *Advances in Information Retrieval*, pages 400–410, Cham. Springer International Publishing.
- Negi, S. and Buitelaar, P. (2017). Suggestion mining from opinionated text. *Sentiment Analysis in Social Networks*, pages pp. 129–139.
- Negi, S., Daudert, T., and Buitelaar, P. (2019). SemEval-2019 task 9: Suggestion mining from online reviews and forums. In Jonathan May, et al., editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 877–887, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- Negi, S. (2019). *Suggestion mining from text*. Ph.D. thesis, National University of Ireland Galway, Galway, Ireland.
- Padmakumar, V. and He, H. (2021). Unsupervised extractive summarization using pointwise mutual information. In Paola Merlo, et al., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online, April. Association for Computational Linguistics.
- Papay, S., Klinger, R., and Padó, S. (2020). Dissecting span identification tasks with performance prediction. In Bonnie Webber, et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4881–4895, Online, November. Association for Computational Linguistics.
- Ramanand, J., Bhavsar, K., and Pedaneekar, N. (2010). Wishful thinking - finding suggestions and ‘buy’ wishes from product reviews. In Diana Inkpen et al., editors, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 54–61, Los Angeles, CA, June. Association for Computational Linguistics.
- Ravikiran, M. and Chakravarthi, B. R. (2022). Zero-shot code-mixed offensive span identification through rationale extraction. In Bharathi Raja Chakravarthi, et al., editors, *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 240–247, Dublin, Ireland, May. Association for Computational Linguistics.
- Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, 2nd edition.
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London, Jan.
- Shutova, E., Sun, L., Darío Gutiérrez, E., Lichtenstein, P., and Narayanan, S. (2017). Multilingual metaphor processing: Experiments with semi-supervised and unsu-



- pervised learning. *Computational Linguistics*, 43(1):71–123, April.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.
- Viswanathan, A., Venkatesh, P., Vasudevan, B. G., Balakrishnan, R., and Shastri, L. (2011). Suggestion mining from customer reviews. In *Americas Conference on Information Systems*.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Zhou, Q., Zhang, Z., Wu, H., and Wang, L. (2019). ZQM at SemEval-2019 task9: A single layer CNN based on pre-trained model for suggestion mining. In Jonathan May, et al., editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1287–1291, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

# Beyond names: how to label gender automatically in CMC data?

Pasi Fränti, Juhani Järviö, Mehrdad Salimi, Irene Taipale, Mikko Laitinen, Rahel Albicker, Chunyan Nie, Masoud Fatemi, Paula Rautionaho

University of Eastern Finland

E-mail: mikko.laitinen@uef.fi

## Abstract

Large-scale data from social media offers numerous benefits for research, but one significant and widely known limitation is the lack of detailed social background information. This gap poses a serious challenge for fields such as sociolinguistics and the study of language variation and change, where demographic and contextual information are crucial. A commonly used approach in computer-mediated communication (CMC) research has been to infer gender from users' names. However, a variety of other methods have emerged in recent years, drawing from advances in machine learning. This presentation reviews the current state of social media data enrichment and introduces a generalizable method that integrates various types of background information. Enriched data can train machine learning models to label social media user accounts with more accurate gender information.

**Keywords:** demographic prediction, Twitter data, data labeling, social media

## 1. Introduction

Social media datasets are widely used in sociolinguistic research, yet they typically lack the most basic demographic variables, such as age, gender, or occupation.<sup>1</sup> As a result, socially conditioned variation remains largely inaccessible in studies that draw on large-scale social media corpora. We use datasets of 343,149 Twitter users from Australia, the UK, and the USA, with message histories from 2006 to 2023 (Laitinen et al., 2025).<sup>2</sup> So far, we have enriched the data with network information (Laitinen & Fatemi, 2024), and our current focus is on inferring gender information from user profiles. While this application does not provide a standardized format for users to express their gender, many profiles contain sufficient direct or indirect cues to make gender inference feasible. To illustrate, Figure 1 provides information pointing toward a male gender identity. These include a profile picture, a traditionally male first name in the Anglo-American culture, and a pronoun declaration (*he/him*).

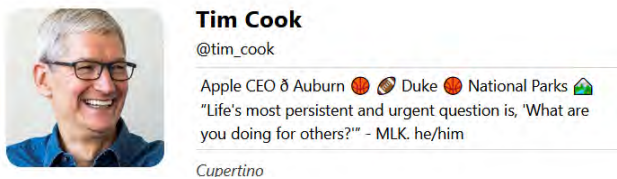


Figure 1: Example profile of a known public figure

This is not always the case with our dataset, which is also too large to annotate manually. If it took a human 10 seconds to infer the gender of a user, the full annotation task would require around 1,000 hours of work for a single person. Many user profiles also lack clear indicators, such as pictures or names, and may contain contradictory information, making them much more difficult to annotate.

For these reasons, we rely on computational methods to assist with labeling.

Previous studies have employed a variety of methods, ranging from using users' names (Coats, 2016, 2019) and profile pictures (Yildiz et al., 2017) to analyzing textual information in user profiles (Wang et al., 2019). A commonly used method in computer-mediated communication (CMC) studies is the use of names, as first-name databases are widely available across many languages (Lupo et al., 2024). For example, the UK Office for National Statistics (ONS) provides data on names given to newborns between 1996 and 2021, along with associated gender information. These databases are generally consistent, with only a small minority of names assigned to more than one gender. However, naming practices are highly heterogeneous, and a significant share of accounts may remain unlabeled. This article surveys the state of the art to support more accurate labeling.

We mostly focus on binary classification of gender for the results. This is a practical decision, as most methods that have high coverage only make binary predictions: either male or female. Three of the methods we consider also offer a non-binary category: pronoun declarations, keyword search, and crowdsourcing. However, none of them has enough coverage to be useful on a large scale. We acknowledge the limitations of binary gender classification, which does not capture the complexity and diversity of identities (Keyes et al., 2021; Simeoni et al., 2024). However, due to the scale of the dataset and the constraints of current computational methods, we settled on binary categorization at this stage.

<sup>1</sup> We wish to thank the two anonymous reviewers for their comments on an earlier version of this article.

<sup>2</sup> Twitter's name was changed to X recently. However, our

datasets were collected before this change. For consistency and accuracy, we use the original name to refer to our dataset.

## 2. Previous studies

Gender (Yildiz et al., 2017) and age (Sloan et al., 2015) prediction have received considerable attention in recent research, with Twitter serving as the primary data source (Lupo et al., 2024; Mislove et al., 2011; Sloan et al., 2015; Tonglet et al., 2024; Yildiz et al., 2017). Previous studies have also focused on predicting other demographic factors, such as location (Lupo et al., 2024; Sloan et al., 2013; Wang et al., 2019), occupation (Sloan et al., 2015), and social class (Sloan et al., 2015). However, some demographic attributes are more readily predictable than others. For example, it has been argued that age is easier to predict than occupation (Sloan et al., 2015) but more challenging than gender (Lupo et al., 2024; Wang et al., 2019).

A key challenge in using social media data is its lack of demographic representativeness, particularly concerning geography, gender, and race (Mislove et al., 2011; Wang et al., 2019). There is a known bias toward males in Twitter data compared to actual populations (Dixon, 2024; Gombert et al., 2025; Mislove et al., 2011; Yildiz et al., 2017). Twitter users are also, on average, significantly younger than the population at large (Sloan et al., 2015), which affects the ability to generalize social media results. Demographic prediction offers a potential means to mitigate this limitation. When demographic information is available, researchers can address sampling biases such as overrepresentations of young men.

In manual labeling, annotators typically rely on a combination of profile-based signals such as the username, display name, profile description, and profile picture. Message content has also been used to infer user demographics (Yildiz et al., 2017). The scale of annotation varies across studies. For instance, in Wang et al. (2019), the labeling was conducted by three individuals, while Yildiz et al. (2017) involved over 1,000 annotators. Manual labeling allows nuanced interpretation of multiple cues with high accuracy but is time-consuming and does not scale well.

Automatic labeling methods have been developed to address these challenges. Common approaches include image processing based on profile pictures or searching for gendered lists of names and keywords. Some studies have also applied regular expressions to users' profile descriptions to detect age (Lupo et al., 2024; Sloan et al., 2015). These methods are scalable. However, some of these have limited coverage; e.g., a face could be detected only in about 30% of the profile pictures (O'Connor et al., 2024).

The effectiveness of these approaches varies depending on the task and data quality. Crowdsourcing has been reported to work better for age detection than computer-based methods (Yildiz et al., 2017), whereas another more recent study reports that computer-based methods can already predict gender with high accuracy (Wang et al., 2019).

The most common approach to gender prediction is to use lists of first names (Sloan et al., 2013). In this approach, names are classified based on the probability of being

associated with one gender. For example, if 98 of 100 babies named *Lennox* are male and two are female, the probability of the name being male is 98%. Following Mislove et al. (2011), we use a 95% probability threshold for choosing names. However, accuracy varies between languages. For instance, Italian names are more gender-specific, allowing for more reliable prediction (Lupo et al., 2024). In general, this approach does not provide high coverage: Mislove et al. (2011) managed to predict the gender of only 64.2% of users.

Machine learning models based on text have also been used for gender prediction. Lupo et al. (2024) consider Bag of Words, TF-IDF, word embeddings, and topic modeling for feature extraction. Using the M3 inference tool, Gonzales (2024) utilizes grammatical features for age and gender prediction. Large language models (LLMs) and machine learning classifiers such as Random Forest and XGBoost have been applied with promising results (Tonglet et al., 2024; Wang et al., 2019). Tonglet et al. (2024) report 92% accuracy for gender detection with M3 for 82% of users in their dataset. While these methods can leverage different data like text and images, they have comparability issues due to a lack of standardized evaluation datasets (O'Connor et al., 2024).

A challenge in gender and age prediction is the lack of standardized reporting on data collection and labeling procedures (O'Connor et al., 2024). We aim to address this by providing a detailed account of our data collection and annotation process. Our computer software is publicly available as open source to support transparency and replication (see section 6 below).

## 3. Data

Our data consists of profile information and user-generated messages of 343,149 Twitter users, mostly from the USA, UK, and Australia. Data gathering is explained in more detail in Laitinen et al. (2025). The profiles were collected in 2023 using the now-closed Academic API, which limited the data collection to a maximum of 3,200 messages per user, spanning from 2006 to 2023. Network data was also collected based on ego networks, centered around 5,773 ego users. However, we focus here on individual profiles for predicting gender and aim at combining network data with gender/age data later.

The profile information includes username, display name, description, and location fields. All of these were used for the methods described in the following sections. The location field is not useful as an actual geographical datapoint, since it is a free-text field, and according to Mislove et al. (2011), only 9% of self-announced user locations can be matched to a real location by the Google Maps API. Because of the age of the data and the changes in the API, retrieving profile images at a massive scale is difficult, meaning that images were not used in this paper.

User language is available for each post, provided by the Twitter API, and is used to estimate each user's primary language. In our data, primary language means that at least

75% of the user’s tweets were in a single language. Table 1 shows that 82% of users write mainly in English.

Language	Users	
English	280,044	82 %
Multiple	46,915	14 %
Spanish	3,493	1 %
Japanese	2,190	1 %
No linguistic content	1,470	0 %
<b>Top 5 languages</b>	<b>334,112</b>	<b>97 %</b>
<b>All languages</b>	<b>343,146</b>	<b>100 %</b>

Table 1: Top 5 languages as the number of users whose tweets were 75% in the listed language.

Information on user location is limited as it is available only for a subset of tweets. Only about one-third of users have any country-level location data. Of those, 83% are from the three countries from which the original data was gathered: the US, UK, and Australia.

To fix the gaps in the profile data, we set up a crowdsourcing effort to manually label a small subset. This crowdsourcing used a combination of profile data and user-generated messages as the stimuli for our students, who manually labeled some 2,800 accounts. This data will be used as the gold standard to evaluate the accuracy of automated labelling.

#### 4. Parameters for computer-based labelling

For automatic labeling, we use a combination of three parameters:

- Name information
- Self-declaration of gender
- Keywords in user profiles

These methods were chosen for their simplicity and presumably high accuracy for automated labeling.

##### 4.1. Name-based classification

One of the most widely used methods is to tokenize usernames and match the output against a priori collected first name lists with gender information (Coats, 2016; Mislove et al., 2011; Sloan et al., 2013).

We acquired first name lists from four national government organizations: the UK, US, Australia, and Canada to increase accuracy (Attorney-General’s Department, Government of South Australia, 2013; Office for National Statistics, 2022; Statistics Canada, 2023; U.S. Social Security Administration, 2024). Their statistics are summarized in Table 2.

Region	Years	Names	Average M F confidence	
UK	1996–2021	36,043	99 %	99 %
USA	1923–2023	101,785	98 %	99 %
Australia	1944–2014	51,518	99 %	99 %
Canada	1991–2023	16,444	98 %	99 %
<b>Unique names</b>	<b>M:43,441</b>	<b>F:75,655</b>	<b>98 %</b>	<b>99 %</b>
			<b>M</b>	<b>F</b>

Table 2: Summary of the name datasets. Average confidence is the probability that a name is a specific gender, averaged across all names.

We also employ a list of last names from the US 2010 census data (U.S. Census Bureau, 2016) to filter out names that are predominantly surnames, even if they are occasionally used as first names. One such example is *Williams*, which occurs 1,625,252 times as a surname compared to just 5,295 times as a first name.

The datasets contain names, counts, and the gender information (male or female) for each name, excluding names with <10 instances. Names that can be given both to girls and boys add uncertainty to the method (e.g., *Riley*, with 51% male). We tokenize usernames and profile names using the NLTK natural language toolkit Python package, along with various manual fixes to parse more difficult names. We match each token to the combined name list and, in total, found a match for 275,143 users (80%), of which 152,612 (45%) were gendered based on first names with an average probability of one gender >95%.

Mislove et al. (2011), using the 1,000 most common male and female names from the Social Security Administration dataset, found a match for 64% of users, of which 72% had a male name. Sloan et al. (2013) used a database of 40,000 names from 54 countries globally. They found a first name for 48% of the users. Both Mislove and Sloan used only the first name item.

##### 4.2. Pronoun declarations

The second approach builds on an increasing tendency of users to self-declare their gender identity in the profile info (e.g., *she/her*). Recent studies show that this is an emerging way of identifying oneself (Jiang et al., 2023; Tucker & Jones, 2023).

We found this method to be both the most reliable and the easiest to detect. Self-declaration also extends to non-binary genders.

To implement it, we developed a list of regular expressions designed to match known English pronoun declarations. This list includes 200 variations, with the most frequent ones shown in Table 3. Remarkably, 94% of all pronoun declarations fall into the ten most common patterns. To identify these declarations, we tokenized the text by extracting words separated by standard delimiters and

matched them against our list. All matches were manually reviewed to eliminate false positives. Each declaration was then categorized into one of four groups: female (F), male (M), non-binary (NB), or uncategorized (U).

Declaration		Count	
She / her	F	6,178	45 %
He / him	M	3,519	26 %
They / them	NB	926	6 %
She / they	F	835	6 %
He / they	M	430	3 %
<b>Top 5</b>		<b>11,888</b>	<b>91 %</b>
<b>All</b>		<b>13,447</b>	<b>100 %</b>

Table 3: Top 5 most common pronoun declarations.  
(Taken from a total 343k user dataset.)

The limitation is low coverage, as only about 5% of users in our data include pronoun declarations. The method is also limited to English-language pronouns. The main advantage of pronoun declaration is its high accuracy.

Jiang et al. (2023) reported an increase of 33% in the number of tweets made by users with pronouns from 2020 to 2021 (2.86%  $\rightarrow$  3.82%). Tucker & Jones (2023) write that pronoun usage peaked by 2022 to around 4%–5%. Our results show the same for profiles fetched in 2023. Both studies also show a similar distribution of gender categories to ours, with female pronoun declarations being the majority. Table 4 shows the gender distribution in pronoun declarations.

Unique users	Count	
Female	7,445	55 %
Male	4,359	32 %
Non-binary	962	7 %
Unknown	681	5 %
<b>Total</b>	<b>13,447</b>	<b>100 %</b>

Table 4: Counts of users with pronoun declarations. The coverage remains relatively low: 13,447 / 343,149 = 4%.

### 4.3. Keyword search

The third approach is to find gender-related keywords in the profile description (Emmery et al., 2017; Jurgens et al., 2017). We experimentally compiled three keyword lists targeting gendered terms associated with female, male, and non-binary identities (e.g., *mother*, *husband*, and *enby*). These keywords were used in regular expression-based searches of users’ profile descriptions to identify gender-related self-descriptions. Unfortunately, like pronoun declarations, keyword search also has low coverage, with only 6% of profiles labelled as one of the three gender categories: F, M, NB.

## 5. Observations

The gender labelling task has two objectives:

1. *Coverage*: to label as many profiles as possible
2. *Accuracy*: to ensure the labels are accurate

We measure the success of the different methods by their coverage and accuracy. *Coverage* is the number of users that are assigned a label, while *accuracy* is the proportion of labeled users for whom the predicted gender matches the true gender. Both are reported as a proportion between 0 and 100%.

In computer science, their joint optimization is known as a two-objective optimization problem. Perfect results for both metrics are rarely achievable simultaneously: improving one typically compromises the other.

Reducing the problem to a single-objective optimization, a simple goal would be to label all users (100% coverage) and find a method to maximize accuracy. It would also be possible to achieve high accuracy by labelling only the users that we are certain of. This could give close to 100% accuracy at the cost of low coverage.

In practice, perfect accuracy is unattainable even for a subset of users. Automatic methods can be surprisingly good but are far from perfect. Even humans may fail, as user profiles may lack obvious gender-related clues such as a name and profile picture. A realistic goal is therefore to find a good compromise by giving more weight to the more important objective.

Next, we report the observed values for the two objectives when labeling a subset of our dataset using three automatic methods. We lack ground truth labels and therefore use the human-annotated genders as a gold standard.

The results are reported in Table 5. Our first observation is that names offer the widest coverage (44%) but have the lowest accuracy (72%). In contrast, pronoun declaration and keyword-based methods show higher accuracy (96% and 82%) but considerably lower coverage (3% and 6%).

Table 5 also includes results obtained with the M3 method (Wang et al., 2019), excluding predictions based on profile pictures. M3 uses sophisticated machine learning techniques with word embedding. It achieves higher coverage (63%), but lower accuracy (65%) compared to name-based predictions alone (72%). Its accuracy can be increased from 65% to 79% by raising its confidence threshold from 50% to 95%, although this reduces coverage to 28%. Including profile pictures would increase accuracy, but as explained above, we did not use them for this test due to the scale of the data.

We also note that the methods display gender-related biases, as the proportion of users classified as female is systematically higher with the automatic methods than in the manually annotated sample. For example, name-based and keyword-based predictions identify 43% and 46% users as female, respectively. Women seem to use pronoun declarations more often than men in their profiles, as 63% of the detected declarations are female. This result differs

considerably from the usual male-female ratio reported elsewhere. For example, Dixon (2024) reported 60% male, and our crowdsourcing tool shows 70% male.

Feature	Female labels	Coverage	Estimated accuracy
Names	43 %	152,612	72 %
Pronouns	63 %	11,729	96 %
Keywords	46 %	22,248	84 %
Manual	35 %	1,888	-
M3 (50 %)	38 %	217,563	65 %
M3 (95 %)	26 %	96,322	79 %

Table 5: Results table from several methods combined. Manual labelling was used to measure estimated accuracy.

M3 was split into two: one with a simple majority (the bigger probability is the label), and one with a 95% threshold, where any users with a lower probability were labeled as undetermined.

## 6. Conclusions and next steps

Incorporating social background information into large-scale CMC data may initially appear straightforward. However, several factors make this a complex and nuanced task. From a social sciences and humanities perspective, many social background variables, such as gender, age, and class, are not fixed traits but socially constructed categories. Individuals perform and express these identities in diverse ways. This complexity is especially evident on social media, where users curate and perform their identities through varied and often ambiguous signals.

From a computational perspective, the challenge becomes methodological: how can we ensure both high coverage and high accuracy in labeling? Prioritizing high accuracy reduces coverage, while maximizing coverage compromises precision. Effective research must strike a balance between these competing goals. Our approach addresses this by constructing a dataset that integrates multiple parameters to support robust demographic inference.

This article reviews the state-of-the-art in socio-demographic user labeling, drawing on recent advances not only in CMC research but also in machine learning and related fields. Building on interdisciplinary work, we outline a methodology for constructing a testing dataset that incorporates three gender-related parameters.

We also built a crowdsourcing platform, designed for flexible task configuration and adaptable for a wide range of annotation needs. This platform underpins the creation of our testing dataset, which will be our ground truth. It can be used to evaluate our automatic methods, and other ML-based methods, such as the M3 model (Wang et al., 2019).

Our immediate goal is to complete gender labeling, aiming to maximize coverage while maintaining acceptable

accuracy for all 343,149 users in the dataset. In the future, we plan to extend our work to include additional demographic variables, such as age groups (e.g., <29 years, 30–50 years, 51+). Notably, our dataset already includes network parameters that capture the degree of user connectivity. When combined with gender and age data, this enables large-scale sociolinguistic analysis.

The resulting data will be made public for reuse and further research once we have annotated all our data. The current code for the presented automatic labeling is available at <https://cs.uef.fi/comet/code/>.

## 7. References

- Attorney-General’s Department, Government of South Australia. (2013). *Most popular Baby Names (1944–2013)* (Version 2016) [Statistical dataset; CSV]. data.sa.gov.au.  
<https://data.sa.gov.au/data/dataset/popular-baby-names>
- Coats, S. (2016). Grammatical frequencies and gender in Nordic Twitter Englishes. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. University of Ljubljana Academic Publishing, pp. 12–16 [https://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016\\_Coats\\_Grammatical-Frequencies-and-Gender.pdf](https://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Coats_Grammatical-Frequencies-and-Gender.pdf)
- Coats, S. (2019). Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics*, 42(1). <https://doi.org/10.1017/S0332586519000039>
- Dixon, S. J. (2024, May 22). *Distribution of X (formerly Twitter) users worldwide as of January 2024, by gender*. Statista.  
<https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/>
- Emmery, C., Chrupała, G., & Daelemans, W. (2017). Simple Queries as Distant Labels for Predicting Gender on Twitter. *Proceedings of the 3rd Workshop on Noisy User-Generated Text*, 50–55.  
<https://doi.org/10.18653/v1/W17-4407>
- Gombert, A., Sánchez-López, B., & Cerquides, J. (2025). Jekyll institute or Mrs Hyde? Gender identification with machine learning. *Engineering Applications of Artificial Intelligence*, 144, 110087.  
<https://doi.org/10.1016/j.engappai.2025.110087>
- Gonzales, W. D. W. (2024). When to (not) split the infinitive: Factors governing patterns of syntactic variation in Twitter-style Philippine English. *English Language & Linguistics*, 28(2), 305–339.  
<https://doi.org/10.1017/S1360674323000631>
- Jiang, J., Chen, E., Luceri, L., Murić, G., Pierri, F., Chang, H.-C. H., & Ferrara, E. (2023). *What are Your Pronouns? Examining Gender Pronoun Usage on Twitter*.  
<https://doi.org/10.36190/2023.02>
- Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Writer Profiling Without the Writer’s Text. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics*

- (Vol. 10540. Springer International Publishing, pp. 537–558 [https://doi.org/10.1007/978-3-319-67256-4\\_43](https://doi.org/10.1007/978-3-319-67256-4_43)
- Keyes, O., May, C., & Carrell, A. (2021). You Keep Using That Word: Ways of Thinking about Gender in Computing Research. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 39, pp. 1–23. <https://doi.org/10.1145/3449113>
- Laitinen, M., & Fatemi, M. (2024). Testing the weak-tie hypothesis with social media. *11th Conference on Computer-Mediated Communication and Social Media Corpora*, pp. 46–51. [https://shs.hal.science/halshs-04673776/file/241007\\_CMC\\_Proceedings\\_DOI.pdf#page=60](https://shs.hal.science/halshs-04673776/file/241007_CMC_Proceedings_DOI.pdf#page=60)
- Laitinen, M., Rautionaho, P., Fatemi, M., & Halonen, M. (2025). Do we swear more with friends or with acquaintances? F#ck in social networks. *Lingua*, 320, 103931. <https://doi.org/10.1016/j.lingua.2025.103931>
- Lupo, L., Bose, P., Habibi, M., Hovy, D., & Schwarz, C. (2024). *DADIT: A Dataset for Demographic Classification of Italian Twitter Users and a Comparison of Prediction Methods* (No. arXiv:2403.05700). arXiv. <https://doi.org/10.48550/arXiv.2403.05700>
- Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), Article 1. <https://doi.org/10.1609/icwsm.v5i1.14168>
- O'Connor, K., Golder, S., Weissenbacher, D., Klein, A. Z., Magge, A., & Gonzalez-Hernandez, G. (2024). Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review. *Journal of Medical Internet Research*, 26(1), e47923. <https://doi.org/10.2196/47923>
- Office for National Statistics. (2022). *Baby names in England and Wales: From 1996 (Version 2022)* [Statistical dataset; XLSX]. Office for National Statistics (ONS). <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/babynamesinenglandandwalesfrom1996>
- Simeoni, F., Menéndez-Blanco, M., Vyas, R., & De Angeli, A. (2024). Querying the Quantification of the Queer: Data-Driven Visualisations of the Gender Spectrum. *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 3243–256. <https://doi.org/10.1145/3643834.3660695>
- Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLOS ONE*, 10(3), e0115545. <https://doi.org/10.1371/journal.pone.0115545>
- Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, 18(3), 74–84. <https://doi.org/10.5153/sro.3001>
- Statistics Canada. (2023). *First names at birth by sex at birth, selected indicators* [Dataset]. Government of Canada. <https://doi.org/10.25318/1710014701-ENG>
- Tonglet, J., Jehoul, A., Reusens, M., Reusens, M., & Baesens, B. (2024). Predicting the demographics of Twitter users with programmatic weak supervision. *TOP*, 32(3), pp. 354–390. <https://doi.org/10.1007/s11750-024-00666-y>
- Tucker, L., & Jones, J. (2023). Pronoun Lists in Profile Bios Display Increased Prevalence, Systematic Co-Presence with Other Keywords and Network Tie Clustering among US Twitter Users 2015–2022. *Journal of Quantitative Description: Digital Media*, 3. <https://doi.org/10.51685/jqd.2023.003>
- U.S. Census Bureau. (2016). *Frequently Occurring Surnames from the 2010 Census* [Dataset]. U.S. Census Bureau; U.S. Census Bureau. [https://www.census.gov/topics/population/genealogy/data/2010\\_surnames.html](https://www.census.gov/topics/population/genealogy/data/2010_surnames.html)
- U.S. Social Security Administration. (2024). *Baby Names from Social Security Card Applications—National Data* [CSV]. U.S. Social Security Administration. <https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data>
- Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *Proceedings of the 2019 World Wide Web Conference*, pp. 2056–2067. <https://doi.org/10.1145/3308558.3313684>
- Yildiz, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, 37(46), pp. 1477–1514. <https://doi.org/10.4054/DemRes.2017.37.46>

# **“I expected better from you, Mr. King”: Feminist resistance and reader critique in the subreddit r/MenWritingWomen**

**Marie Flesch, Heather Burnett**

Université de Lorraine, CNRS–Université Paris Cité  
marie.flesch@univ-lorraine.fr, heather.burnett@cnrs.fr

## **Abstract**

This study adopts a multi-layered approach to understanding the dynamics of gendered and feminist discourse in the subreddit r/MenWritingWomen, a Reddit forum dedicated to highlighting and critiquing how men describe women in literature and media. First, drawing on feminist critical discourse analysis, we focus on a sample of the posts themselves, examining how gender inequality is discursively constructed through stereotypical, objectifying, or reductive representations of women by male authors. Then, our analysis of a subset of comments about Stephen King, the most frequently cited writer in our sample, captures how Redditors react to, critique and defend these representations. Finally, using quantitative corpus linguistics methods, we explore the sociodemographic dimensions of participation, by identifying in the whole corpus users who self-disclose their gender in comments (e.g. “I’m a woman”).

**Keywords:** critical discourse analysis, Reddit, feminism, folk literary criticism, Stephen King

## **1. Introduction**

Reddit has often been characterized as a site of antifeminist activity, a breeding ground for the “manosphere” and a platform where misogynistic discourse flourishes (Massanari, 2017; Farrell et al., 2019; Helm et al., 2024). Yet, even within this largely masculinist ecosystem, “pockets of resistance and progressivism” emerge, challenging its dominant discourse (Massanari, 2019). Some of these spaces are openly feminist and progressive, such as r/feminism, or r/TwoXChromosomes, which states in its rules that “direct responses [...] should come from feminists and must reflect a feminist perspective”. Other communities are not explicitly feminist, but still do the “the work of everyday feminism”, often relying on humour (Massanari, 2019) (p. 19). One such space is r/MenWritingWomen. Created in October 2017, the subreddit highlights how male authors depict women in stereotypical, objectifying, or absurd ways. Users post excerpts from books, films, television, and other media, and commenters respond with critique, humour, and reflexive commentary. Thus, without declaring itself feminist, the subreddit aligns itself with a long tradition of feminist critique (Sundén and Paasonen, 2021): the critical examination of portrayals of women in writing by men, which emerged in the 1960s and 1970s (Eagleton, 2007). r/MenWritingWomen constitutes both a digital “sexism archive” (Ahmed, 2015), i.e. a collection of problematic representations of women, and a metalinguistic space where critique is collectively developed. Looking qualitatively at the (now inactive) Twitter account spawned by the subreddit, Sundén and Paasonen (2021) show that, far from the “feminist killjoy” trope (Ahmed, 2020) (p. 50), users employ the absurd as a feminist strategy, expressing feminist emotions of anger, rage, frustration, and bewilderment. Their study also reveals the limitations of the space, marked by a white, middle-class perspective.

The present paper builds on these insights, proposing a multi-layered approach to understanding the dynamics of gendered and feminist discourse in r/MenWritingWomen. It aims to answer several research questions: what are the

types of sexist representations discussed on the subreddit? What discursive strategies do commenters use to react to or critique the excerpts? Does the project of the community receive pushback? And who are the users who participate on the subreddit? Our analysis is divided into three interconnected parts. The first explores how women are represented by male authors in a corpus of 440 popular posts, focusing on the recurring sexist tropes used to construct women characters. This analysis draws on feminist critical discourse analysis (CDA) (Lazar, 2007) to reveal gendered stereotypes, from objectification of women and naturalization of women’s inferiority to women being defined by their relations with men. The second part examines how users of the subreddit respond to these representations in a subset of 13 discussion threads focused on the most frequently featured writer in our sample, horror author Stephen King, by using various discursive strategies (irony, parody, ethical critique, defence of author, etc.). Finally, the third part investigates the sociodemographic profile of the subreddit, using quantitative methods to extract information about the gender of Redditors.

## **2. Dataset and tools**

The dataset used in this study was obtained by scraping Reddit with the now defunct Pushshift API. It contains 829,732 comments and 15,251 posts dating from October 2017 to January 2023. It was pre-processed by removing duplicates, and posts and comments by Redditors who had deleted their account. Analyses were conducted with R version 4.4.1 (RCoreTeam, 2024).

## **3. Discursive representations of women**

### **3.1. Sample**

The sample of posts was created by selecting 1) posts that had the “Quote” tag (or “flair”, on Reddit), indicating that they contain an excerpt from a book, movie, etc. (as opposed to posts that contain commentary or satirical content) 2) posts that had a popularity score superior to 2500 points. 530 posts matched these two criteria. Each was accessed on



the subreddit using the URL provided in the metadata, so that we could view the images containing the excerpts. 49 posts had been removed or deleted and could not be coded; 41 posts were coded as “irrelevant”, because they contained commentary, satire, or memes instead of original excerpts from books and other media. The final dataset contains 440 posts.

### 3.2. Coding of posts

Coding of posts and comments was conducted by the first author of the paper. For each post, we coded the type of media discussed, the name of the author, when it was specified in the post title or in a comment, and the main discursive representation of women in the excerpt. The coding schemes of discursive representations was developed using an corpus-driven, iterative process informed by principles of feminist discourse analysis (Lazar, 2007), which emphasizes how language reflects and reinforces gendered hierarchies. The coding thus focuses on how language constructs, normalizes and perpetuates gender inequalities, with categories that capture discursive representations of gendered power relations. A first phase of open coding (without predefined categories) allowed us to identify recurring patterns; then, the initial categories were revised to merge redundant categories, and the resulting coding scheme was applied to the full dataset. The final coding scheme contains seven categories (Table 1).

### 3.3. Results

Representations of women discussed in the sample come from various sources. Most of the posts in the sample feature excerpts from literature (n=250), followed by social media (n=60), press (n=26), nonfiction books (21), movie and TV (n=16), and comics (n=17). The rest of the excerpts discussed (n=49) are from video games, recipes, advertisements, book covers, song lyrics, etc. Posts frequently, but not always, specify the name of the author. Authors of literary works the most frequently featured are Stephen King (n=20, or 8.85% of all posts in the literature section), Haruki Murakami (n=7), Charles Bukowski (n=4) and Ned Vizzini (n=4). The most frequent discursive representation of women in excerpts is, by far, objectification, with 293 posts, or 66.59% of all posts in the sample. Since this category is heterogeneous, we also coded for subframes (anatomical focus, oversexualization, unrealistic descriptions of women’s bodies or sexuality, etc.).

The anatomical focus is the most frequent subframe (n=198); Redditors frequently underline breasts descriptions (visually, in the image or screen capture, or through discourse, in the title of their post). The prominence of breast descriptions in the sample of posts is a reflection of r/MenWritingWomen’s central satirical critique; the description of the subreddit opens with the phrase “She breasted boobily down the stairs....”. Our analysis of posts shows that this deliberately absurd expression is not just a joke, but a cutting commentary on a recurring narrative pattern. Male authors repeatedly reduce female characters to their breasts, transforming them from complex individuals into fetishized caricatures. A Redditor, citing an excerpt from James Patterson (first row in Table 1) ti-

ties their post : “I know you’re reading a murder mystery but first, boobs”. Breasts are assigned personalities (“Her friendly nipples jiggled”, Philip E Dick), emotions (“Her breasts, of which she was normally proud, had withdrawn into themselves, as if depressed”, Jeffrey Eugenides; “Her big breasts, which had never suckled a child of her own, felt a merciless compassion”, Graham Greene) and actions (“her bosom changed from a promise to a threat”, Ross MacDonald). Even in death, women are defined by their breasts, like in this description of a corpse by an unspecified author (“They [her breasts] had perhaps lost some of their tone post-mortem, they sagged to this side and that”), or in this description of a character who has lost two friends: “She mourned their lovely breasts - breasts that have vanished without a trace” (Haruki Murakami). Even more disturbing to the Redditors is the focus on childlike imagery (“her breasts small and childlike”, James Herbert), and on the breasts of children (“She was a teenager who had a full set of breasts for milking”, Garth Stein).

## 4. Discursive strategies used by Redditors to discuss Stephen King’s writing of women

### 4.1. Sample

This section moves from the discursive representations of women by men to the discursive strategies used by Redditors to comment on the extracts posted on the subreddit. We decided to focus on Stephen King, the most frequently cited author in our sample of posts, and created a random sample of 13 discussion threads dating from August 2019 to May 2022 (13 posts, 1168 comments). The titles, dates and links to the threads are in the table in Appendix A.

### 4.2. Coding of comments

The coding scheme for comments is theoretically aligned with the coding of posts. It highlights how Redditors recognize, critique, and resist the sexist representations in posts, while also accounting for disagreement and pushback. Multi-labelling was allowed, as many comments use several strategies. The coding schema consists of six umbrella categories:

- **Critical interpretation:** ethical critique (frames text as ethically harmful), aesthetic critique (separates style from content), author critique (references to the author’s legacy or public image).
- **Discursive resistance:** irony/sarcasm/mock praise/humour, parody/satire, meta-commentary referring to common critiques of problematic content, textual intervention where Redditors propose edits of problematic content.
- **Identity and positioning:** subjectivization by gendered - or other- self-positioning, emotional responses, expression of feminist alignment.
- **Discursive negotiation/pushback:** defence of text/author, contextualization, criticism or hostility towards community.
- **Intertextuality:** community references (in-jokes, references to the norms of the subreddit), references to other texts, authors or famous people.

Discursive representation	Description	Example	Frequency
Objectification	Women are reduced to body parts, for consumption by men.	“Kimberly was longed-legged as well, with firm, sculpted breasts and a glowing tan. Her nipples were already erect.” (James Patterson, <i>The Midnight Club</i> )	293
Relational reduction	Women’s value is defined in relation to men; includes women’s rivalry and lesbophobia.	“The only negative was the nasty rumour. Surely she couldn’t go for women. She was too perfect [...]. She was destined to be a trophy wife!” (John Grisham, <i>The King of Torts</i> )	48
Naturalized inferiority	Women’s inferiority is framed as being natural or biological, justifying inequalities.	“Comparing man and woman in general one may say: woman would not have the genius for finery if she did not have the instinct from the secondary role.” (Friedrich Nietzsche, <i>Beyond Good and Evil</i> )	27
Pathologized femininity	Women’s emotions and behaviors are framed as irrational or excessive.	“They [space pods] were usually christened with feminine names, perhaps in recognition of the fact that their personalities were sometimes slightly unpredictable.” (Arthur C. Clarke, 2001: <i>A Space Odyssey</i> )	27
Sexual danger/control	Women’s sexuality is framed as dangerous; includes coercive or manipulative sex tropes.	“All women love semi-rape. They love to be taken.” (Ian Fleming, <i>The Spy Who Loved Me</i> )	22
Devalued/deviant body	Women’s bodies are devalued because of their age, disability, or racialization.	“Madame Danglars, whose beauty was quite remarkable in spite of her thirty-six years” (Alexandre Dumas, <i>The Count of Monte-Cristo</i> ).	11
Other/ambiguous	Satire, memes, commentary, etc.	–	12

Table 1: Description of the coding scheme for posts.

- **Not relevant:** non evaluative, ambiguous.

### 4.3. Results

The critical interpretation category is the most represented in comments (n=510; see also Figure 1), which shows the subreddit’s investment in highlighting the implications of problematic writing. Some comments target technical aspects of the writing (e.g. “What on earth is happening with that godawful comma splice”; “Did he have to mention that she was sitting up twice?”). Critique focuses however mostly on thematic content, particularly the objectification of women and girls, in the excerpts discussed (“Yeah I always thought this was fucking creepy and honestly pedophilic as hell”) or in Stephen King’s writing in general (“It’s crazy that this man is one of the best selling authors of all time. So many people are influenced by the casual misogyny in his stories.”). Redditors use a variety of affective registers, including anger/exhaustion/disgust (n=275; “Ew ew ew!”, “Gross”, “Da fuq did I just read?”), and sarcasm and humour (n=257) (“That’s art!”, “Stephen King and unnecessarily talking about tits, name a more iconic duo.”). The frequent recourse to humour can be seen as a rhetorical strategy but also as a community-building tactic. In particular, reversal seems to be a key strategy, used to underline the misogyny of King’s descriptions. Redditors imagine alternative scenarios where men and boys are depicted the way women and girls are, either through commentary (“He loves to describe every woman’s breast situation, but never bothers to mention how the boys are hanging on the men.”) or parody (“A boy once afraid of lightning



Figure 1: Treemap chart showing the proportion of discursive strategies (including sub-categories) in the Steven King’s sample.

now stands before me at the age of 12, with ripe, pendulous balls swinging to and fro within the confinement of his basketball shorts.”). These exaggerated parodies reflect a broader feminist tradition of using satire to denaturalize sexist tropes (Fahs, 2019; Sundén and Paasonen, 2021).

Other comments use comparison to controversial authors and political figures as a strategy, positioning King’s writing within a broader tradition of patriarchal and literary power. The most frequent comparison is to Donald Trump (n=18), whose candidacy at the 2016 election the writer opposed (“Is the narrator Donald Trump?”, “Donald Trump wrote another book?”). Vladimir Nabokov is

also mentioned four times (“sounds like lolita”, “Written by Vladimir Nabokov...”).

Not all comments are negative: 201 comments show appreciation for the writer or defend his writing. A number ( $n=73$ ) recognize the validity of criticism, revealing an ambivalent stance (“stephen king: Great horror, not great character descriptions.”; “It’s good writing with gross content.”); they reflect what popular culture researchers have called “critical fandom”, i.e. the political and critical engagement of fans (Andersen and Jensen, 2020; Svegaard, 2015).

Three main arguments emerge both in comments that criticize the writer and in comments that defend him (coded as “discursive negotiation/pushback”,  $n=320$ ). The first is biographical: 40 comments refer to the drug and alcohol addiction the writer struggled with, stating in his memoir that he barely remembers writing his novel *Cujo* (King, 2000). Some Redditors use this to explain or excuse the misogyny (“He was so drugged up at that point, the structure of his brain probably looked a lot more like jelly than an actual brain.”), while others reject this rationale (“Okay about the whole ‘Cocaine’ thing. Drugs are amplifiers or dampeners NOT excuses. That shit came from him. The author. King.”). The second argument used by Redditors to defend King is the contextualization of women and girls’ descriptions, mostly by pointing out that the point of view is not that of the writer (“Another case of this being from the POV of a male character”, “This one is from the perspective of a little boy.”). The validity of the argument is vehemently rejected by some (“‘Stephen king is writing from the perspective of gross men and that’s why it sounds so gross!’ Yeah and maybe he should write fewer gross men??? Just saying??”, “It’S bEcAuSe ThE pOv ChArAcTeR iS a WeIrDo!”). Finally, some comments explain the “bad” writing of women’s characters by the fact that King wrote many (65) novels (“This is what happens when you write books way too fast”).

Many comments show that Redditors are aware that King is a “staple” of the subreddit (“As much as I love his outlandish stories, we should have our own tag for Stephen King.”), one Redditor even addressing the writer with the “u/” prefix used in Reddit usernames, as if he were a member of r/MenWritingWomen (“u/stephenkinghere what happened here?”). In total, 27 comments address the author directly (“Seriously, King?”, “What the fuck Stephen King. Why can’t you be normal?” “Really Stephen, we don’t need this much detail. This is why I quit your books.”). Finally, while most comments in the “discursive negotiation/pushback” are aimed at King, a number ( $n=44$ ) are hostile or critical of the subreddit itself (“You people need a new hobby”, “Wow this sub is getting whiny. Should they just not include women at all now?”). These comments show that that feminist critique can be contested in the subreddit, revealing its ambivalent relationship to feminism.

## 5. Who’s critiquing?

In order to explore the sociodemographic make-up of the forum, we queried the corpus using the key expression *I am a(n) / I’m a(n)* using the Quanteda R package (Benoit et al., 2018). The 12,740 concordance lines obtained, pro-

duced by 8936 unique Redditors, were manually inspected and coded as being written by women, men, and non-binary individuals, with statements such as “I’m a guy who used to be shy” and “I’m a hetero woman”. The non-binary category is an umbrella category that includes several gender identities (non-binary, genderfluid, agender, etc.). During this process, we uncovered a lot of satirical, ironic and fictional statements, that we were careful not to code as indicative of a gender identity. We inferred gender identities for 3021 Redditors, or 1.74% of all 173,898 Redditors in the corpus; these Redditors wrote 99,644 comments, or 11.79% of all comments in the corpus. In this subset of Redditors, there are 1642 men (or 54.35% of Redditors that disclosed their gender identities), 1362 women (45.08%) and 17 non-binary individuals (0.56%). The median number of comments per person is higher for women (38.51, *IQR* = 95.06) than for men (28.35, *IQR*=78.01). This difference is significant according to the Wilcoxon-Mann-Whitney (one-tailed test,  $W=1302912$ ,  $p<0.001$ ). Among prolific Redditors (who wrote more than 50 comments), women are slightly over-represented (417 women, 352 men, 5 non-binary individuals). The length of comments is not correlated with gender; in other words, women did not write longer comments than men (Quasi-Poisson regression model with author as a random effect,  $p=0.11$ ).

## 6. Conclusions

In this paper, which combines feminist critical discourse analysis with corpus linguistics methods, we examined 440 of the most popular posts on r/MenWritingWomen between 2017 and 2022, as well as a subsample of 13 discussion threads about Stephen King, and presented a sociodemographic analysis of the subreddit. We have shown that objectification, and particularly the focus on breasts, is by far the most frequent representation of women in our sample. It confirms this to be the central – or at least the most popular – critique of the subreddit, aligned with its description/motto “She breasted boobily down the stairs.....”.

Looking more closely at comments about Stephen King, we found that Redditors criticize misogynistic content in various ways, using anger, disgust, humour, irony, comparisons, or direct addresses to the author. A number of comments defend the author or contextualize the excerpts; many are ambivalent and complex, showing that critique and defense often coexist. Finally, our sociodemographic analysis showed that both women and men actively contribute to the subreddit. By contrast, r/TrollXChromosomes, another subreddit doing “everyday feminism” through humour, has a less diverse commenters’ base (68.3% women, according to Burkhart (nd)). This finding may be interpreted in two different lights. If we look at Reddit as a men-dominated, often hostile environment to women, the subreddit is characterized by the notable presence of women. If we look at the subreddit’s project as a feminist one, such a strong presence of men may be surprising.

Taken together, our qualitative and quantitative findings suggest that r/MenWritingWomen cannot be straightforwardly characterized as a feminist space. Not all Redditors who comment on the subreddit engage in feminist cri-

tique; the main focus of critique in our sample (women's breasts) is quite narrow; and other axes of inequality, for example related to questions of race, ethnicity, class or queerness, are largely sidelined. In the Steven King's sub-sample, a Redditor expresses their frustration with this narrow lens: "I'm more concerned about how female characters behave and see the world, to me that's 'men writing women'. Am I completely misunderstanding the sub and its purpose?". This comment highlights a disconnect between the subreddit's stated mission and the kind of critique that is the most visible and rewarded. It seems that broader Reddit norms, like the focus on women's bodies and the reward system (karma) that valorizes sensational content (Richterich, 2014), may shape and narrow feminist critique. *r/MenWritingWomen* is thus an ambivalent space, that operates both as a feminist consciousness-raising platform and a more "typical" subreddit.

## 7. Acknowledgements

This work has received funding from the French National Research Agency (Agence Nationale de la Recherche) under the "ANR-24-AERC-0011-01" project; and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 850539).

## 8. References

- Ahmed, S. (2015). Introduction: Sexism - a problem with a name. *new formations: a journal of culture/theory/politics*, 86(1):5–13.
- Ahmed, S. (2020). *The promise of happiness*. Duke University Press.
- Andersen, T. F. and Jensen, T. (2020). Tintin and the adventure of transformative and critical fandom. *Participations. Journal of Audience Reception Studies*, 17(2).
- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.
- Burkhart, B. (nd). Subreddit gender ratios.
- Eagleton, M., (2007). *Literary representations of women*, pages 105–119. Cambridge University Press.
- Fahs, B. (2019). Reinvigorating the traditions of second-wave radical feminism: Humor and satire as political work. *Women's Reproductive Health*, 6(3):157–160.
- 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 - WebSci '19*, pages 87–96, Boston, Massachusetts, USA. ACM Press.
- Helm, B., Scrivens, R., Holt, T. J., Chermak, S., and Frank, R. (2024). Examining incel subculture on reddit. *Journal of Crime and Justice*, 47(1):27–45.
- King, S. (2000). *On writing: A memoir of the craft*. Simon and Schuster.
- Lazar, M. M. (2007). Feminist critical discourse analysis: Articulating a feminist discourse praxis. *Critical discourse studies*, 4(2):141–164.
- Massanari, A. (2017). #gamergate and the fapping: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media Society*, 19(3):329–346.
- Massanari, A. L. (2019). "come for the period comics. stay for the cultural awareness": reclaiming the troll identity through feminist humor on reddit's *r/trollxchromosomes*. *Feminist Media Studies*, 19(1):19–37.
- RCoreTeam. (2024). *R: a language and environment for statistical computing*. Vienna, Austria.
- Richterich, A. (2014). 'karma, precious karma!' karmawhoring on reddit and the front page's econometrisation. *Journal of Peer Production*, 4(1).
- Sundén, J. and Paasonen, S. (2021). "we have tiny purses in our vaginas!!! thanksforthat": absurdity as a feminist method of intervention. *Qualitative Research Journal*, 21(3):233–243.
- Svegaard, S. F. (2015). Critical vidders: fandom, critical theory and media. *Academic Quarter/Akademisk Kvarter*, 11:104–114.

## A Appendix

title	date	comm.
Ah yes, a completely normal way to describe one's DAUGHTER.	2019-08-28	176
You would think this sexy corpse was satire...	2019-11-04	93
He's gotta be taking the piss out of himself at this point...right? [Stephen King, Doctor Sleep]	2020-03-22	91
A man spends time with his mother for the first time since he left home. He wonders why their dynamic seems different. Could it be the size of her breasts? (The Stand, Stephen King)	2020-04-14	90
How do your nipples react to anti-semitism? [IT by Stephen King]	2020-05-20	64
steven king strikes again	2020-09-14	80
Her orifices aren't welcoming enough	2021-05-26	75
Remember when Stephen King wrote about a sexually abused 12 year old having sex with all her friends (and having an orgasm from two of them)?	2021-07-06	226
Yes, because I really need to know that the lost 9yr old doesn't have breasts. Thanks, Stephen King.	2021-07-11	79
I, too, experienced menopausal depression prior to my first day of kindergarten (from Pet Sematary by Stephen King)	2021-07-29	37
Whyyyy must he write like this?! From Skeleton Crew by Stephen King	2022-03-24	45
Good ol' Stephen King and the sand vagina (Roadwork)	2022-04-04	63
Stephen King - The Shining	2022-05-13	49

Table 2: Sub-sample of discussion threads about Stephen King analyzed in section 4., with hyperlinks to the posts on Reddit, dates and number of comments per thread in our corpus (NB: these numbers may not match the number of comments on the website, as scraping was performed several years ago).

# OMG! Why discourse markers thrive in interactive social media writing

Reinhild Vandekerckhove

E-mail: reinhild.vandekerckhove@uantwerpen.be

## Abstract

The present paper examines the use of *omg* ('oh my god') in online teenage talk. Both the quantitative and qualitative analyses reveal that *omg* has 'bleached' and grammaticalized/pragmaticalized from an interjection expressing strong surprise or shock into a discourse marker with a primarily metatextual function, serving both the interpersonal connection between the interlocutors and the organization of the conversation. It is argued that interactive social media writing provides a fertile ground for the transformation of interjections into discourse markers that perform functions highly comparable to those of emoticons and are therefore well-suited to the demands of this type of communication. Apart from their functional parallels, the female preference for *omg* emerged as another striking similarity between this discourse marker and the use of emoticons.

**Keywords:** adolescent social media writing, discourse markers, pragmaticalization

## 1. Introduction

Two Flemish teenage girls are having a Messenger chat on going to the movies, when a third friend joins them by posting the following message: *Omg ja kwil vrijdag cine* 'omg yes I want cinema on Friday'. The initialism *omg* ('oh my god') appears to highlight her agreement with the suggestion made by the other girls earlier in the conversation. The present paper is going to focus on this use of *omg*. By discussing the appearance and behavior of *omg* in online teenage chat, we aim to demonstrate that interactive social media writing provides a fertile breeding ground for this type of discourse markers.

Before analysing the occurrences of *omg* in Flemish teenage chat, we take a look at its origin. The abbreviated form is clearly derived from the full interjection 'oh my god'. Cambridge Advanced Learner's Dictionary & Thesaurus defines "(Oh my) God!" as informal "idiom" "used to emphasize how surprised, angry, shocked, etc. you are"<sup>1</sup>. Judging from the exclamation mark, the utterance is considered to have an exclamatory character. This is confirmed by Oxford English Dictionary (OED) that lists the abbreviated form "OMG", labels it as an interjection and renders it in full as "oh my God!", again with exclamation mark<sup>2</sup>. Interestingly OED adds: "frequently in the language of electronic communications".

The *omg* produced by the Flemish teenager who seems to like the suggestion of her friends can hardly be considered an exclamation or expression of surprise. It has clearly undergone some semantic bleaching which might make it a pragmatic marker in much the same way as for instance *lol* ('laughing out loud'), which has been found to have undergone a grammaticalization process in informal CMC (McWhorter, 2016, 128), "in the pragmatic department, modal wing" (McWhorter, 2016, 218). It is precisely this grammaticalization or pragmaticalization process which makes it a typical discourse marker (see Narrog & Heine, 2011, 4 for literature on pragmaticalization as a sub-process of grammaticalization). Still discourse markers, in view of the inherent fluidity of the category, are hard to define.

Heine et al. (2021, 6) offer a good starting point by acknowledging the context-dependent multifunctional character of discourse markers: "Discourse markers are (a) invariable expressions which are (b) semantically and syntactically independent from their environment, (c) set off prosodically, from the rest of the utterance in some way, and (d) their function is metatextual, being anchored in the situation of discourse and serving the organization of texts, the attitudes of the speaker, and/or speaker-hearer interaction." Since we are dealing with social media writing, and not with speech, prosody does not come into play. However, Onodera (2011, 620) offers an interesting alternative for (c) in the definition above by claiming that a universal characteristic of discourse markers is their "initialness", i.e. they tend to occur in initial position. While this generalization is difficult to sustain, both for speech and written CMC (see e.g. Degand & van Bergen 2018 for a discussion of utterance-final discourse markers in both face-to-face conversations and Instant Messaging), this characteristic does seem to make sense for *omg*, which – as will be shown below – relates to its pragmatic function (see also Aijmer 2013, 16 for the relation between placement and function of pragmatic markers).

In what follows we will examine both in a quantitative and qualitative way to what extent occurrences of *omg* in informal Flemish teenage chat suggest that it has shifted from an exclamatory interjection with a specific meaning into a more general discourse marker. The discussion presented in this introductory section will serve as our frame. However, we start with some explanation on the database.

## 2. Data and coding

The data are extracted from an anonymized corpus of private social media conversations produced on WhatsApp and Facebook Messenger in 2015-2016 by Flemish teenagers aged 13-20. The data were collected via secondary schools. On a voluntary basis, students donated conversations produced outside the school context and provided relevant metadata on their social profile. The

<sup>1</sup> <https://dictionary.cambridge.org/dictionary/english/oh-my-god>

<sup>2</sup> [https://www.oed.com/dictionary/omg\\_int](https://www.oed.com/dictionary/omg_int)

corpus consists of 456 751 posts (2 653 924 tokens). For more information about the data collection, we refer to Hilte et al. (2020)<sup>3</sup>. Since gender will surface as an important variable in the analysis below, we present the gender distribution in terms of posts in table 1. The gender balance in terms of participants is less skewed: the corpus contains conversations produced by 724 adolescents who identified as female (51,59%) and 674 adolescents who identified as male (48,21%):

variable	levels	posts
gender	girls	301 189 (65.94%)
	boys	155 562 (34.06%)

Table 1: gender distribution in the entire corpus

From this corpus we extracted all posts that contained tokens of *omg* (including expressive variants like *omggg* and *ooooomg*)<sup>4</sup>. This search rendered 2179 tokens of the target form. Since “pragmatic markers get their meaning in interaction with the context” (Aijmer 2023, 13), we added the two preceding posts and the post following the selected posts. Full variants proved rare in the corpus. E.g., there are 10 tokens of *o my god* and 9 of *oh my god*. They are excluded from the analysis.

In view of the focus of the present study and the frame presented in the introduction, we coded the data for two parameters: the ‘initialness’ of *omg* and its exclamatory character. Initialness was operationalized as ‘post-initialness’, i.e. *omg* opening the post produced by the chatter. In order to determine whether *omg* has an exclamatory character or not, we coded for three features: (1) the occurrence of one of more exclamation marks, or a combination of question and exclamation marks after *omg* itself or closing the sentence or phrase in which *omg* occurs, (2) the use of capital letters when rendering *omg* (*OMG*), (3) the presence of letter repetition when rendering *omg* (e.g. *omggg*). The latter two are commonly used expressive markers in social media writing (see e.g. Hilte et al. 2018) that are considered proxies for exclamation, since they suggest at least some degree of emphasis. The results of this coding will be dealt with in a quantitative way (par. 3.1). Apart from that we also need a qualitative look at the data to assess the exact pragmatic function of the initialism (par. 3.2).

### 3. Results

#### 3.1. Quantitative analysis

The extent to which occurrences of *omg* have an exclamative character may indicate how closely the the interjection still adheres to its original meaning and function, namely an exclamation of (strong) surprise or even ‘shock’ (see OED).

Table 2 renders the results of the coding for this parameter, both in absolute and relative numbers. The proportions of tokens marked by the different features by which we chose

to operationalize ‘exclamativity’ appear highly comparable:

Exclamation marks	Capitalization	Letter repetition
222 (10,19%)	249 (11,43%)	270 (12,39%)

Table 2: Exclamativity marking of the *omg*-tokens

30,38% (662/2179) of the *omg*-occurrences are accompanied by at least one of the exclamativity markers. In other words, less than one third of the tokens have some exclamative/emphatic character, two thirds do not. Only 3,63% (79/2179) of the tokens bear two exclamative features (see e.g. the *OMG!* In the title of this paper). Strikingly, none of the 2179 *omg*-occurrences combines the three features. Moreover, it should be added that this quantitative operationalization of exclamativity, not taking into account the context, inevitably produces false positives. E.g.: When a sixteen-year-old girl posts *HAHAHAHAHA OMG XD* because she finds the previous posts of her friend very funny and she simply wants to stress this, we can hardly say this is an exclamation, let alone an exclamation expressing surprise. However, the capitalization does add to the emphatic character of the utterance. Nevertheless, as can be deduced from table 2, the exclamatory and highly emphatic character of the original interjection *oh my god* seems largely subordinate in these chat data, suggesting that the compressed form *omg* might indeed have undergone semantic bleaching in informal CMC-writing to the extent that it has become a more general discourse marker. A closer look at the conversational context will reveal how this functions (par. 3.2).

If *omg* has indeed pragmaticalized into a general discourse marker, it is more likely to occur at “transitions in the discourse” (Aijmer 2013, 7) and – following Onodera (2011, 620) – may be especially favoured at the beginning of posts. Table 3 confirms this is the case in the Flemish adolescent conversations:

Initial position	non-initial position
1685 (77,3%)	494 (22,7%)

Table 3: Position of *omg*-tokens within chat posts

*Omg* is much more frequent in initial position than in non-initial position. Posts that only consist of the target form, unaccompanied by any other words, were also included in the first category. However, these were much less frequent than multiple-word posts.

Finally, though this was not the initial focus of the study, we were struck by a distinct gender pattern when coding the data: The proportion of chatters using at least one token of the target form amounts to nearly one third of the female adolescents (237/724), as opposed to only 7,6% (51/674) of their male peers (see table 4). The difference is significant ( $X^2 = 135.17$ ,  $p < .0001$ ). Moreover, women

<sup>3</sup> Ethical clearance for data collection and secure data storage within research group CLiPS was given by the Ethical Advisory Committee for Social and Human

Sciences of the University of Antwerp.

<sup>4</sup> With sincere thanks to Lisa Hilte for extracting the data.

appear to use it much more abundantly: 94,4% of the *omg*-tokens are produced by female adolescents, which means that the gender balance is much more skewed than the overall gender distribution in the corpus (see table 1) and that *omg* is predominantly a marker of female speech.

	Female adolescents	Male adolescents
<i>omg</i> -users	32,7%	7,6%
<i>omg</i> -tokens	94,4%	5,6%

Table 4: Use of *omg* related to gender of the chatters

### 3.2 Qualitative analysis

In the present section we focus on the meaning and function of *omg* and the extent to which it has ‘bleached’ and pragmaticalized from an interjection expressing strong surprise into a discourse marker the function of which is primarily “metatextual”, “serving the organization of texts, the attitudes of the speaker, and/or speaker-hearer interaction” (Heine et al., 2021, 6, see full definition in par. 1).

First of all, we observe that the original use of the interjection has not entirely disappeared: the adolescent corpus still contains examples in which *omg* appears in exclamatory utterances expressing strong surprise or even shock. E.g.:

- (1) *Meent gy da nu srs! Omg! Gy bent ni te doen!* ‘Do you really mean this, serious! Omg! You are totally impossible!’ (14-year-old girl)

Generally speaking, however, *omg* primarily serves the organization of the interaction, e.g. by marking the turn-taking, and/or the interpersonal connection. With respect to turn-taking, *omg* often marks the switch from one chatter to another, both when it is in post-initial position, or when it occurs in isolation, in that case often followed by another post of the chatter who has taken over:

- (2) Girl 1: *Ik voel het* ‘I feel it’  
 Girl 1: *Hij zoekt ruzie* ‘he’s looking for a fight’  
 Girl 2: *omg*  
 Girl 2: *hahahahahha*

At the same time, this use of *omg* shows engagement on the part of the addressee toward the interlocutor. That seems to be the main function of *omg*: it is a backchanneling device which signals attention for the message of the interlocutor, which means that *omg* in teenage chat tends to be the equivalent of ‘I hear you’ or of an enthusiastic nod, since the attitudinal dimension in the metatextual function as described by Heine et al. (2021, see definition above) is seldom completely absent. However, the strong emotional involvement inherent to the original interjection ‘oh my god’ is mostly absent. The post which opened the present paper (see par. 1) serves as a good example of this enthusiastic backchanneling, and so does the following extract:

- (3) Girl 1: *Als we naar ieper gaan..?* ‘If we go to Ypres..?’  
 Girl 1: *Zouden we gene mc do passeren?*

‘Wouldn’t we pass a McDonald then?’

Girl 2: *Omg ja mss wel? :D* ‘omg yes maybe indeed? :D’

In example (4) the addressee acknowledges the hilarity of the situation outlined by her friend:

- (4) Girl 1: *ja haha da was echt gieren* ‘yes haha that was really funny’

Girl 1: *en dan kwamen we xnaamx<sup>5</sup> tegen met dieje kerel weete nog?* ‘and then we bumped into xname with that dude, remember?’

Girl 2: *Omg ja das waar haha* 🤔 ‘omg yes that’s right’

Girl 1: *Busted*

Finally, in example (5) ‘girl 2’ makes clear she registers and appreciates the little joke of her friend, though from her final reaction we can deduce she might have considered it somewhat silly. They are discussing a camp they will build, and how they will decorate it:

- (5) Girl 1: *mss versiering of zo voor in het kamp?* ‘maybe some decoration for the camp?’

Girl 2: *gwn stokken en blaadjes, niet?* ‘just branches and leaves, no?’

Girl 1: *ma das zo droog* ‘but this is so dry (meaning: this is so uninspired)’

Girl 2: *Niet als het regent* ‘not when it rains’

Girl 1: *hha omg*

Girl 1: *humor van xnaamx* ‘humor of xname’

In all of these cases *omg* primarily expresses (positive/enthusiastic) confirmation, thus behaving like a typical discourse marker serving the interpersonal connection, with the “speaker-hearer interaction” in the definition of Heine et al. (2021, 6) obviously being replaced by writer and addressee. In most of these cases *omg* typically occupies the post-initial position. Onodera (2011, 623) claims that a discourse marker “must appear before what is to be informed. This is what the marker highlights, and it is the upcoming speaker’s intention/action/subjective strategy.” While this post-initial position may be overstated by Onodera (2011) (our corpus contains quite a lot of counterexamples (see table 3) and Degand & van Bergen 2018 discuss the role of utterance-final discourse markers in Instant Messaging), *omg* indeed highlights the writer’s intention and that intention mainly consists in giving the interlocutor the idea that one is genuinely interested and engaged in what they are posting. At the same time, it contributes to the ‘text organization’ by “elaborating on the preceding discourse” (Heine et al, 2021, 8) and/or marking the turn-taking.

## 4. Discussion

The chat conversations of the Flemish adolescents confirm that *omg* has undergone a development comparable to the shift documented for *lol* (‘laughing out loud’) by McWhorter (2016, 128) in informal CMC-writing: it has grammaticalized, or more specifically pragmaticalized,

<sup>5</sup> Names are replaced by xnaamx (xname, in English)



into a discourse marker. It displays the typical metatextual functions associated with discourse markers, serving both the interpersonal connection between the interlocutors, by highlighting (mainly positive) confirmation and engagement with (the messages of) the other(s), and the organization of the conversation, by signaling one is elaborating on the preceding discourse and by marking the turn-taking. Regarding the latter function, its predominantly post-initial position is clearly helpful in that respect. However, this does not exclude the use of utterance-final discourse markers as turn-marking devices: As stated above, discourse markers tend to occur at “transitions in the discourse” (Aijmer 2013, 7). Moreover, Degand and Van Bergen (2018, 19) demonstrate that utterance-final discourse markers in Instant Messaging, signal “that the floor is open to the addressee”. In other words, discourse markers occurring at the boundaries of posts/messages in informal interactive online writing seem very useful either in marking the turn-taking (post-initial) or the turn-yielding (post-final).

Apart from that, it is not surprising that social media writing provides a fertile ground for the transformation of interjections like *lol* and *omg* into discourse markers. Previous research has shown that chat participants commonly use a variety of expressive markers in informal, interactive online writing to compensate for the lack of body language and other non-verbal cues, such as those used to signal engagement with interlocutors, in face-to-face communication (see e.g. Hilte et al. 2018). Especially emoticons often serve this social function. Spina (2019, 346), for instance, notes: “Emoticons are a kind of relational icon (Asteroff, 1987), promoting rapport” (see also Spina 2017, 25). Strikingly, a discourse marker like *omg* has a lot in common with this emoticon-function, when compensating for the absence of non-verbal signs of engagement and empathy. Another parallel relates to the gender distribution of both features: Previous research has shown that women tend to use (much) more emoticons than men (e.g. Hilte et al. 2018, Schwarz et al. 2013). Similarly, *omg* is also primarily characteristic of female interaction. Since the early days of sociolinguistics, scholars have emphasized the connective dimension in women’s discourse (e.g., Tannen 1990) and this very function appears to prevail in the way the Flemish adolescent chatters employ *omg*. Finally, both emoticons and discourse markers like *omg* can have a structural function. While emoticons tend to prefer post-final position, often replacing punctuation marks (see Spina 2019), *omg* has a preference for the initial position, marking turn-taking and elaboration on previous discourse, which makes both features perfectly complementary.

Taken together, these observations indicate that *omg*, like *lol*, has evolved into a discourse marker well-suited to the demands of interactive social media writing and that, consequently, other interjections or similar expressions may follow a similar path, stimulated by the dynamics of written online communication.

## 5. References

- Aijmer, K. (2013). *Understanding pragmatic markers. A variational pragmatic approach*. Edinburgh: Edinburgh University Press.
- Hilte, L., Vandekerckhove, R. & Daelemans, W. (2018). Expressive markers in online teenage talk: a correlational analysis. *Nederlandse Taalkunde* 23(3), pp. 293–323.
- Hilte, L., Vandekerckhove, R. & Daelemans, W. (2020). Linguistic accommodation in teenagers’ social media writing: convergence patterns in mixed-gender conversations. *Journal of Quantitative Linguistics* 29(2), pp. 241–268.
- Asteroff, J. F. (1987). *Paralanguage in electronic mail: A case study*. PhD dissertation, Columbia University, New York.
- Degand, L. & van Bergen, G. (2018). Discourse Markers as Turn-Transition Devices: Evidence From Speech and Instant Messaging. *Discourse Processes: a multidisciplinary journal*, 55(1), pp. 47-71.
- Heine, B., Kaltenböck, G., Kuteva, T. & Long, H. (2021). *The Rise of Discourse Markers*. Cambridge : Cambridge University Press.
- McWorther, J. (2016). *Words on the Move: Why English Won't - and Can't - Sit Still (Like, Literally)*. New York: Picador.
- Narrog, H. & Heine, B. (2011). Introduction. In H. Narrog & B. Heine, *The Oxford Handbook of Grammaticalization*, pp. 1–16. Oxford: Oxford University Press.
- Onodera, N.O (2011). The grammaticalization of discourse Markers. In H. Narrog & B. Heine, *The Oxford Handbook of Grammaticalization*, pp. 614–624. Oxford: Oxford University Press.
- Schwartz, H.A., Eichstaedt, J.C., Kern M.L., Dziurzynski L., Ramones S.M., Agrawal, M, Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P & Ungar, L.H.(2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS One* 8(9), <https://doi.org/10.1371/journal.pone.0073791>
- Spina, S. (2017). Spina, Stefania (2017). Emoticons as multifunctional and pragmatic Resources: a corpus-based Study on Twitter. In E.W. Stemle & C. R. Wigham (Eds.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*, pp. 25–29.
- Spina, S. (2019). Role of Emoticons as Structural Markers in Twitter Interactions. *Discourse Processes* 56(4), pp. 345-362.

# Emoji and Emoticon Use in Online Dating Profiles and Chats: A Corpus Study into Functions and Categories

Lieke Verheijen, Tess van der Zanden

Radboud University, Utrecht University (the Netherlands)

E-mail: lieke.verheijen@ru.nl, t.vanderzanden@uu.nl

## Abstract

Online dating has become mainstream in today's society. Computer-mediated communication through online dating platforms is highly multimodal. An increasingly pervasive visual element of computer-mediated dating is emoji. This paper presents the first large-scale quantitative corpus analysis into the use of emoji and emoticons in dating profiles ( $n = 79$ ) and dating chats ( $n = 125$ ). All emoji and emoticons in the corpus were coded for their categories and functions. Findings reveal that emoji are used in significantly different ways in dating profiles as compared to dating chats. These results seem to correspond to the different purposes of these communicative contexts, namely self-presentation and impression management in profiles versus emotion expression and social connectedness in chats. Gender differences were also explored: men and women were found to differ in their use of emoji and emoticons in our online dating corpus, with women using more of such graphicons, especially for expressing emotions.

**Keywords:** emoji, emoticons, online dating, dating profiles, dating chats

## 1. Introduction

In today's digital society, many people search for romantic or sexual partners through online dating, both for long-term committed relationships and for casual flings. The popularity of online dating has surged: it has become a mainstream way to meet new people and form relationships. Over 350 million people worldwide take part in online dating (Cloudwards, 2025). There is a plethora of online dating platforms available – websites as well as apps, with Tinder being the most popular in the US, Europe, and the Netherlands (ibid.; Statista, 2024).

Dating platforms involve two main communication contexts. First, online dating profiles, which daters use to present themselves, to manage other daters' impressions about themselves, and to decide if others spark sufficient interest to connect with. Second, online dating chats, in which daters with a mutual interest or 'match' get better acquainted through instant messaging. Similar to other genres of computer-mediated communication (CMC), computer-mediated dating has become more multimodal: self-presentation on online profiles and conversations via online chat involve various visual elements next to verbal elements (i.e., text). Such visual elements or 'graphicons' include videos, photos/images, GIFs, and stickers, but also emoji and emoticons (Herring & Dainas, 2017; Konrad et al., 2020). This increasing visualisation of online communication as well as online dating reflects the ever-changing technological affordances and constraints of social media platforms and dating apps.

Emoji are a salient visual element in online dating. Self-reported usage frequencies are discrepant, but according to US survey research, most single online daters (between 62% and 97%) report to use emoji in courting potential dates (Gesselman et al., 2019), which underscores their ubiquity in online dating. Emoji are a 'multifunctional resource' in CMC (Verheijen & Mauro, 2025): they can function as paralinguistic cues, to compensate for the lack of non-verbal cues (facial expressions and gestures) in writing, help convey tone and sentiment, and can make written text more informal, playful, and warm, among other things (for extensive reviews on emoji research, see Bai et

al., 2019; Chiang & Gomez-Zara, 2024; Manganari, 2021; Setyawan & Musthafa, 2024; Tang & Hew, 2019). Motives for using emoji and emoticons in messages of singles to potential dates, according to Gesselman et al. (2019), include adding a touch of personality, expressing feelings, speed and convenience, and mirroring one's interlocutor. Studying emoji and emoticon usage in the online dating context can provide valuable insights into the constantly evolving landscape of CMC and, in particular, its impact on digital relationship dynamics. Understanding how these graphicons are used in online dating sheds light on how people use such visuals to present and express themselves for building connections online.

However, so far, few studies have been conducted on emoji in online dating. These revealed that emoji use with potential partners is strongly associated with more romantic and sexual interactions (Gesselman et al., 2019), that matching/mirroring of and accommodation with emoji can be used as a flirting strategy (Nexo & Strandell, 2020), and that emoji use in first-date online chats is indeed expected to adhere to social norms of reciprocity/mirroring ('non-verbal attunement'), but also norms of appropriate intensity (Stein, 2023). Other studies zoomed in on specific emoji and showed that winking emoji serve as ambiguous flirtation cues in Tinder chats (Gibson, 2024), that certain emoji can be strategically used to signal relational intentions (friendship: 😊, romantic intent: 😍 or ❤️, sexual intent: 😏 or 🍆) (Rodrigues et al., 2022), and that certain emoji (🍆🍆) are commonly interpreted in a sexually suggestive manner, although interpretations rely heavily on context (Thomson et al., 2018; Weissman, 2019).

The present study adopts a methodologically novel approach to examining emoji and emoticon use in online dating: while previous research on emoji in online dating was survey-based (Gesselman et al., 2019; Thomson et al., 2018), used focus groups (Nexo & Strandell, 2020), conversation analysis (Gibson, 2024), or had an experimental design (Rodrigues et al., 2022; Stein, 2023; Weissman, 2019), this paper reports on the first quantitative corpus analysis of emoji and emoticons in online dating. We take a much broader look by studying all emoji and

emoticons, so not only those used to euphemistically refer to body parts (e.g., 🍌 🍑 🍆 🍈 🍉) or other sexual innuendos (e.g., 🍷 🍸 🍹 🍺 🍻).

Moreover, although much prior research in the context of online dating focuses on self-presentation on dating profiles (see., e.g., Van der Zanden et al., 2021, 2022, 2024), such research has not taken into account emoji. In contrast, research in the context of online dating that does focus on emoji has hitherto zoomed in on chat communication (Gibson, 2024, Nexo & Strandell, 2020: Tinder chats; Rodrigues et al., 2022: Facebook and Tinder chats; Gesselman et al., 2019, Stein, 2023, Weissman, 2019: text messages), with the exception of Nexo and Strandell (2020) who included a self-presentation task using emoji in their study, but this involved non-naturalistic data collected under experimental conditions. We combine these two strands of research into computer-mediated dating – (a) research into online dating profiles and (b) research into emoji in online dating chats – by analysing emoji and emoticon use in chats *and* profiles.

This study aims to present a comprehensive and comparative picture of online daters’ actual emoji and emoticon use in dating profiles and chats, with data collected from real-world settings. We quantitatively classified online daters’ usage into different categories (which emoji are used?) and different functions (why are those emoji and emoticons used?) and qualitatively explored how these graphicons are employed with such functions in online dating.

## 2. Method

### 2.1 Data sample

The sample used for this corpus study consists of 125 dating chats<sup>1</sup> (of 35 individual respondents) and 79 dating profile texts. All profiles and chats contained at least one emoji or emoticon, with a total of 258 emoji and 23 emoticons in the profiles (on a total of 3,149 tokens) and 393 emoji and 242 emoticons in the chats (on a total of 17,735 tokens): emoji and emoticons thus comprised 8.9% of all tokens in the profiles versus 3.6% in the chats. The data were voluntarily submitted by users of different online dating platforms in the Netherlands. Most profiles and chats were from the popular dating app Tinder, but many other platforms were represented, with more variation in the profile than the chat submissions. More detailed information on the corpus composition can be found in Table 1 on the right.

The majority of profiles (89.9%) and chats (68.0%) were in Dutch; the rest was in English. More data of women were included in our corpus (profiles: 72.2%; chats: 86.4%) than of men. Respondents’ mean age was 30.7 years ( $SD = 9.8$ ) in the profile sample and 21.3 years ( $SD = 3.78$ ) in the chat sample.

Platforms	Profiles	Chats
Tinder	30	75
Bumble	24	39
Breeze	7	
Happn	2	6
Grindr	1	2
Feeld	1	2
Badoo	3	
Lexa	3	
Funky Fish	3	
<b>Total</b>	<b>79</b>	<b>125</b>

Table 1: Corpus composition.

*Note:* Other platforms from which dating profiles were submitted, with a frequency below 3, include Inner Circle, Parship, HER, and SDC, as well as Hinge with a frequency of 1 for dating chats.

### 2.2 Data coding

The emoji and emoticons in the profile texts and chats were coded for categories and functions. For the categories, we followed the categorization of the online emoji reference site *Emojipedia*: Smileys (including all faces), People, Animals & nature, Food & drinks, Activities, Travel & places, Objects, Symbols, and Flags. Nearly all emoticons can be categorized as Smileys. For the functions, we compiled a codebook distinguishing nine functions (building on Verheijen & Mauro, 2025):

- (1) **Emotion addition**, when the emoji/emoticon expresses an emotion that is not expressed in the text (e.g., ‘definitely not 🤔’, ‘no problem :p’);
- (2) **Emotion reinforcement**, when the emoji/emoticon expresses an emotion that is also expressed in the text (e.g., ‘hahaha 😂’, ‘just kidding ;)’);
- (3) **Gesture**, when the emoji/emoticon represents a movement of hand(s) or arm(s) (e.g., ‘bye 👋’, ‘hello 🤖’, ‘👉’);
- (4) **Visualisation of keyword**, when the emoji/emoticon visualises a particular word in the text (e.g., ‘I like bowling 🎳’, ‘love you <3’);
- (5) **Visualisation of general content**, when the emoji/emoticon visualises and disambiguates general content in the text (e.g., ‘I like sports 🏀 🏈 🏊 🏹’);
- (6) **Figurative**, when the emoji/emoticon is used in a non-literal/metaphorical way (e.g., 🚀 to represent success, 💀 to mean dying from laughter or frustration);
- (7) **Substitution, embedded**, when the emoji/emoticon is used to replace (a) word(s) within a sentence or clause, which would be grammatically incomplete without it (e.g., ‘I like 🐶 🐱’);
- (8) **Stand-alone**, when the emoji/emoticon is not part of a textual message but used on its own as a separate message or utterance (e.g., ‘👉’, ‘🤔’);
- (9) **Miscellaneous**, when the emoji/emoticon has another function not identified above.

<sup>1</sup> Note that all these chats took place with instant messaging

software within dating platforms, not on external social media.

### 3. Expectations

For the dating profiles, which are used for presenting oneself to potential dating partners and managing their impressions, we expected to find more substitutions of text, visualisations, and stand-alone emoji [*functions*] to playfully give colour to one's profile and to manage the word limit of dating app profiles, as well as more non-smileys [*categories*] to express daters' hobbies, interests, affiliations, favourite foods, pets, or family composition. For the dating chats, which are used for socially connecting with potential partners, we expected to find more smileys [*category*] and more emoji and emoticons that express emotions [*function*], to compensate for the lack of non-verbal cues such as facial expressions in written computer-mediated communication.

### 4. Results

Two MANCOVAs, controlling for text length, showed main effects of communication context (profiles vs. chats) on emoji/emoticon categories,  $F(9, 193) = 4.10, p < .001, \eta^2 = .160$ , and on functions,  $F(6, 196) = 6.52, p < .001, \eta^2 = .081$ . In both models, text length was a significant covariate (categories:  $F(9, 193) = 24.34, p < .001$ ; functions:  $F(6, 196) = 33.68, p < .001$ ). The following sections report the results of univariate ANCOVAs.

#### 4.1 Categories

70.9% of all 916 coded emoji and emoticons were Smileys. From these 649 smileys, 89.4% appeared in the chats. Five out of nine categories (Activities, Travel & places, Flags, Food & drinks, Objects) were used not at all or hardly at all in the chats. Animals & nature were also used more often in the profiles. Only Symbols (e.g., ❤️ 🕒) and People-related emoji (e.g., 🤔 🧐) occurred about equally often – in terms of absolute frequencies – in the profile sample and the chat sample. Separate ANCOVAs, controlling for text length, showed that smileys occurred significantly more frequently in the chats ( $F(1, 201) = 5.08, p = .025, \eta^2 = .025$ ), whereas all non-smiley categories except for symbols occurred more frequently in the profiles (all  $p$ 's  $< .018, \eta^2$ 's  $> .027$ ).

A single emoji or emoticon could be coded as having more than one function: they can serve several functions simultaneously, for example emotion addition *and* making a gesture (e.g., 'will do 🤔'), visualisation of keyword *and* emotion reinforcement (e.g., 'I love you ❤️'), or making a gesture *and* stand-alone (e.g., '🤔').

Prior to conducting statistical analyses, the two emotion-related functions (emotion addition, emotion reinforcement) and the two visualisation-related functions (visualisation of keyword, visualisation of general content) were both merged into one category, emotion and visualisation, respectively. The residual miscellaneous function ( $n = 11$ ) was not considered for further analyses. This reduced the number of functions from nine to six.

The corpus was coded in its entirety by a student assistant, who was trained in multiple practice sessions how to apply the codebook. In practicing with the codebook, she discussed difficult cases with the authors to finetune it. To make sure that the coding was reliable, 16.5% of the emoji and emoticons in the profiles and 15.8% of the emoji and emoticons in the chats were double-coded (by one of the authors) on their categories and functions. All four coding agreements were 85% or higher, with Cohen's kappas of .78 or higher. Cases of disagreement in coding were discussed by the researchers before deciding upon a final coding. Tables 2 and 3 below show the frequencies and relative proportions of the emoji categories and functions in both communication contexts.

#### 2.3 Data analysis

To examine if the communication context (dating profile vs. dating chat) determined the use of emoji/emoticons in certain categories and functions, two MANCOVAs (Pillai's trace, because the assumption of homogeneity of variance was violated) were run with the frequencies of the nine categories and the six functions as dependent variables, and with text length (operationalized as the total number of tokens) as a covariate. We included this covariate because the chats ( $M = 143.02, SD = 133.28$ ) were generally longer than the profile texts ( $M = 39.9, SD = 52.8$ ), although there was great variation in text length within both communication contexts. Note that we should be cautious with interpreting results for categories and functions with low frequencies.










Categories	Smileys 	People 	Animals & nature 	Food & drinks 	Activities 	Travel & places 	Objects 	Symbols 	Flags 	Total
Profiles	69 (24.6%)	22 (7.8%)	50 (17.8%)	32 (11.4%)	28 (10.0%)	17 (6.1%)	37 (13.2%)	15 (5.3%)	11 (3.9%)	281
Chats	580 (91.3%)	24 (3.8%)	11 (1.7%)	2 (0.3%)	0 (0%)	0 (0%)	3 (0.5%)	15 (2.4%)	0 (0%)	635
Total	649	46	61	34	28	17	40	30	11	916

Table 2: Absolute frequencies and percentages of emoji and emoticon categories, per communication context.

Functions	Emotion (addition + reinforcement)	Gesture	Visualisation (of keyword + general content)	Figurative	Substitution, embedded	Stand-alone	Miscellaneous	Total
Profiles	57 (20.2%)	5 (1.8%)	76 (27.0%)	9 (3.2%)	10 (3.5%)	124 (44.0%)	1 (0.4%)	282
Chats	591 (85.9%)	13 (1.9%)	10 (1.5%)	10 (1.5%)	1 (0.2%)	54 (7.8%)	9 (1.3%)	688
Total	648	18	86	19	11	178	10	970

Table 3: Absolute frequencies and percentages of emoji and emoticon functions, per communication context.



## 4.2 Functions

### 4.2.1. Emotion

Expressing emotion was the most common function of emoji and emoticons in our corpus. Emotions were predominantly expressed using Smileys and sometimes with People. Table 3 shows that using emoji and emoticons to express emotions was much more frequent in chats than profiles,  $F(1, 201) = 7.53, p = .007, \eta^2 = .036$ . Of the 648 times emoji or emoticons were used to add or reinforce emotions, 91.2% ( $n = 591$ ) occurred in chats versus 8.8% ( $n = 57$ ) in profiles. Emoticons are – as the name indicates (*emotion + icon*) – generally used to convey emotions, so it is not surprising that emoticons appeared more frequently in chats ( $n = 242$ ) than in profiles ( $n = 23$ ).

### 4.2.2. Gesture

Emoji barely represented gestures (e.g., 🤔 🙌 🙏) in our corpus: with a total frequency of 18 these made up only 1.86% of all emoji and emoticons used. They occurred five times in profiles and 13 times in chats, which was a non-significant difference,  $F(1, 201) = 0.064, p = .800$ . Some Smileys, but mostly People-related emoji were used to make gestures.

### 4.2.3. Visualisation

Of all 86 emoji used to visualise keywords or general content, the majority appeared in profiles ( $n = 76$ ; 88.4%), with only a small portion in chats ( $n = 10$ ; 11.6%). Most emoji with the function of visualisation in profiles were non-smileys, which visualised Animals & nature, Foods & drinks, Activities, Travel & places, and Objects. For example, the dating profiles included emoji to visualise hobbies and other personal preferences (e.g., 📺 🐶 🍷 🏠 🎮). Emoji and emoticons were used to visualise text significantly more frequently in profiles than in chats,  $F(1, 201) = 17.37, p < .001, \eta^2 = .075$ . If emoji were used in chats for visualisation, these were mostly Smileys or People (e.g., ‘almost holiday 😊 😊’).

### 4.2.4. Figurative

Both profiles and chats showed very little figurative use of emoji, in total only 19 times. Frequencies were low and did not differ significantly in the communication contexts,  $F(1, 201) = 0.81, p = .368$ . Figuratively used emoji in profiles ( $n = 9$ ) mostly represented affiliation with the queer community (🏳️ or 🌈) or indicated being ‘on fire’ (🔥), which were expressed with emoji from the categories of Flags and Animals & nature. Our sample contained only one instance of the eggplant emoji (🍆) to refer to the male reproductive organ. In the chats, 6 out of 10 emoji used figuratively were skulls representing dying from laughter or frustration – all used by the same respondent.

### 4.2.5. Substitution, embedded

Only 1.13% of all emoji were embedded in the text in such a way that they substituted one or more words (e.g., ‘Ready

for 🏠 🌲 🧑’ and ‘sales 🐱’). If used at all, embedded emoji often corresponded with the categories of Animals & nature and Travel & places. The frequencies tentatively suggest that emoji are more often used to replace (a) word(s) within a sentence or clause in profiles ( $n = 10$ ) than in chats ( $n = 1$ ). This difference was significant,  $F(1, 201) = 6.44, p = .012, \eta^2 = .031$ , but should be cautiously interpreted given the low frequencies.

### 4.2.6. Stand-alone

Stand-alone emoji were the second most common in our corpus. Emoji were more often used on their own to replace an entire sentence in profiles ( $n = 124$ ; 69.7%) than in chats ( $n = 54$ ; 30.3%). This observation was confirmed to be statistically significant,  $F(1, 201) = 8.46, p = .004, \eta^2 = .040$ . Stand-alone emoji in profiles were mainly non-smileys for self-presentation purposes, sometimes used in a sequence such as ‘😊 🧑 🌲 🎮’ or ‘🍷 🏠 🎮 🍷 🏠 🎮 🍷 🏠 🎮’. In chats, stand-alone emoji often simultaneously fulfilled the function of adding emotion, for example messages that only consisted of stand-alone smileys like ‘😂 😂 😂’.

## 4.3 Demographic analyses

We conducted additional analyses to tentatively explore the effects of demographic factors on emoji and emoticon use in online dating. Independent samples t-tests with gender revealed that, both communicative contexts taken together, the female respondents ( $M = 9.34, SD = 16.02$ ) used significantly more emoji/emoticons than the male respondents ( $M = 4.19, SD = 3.89$ ) in our corpus,  $t(107.24) = 2.73, p = .004$ , Cohen’s  $d = .36$ . In addition, women used the function of expressing (i.e., adding or reinforcing) emotions ( $M = 6.88, SD = 14.51$ ) significantly more frequently than men ( $M = 1.88, SD = 3.15$ ),  $t(104.93) = 2.97, p = .002$ , Cohen’s  $d = .39$ .<sup>2</sup>

## 5. Discussion

This study is the first large-scale corpus analysis of online daters’ use of emoji and emoticons on dating platforms, a topic on which prior literature is scarce. We have focused on emoji and emoticons in both online dating profiles and online dating chats, with a corpus of naturalistic data collected in the Netherlands, comparing and contrasting usage in terms of categories and functions.

Our results suggest that on **dating profiles**, emoji are used in a thoughtful, deliberate, playful, and creative way for the purposes of self-presentation and impression management within the constraints of online dating platforms. Such constraints include the limits which many dating apps enforce for users’ bios (e.g., maximum 500 characters on Tinder and 300 on Bumble). This corresponds with our analyses of categories and functions. In terms of categories, dating profiles showed a greater use of **People**, **Animals & nature**, **Food & drinks**, **Activities**, **Travel & places**, **Objects**, and **Flags** as compared to dating chats.

<sup>2</sup> Respondents who indicated to be non-binary or of a different gender unfortunately had to be excluded from these analyses for

statistical reasons.

These emoji were used by respondents to present their family composition, pets or favourite animals, favourite foods or beverages, hobbies, interests, values, and affiliations in a concise and visually striking way. In terms of functions, self-presentation with emoji resulted in a greater use of emoji for **visualising** keywords or general content and, possibly for purposes of conciseness, in a greater use of **stand-alone** emoji replacing text and – although with very low frequencies – **embedded** emoji substituting text on profiles than in chats.

As for **dating chats**, our results show that online daters mainly use emoji and emoticons to express emotions in written CMC, compensating for a lack of non-verbal cues such as facial expressions and body language. They do so through a greater use of the category of **Smileys** and the function of adding or reinforcing **emotions**. This aligns with Gesselman et al.'s (2019) findings that one of the key motives for emoji use in dating messages is to express one's feelings. Communication in dating chats, a form of instant messaging, is near-synchronous and aims to resemble face-to-face conversations, in order to increase social connectedness with potential dating partners. Smiley-face and other facial emoji and emoticons add emotions to chats in a non-verbal, visual manner and have become essential to express feelings towards interlocutors, including potential dating partners. These results are in line with prior research (e.g., Chiang & Gomez-Zara, 2024; Novak et al., 2015; Verheijen & Mauro, 2025) on emoji use in instant messaging and social media posts, which point out that emoji often serve as paralinguistic cues or 'emotional markers' to express sentiment.

Our study shows very little figurative or sexual usage of emoji. Where Weissman (2019)'s experiment with American participants focused solely on such emoji and Thomson et al. (2018)'s survey indicated that emoji play a significant role in sexually suggestive messages in Canada, our corpus analysis of dating profiles and chats by Dutch online daters suggest that such messages are in fact not that widespread in computer-mediated dating, at least not in our sample.

Furthermore, exploratory gender analyses reveal that women not only use more emoji and emoticons than men in online dating, but also use them more to convey emotions. These findings are in line with previous corpus research about gendered emoji use in other CMC contexts such as online chat rooms and instant messaging, which have consistently shown that women use more emoji than men (Chen et al., 2018; Fullwood et al., 2013; Koch et al., 2022).

### 5.1 Limitations and suggestions for research

A limitation of this study is that a self-selection bias may have occurred, as online daters voluntarily donated their profiles and chats to our corpus, and could have been selective in what to share with the researchers. This may (in part) underlie the very rare occurrence of emoji used in a figurative, sexually suggestive manner in our corpus.

Another drawback of our corpus is that emoji and emoticon frequencies of some categories and functions

were low. Future research should aim for an even larger corpus of a more demographically varied sample, so that, for instance, gender differences can be more systematically investigated and age differences in emoji and emoticon use in online dating can also be studied.

Further research could follow up on the current study's findings by conducting surveys (cf. Gesselman et al., 2019), focus groups (cf. Nexø & Strandell, 2020), or interviews to learn more about people's motivations for using (specific) emoji and emoticons in online dating communication contexts, especially those in dating profiles: to what extent are users driven by the imposition of character constraints and to what extent are they trying to achieve more playfulness or express their personality?

Finally, future experimental research could study how specific categories of emoji, or originality and creativity in emoji use, in dating profiles affect impression formation. Dating profile texts have to date received little attention in online dating research on emoji. We suggest that follow-up studies explore how emoji interact with other visual cues and textual cues in both dating profiles and dating chats, to deepen our understanding of the interaction between multimodal elements in computer-mediated dating.

## 6. Conclusion

To conclude, this corpus study has provided a more comprehensive overview of which emoji and emoticon categories are used by online daters and with what functions in different online dating contexts. Building on Verheijen and Mauro (2025), we have presented a novel classification of emoji functions, which can be applied to analyses of emoji and emoticon use in other online genres. The present study has filled a research gap by conducting a corpus analysis of emoji and emoticons in online dating, and by including profiles besides chat messages. Our findings highlight that emoji and emoticon use in dating profiles and chats is markedly different, with at the core of these differences the communicative aims of self-presentation and impression management, on the one hand, and emotion expression and social connectedness, on the other hand. Our study has shown that online daters are adept at using graphicons to suit their communication purposes and skilfully take advantage of visual means to digitally connect with potential dating partners. By zooming in on emoji and emoticons in online dating communication through the lens of corpus analysis, this study offers fresh perspectives on how current affordances of CMC are utilized to form relationships online.

## 7. Acknowledgements

We would like to thank our student assistant Nienke Gelderland for meticulously coding our corpus. Our thanks also go to Tila Pronk (Tilburg University) and Lenneke Lichtenberg for their help with the collection of the dating chat sample.

## 8. References

Bai, Q., Dan, Q., Mu, Z., and Yang, M. (2019). A systematic review of emoji: Current research and future

- perspectives. *Frontiers in Psychology*, 10(2221), pp. 1–16.
- Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., and Liu, X. (2018). Through a gender lens: Learning usage patterns of emojis from large-scale android users. *Proceedings of the 27th International World Wide Web Conference* (pp. 763–772).
- Chiang, C., and Gomez-Zara, D. (2024). The evolution of emojis for sharing emotions: A systematic review of the HCI Literature. arXiv:2409.17322.
- Cloudwards (2025, April 25). 35+ Online dating statistics in 2025: Trends, facts and key insights. <https://www.cloudwards.net/online-dating-statistics>
- Fullwood, C., Orchard, L.J., and Floyd, S.A. (2013). Emoticon convergence in Internet chat rooms. *Social Semiotics*, 23(5), pp. 648–662.
- Gesselman, A.N., Ta, V.P., and Garcia, J.R. (2019). Worth a thousand interpersonal words: Emoji as affective signals for relationship-oriented digital communication. *PloS One*, 14(8), e0221297.
- Gibson, W. (2024). Flirting and winking in Tinder chats: Emoji, ambiguity, and sequential actions. *Internet Pragmatics*, 7(2), pp. 249–271.
- Herring, S.C., and Dainas, A.R. (2017). “Nice picture comment!” Graphicons in Facebook comment threads. *Proceedings of the Fiftieth Hawaii International Conference on System Sciences*. IEEE, pp. 2185–2194.
- Koch, T.K., Romero, P., and Stachl, C. (2022). Age and gender in language, emoji, and emoticon usage in instant messages. *Computers in Human Behavior*, 126, 106990.
- Konrad, A., Herring, S.C., and Choi, D. (2020). Sticker and emoji use in Facebook Messenger: Implications for graphicon change. *Journal of Computer-Mediated Communication*, 25(3), pp. 217–235.
- Manganari, E.E. (2021). Emoji use in computer-mediated communication. *International Technology Management Review*, 10(1), pp. 1–11.
- Nexø, L.A., and Strandell, J. (2020). Testing, filtering, and insinuating: Matching and attunement of emoji use patterns as non-verbal flirting in online dating. *Poetics*, 83, 101477.
- Novak, P.K., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PloS One*, 10(12), e0144296.
- Rodrigues, D.L., Cavaleiro, B.P., and Prada, M. (2022). Emoji as icebreakers? Emoji can signal distinct intentions in first time online interactions. *Telematics and Informatics*, 69, 101783.
- Setyawan, H., and Musthafa, B. (2024). Contemporary issues in linguistics: A systematic literature review on emoji and emoticon. *Elsya: Journal of English Language Studies*, 6(3), pp. 315–326.
- Statista (2025, January 31). Online dating usage by brand in the Netherlands as of December 2024. <https://www.statista.com/forecasts/1226871/online-dating-usage-by-brand-in-the-netherlands>
- Stein, J.P. (2023). Smile back at me, but only once: Social norms of appropriate nonverbal intensity and reciprocity apply to emoji use. *Journal of Nonverbal Behavior*, 47(2), pp. 245–266.
- Tang, Y., and Hew, K.F. (2019). Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International Journal of Communication*, 13, pp. 2457–2483.
- Thomson, S., Kluffinger, E., and Wentland, J. (2018). Are you fluent in sexual emoji? 😏: Exploring the use of emoji in romantic and sexual contexts. *Canadian Journal of Human Sexuality*, 27(3), pp. 226–234.
- Van der Zanden, T., Mos, M.B., Schouten, A.P., and Krahmer, E.J. (2021). What people look at in multimodal online dating profiles: How pictorial and textual cues affect impression formation. *Communication Research*, 49(6), pp. 863–890.
- Van der Zanden, T., Schouten, A.P., Mos, M.B., and Krahmer, E.J. (2022). Originality in online dating profile texts: How does perceived originality affect impression formation and what makes a text original? *Plos One*, 17(10), e0274860.
- Van der Zanden, T., and Schouten, A.P. (2024). Creativity, expectancy violations, and impression formation: Effects of novelty and appropriateness in online dating profile texts. *Media Psychology*, 27(6), pp. 842–868.
- Verheijen, L., and Mauro, T. (2025). Emoji use by children and adults: An exploratory corpus study. *Research in Corpus Linguistics*, 13(1), pp. 57–85.
- Weissman, B. (2019). Peaches and eggplants or... something else? The role of context in emoji interpretations. *Proceedings of the Linguistic Society of America*, 4(29), pp. 1–6.

# “Tinder is overrated”: Neoliberal Affective Economies in an Italian Incel Forum

Selenia Anastasi<sup>1</sup>, Maria Natasha Fragalà<sup>2</sup>

<sup>1</sup>University of Rome “La Sapienza”, Department of Communication and Social Science Research

<sup>2</sup>University of Catania, Department of Humanities

<sup>1</sup>selenia.anastasi@uniroma1.it, <sup>2</sup>natasha.fragala@gmail.com

## Abstract

This paper investigates how dating apps are framed within the Italian incel community through a mixed-method analysis of posts from Forum dei Brutti. The study draws on a subcorpus of 2.4 million tokens, extracted from the Incel Data Archive -IDA corpus (Anastasi et al., 2025), and combines keyword analysis via SketchEngine with close qualitative coding of 200 posts (14,514 tokens). Quantitative findings confirm the centrality of Tinder and related terms, while qualitative results show that users interpret dating app dynamics, such as match scarcity and superficial responses, as evidence of systemic exclusion from a *sexual market* governed by women and hegemonic masculinities. These perceptions are shaped by neoliberal and algorithmic logics that commodify intimacy and reinforce the dominance of visibility. A recurring double standard emerges: while users critique women’s sexual selectivity, they apply similar aesthetic criteria themselves. Thus, dating apps appear not as neutral tools, but as infrastructures that intensify and support the redpill ideology and associated misogynistic stances.

**Keywords:** Italian Incels, Forum dei Brutti, Peripheral Manosphere, Dating Apps.

## 1. Introduction<sup>1</sup>

In recent years, *masculinity crisis* has become a central theme in both public and academic debates, prompting a redefinition of traditional gender roles. Within this context, new hybrid forms of masculinities have emerged (Ging, 2019), often developing in digital environments and marked by a sense of victimization in relation to a system perceived as dominated by women. Among these groups, *incel* (involuntary celibate) communities stand out – transnational collectives composed predominantly of white, heterosexual men who interpret women’s emancipation and popularised forms of feminism (Banet-Weiser et al., 2020) as oppressive to men.

Within the so-called Incelsphere, a common narrative suggests that women’s increasing decisional power in the “sexual market” has excluded a segment of men from access to intimate relationships. In this scenario, dating apps, such as Tinder, play a key role by reinforcing hierarchical and deterministic views of sexuality, increasing frustration and rage among non-hegemonic masculinities (Connell, 1987). Yet, as (Sparks et al., 2022) point out, the relationship between the incelsphere and dating apps has remained surprisingly underexplored, particularly outside of English-speaking contexts.

This study aims to address this gap by analyzing a corpus of posts collected from the main Italian incel forum, *Il Forum dei Brutti* (FDB). Founded in 2009 as a self-help space for individuals with non-normative aesthetics or disabilities, the forum has undergone a process of radicalization, adopting the misogynistic rhetoric of the redpill ideology (Anastasi et al., 2025). Today, the FDB receives over 35,000 monthly visits and serves as a notable example of a peripheral manosphere community (Scarcelli, 2021).

Based on a qualitative analysis of 212 posts, the study finds that in 67% of cases, dating apps (and especially Tinder) are described negatively by users, and in 66% they are perceived as ineffective or even harmful in interpersonal relationships. These results are in line with relevant international literature dedicated to the core groups of the Incelsphere. Indeed, our results confirm the circulation of tropes and narratives about dating apps as key instruments of male sexual exclusion while reinforcing a misogynistic narrative based on the idea of women as powerful, deceiving, and superficial.

### 1.1. Structural Misogyny and Neoliberal Logics

Incel ideology is grounded in a rigidly hierarchical understanding of heterosexual desire and relationships, in which women’s sexual emancipation is framed as a principal cause of male marginalisation. According to this perspective, the Sexual Market Value (SMV) – that is, the value assigned to individuals within the sexual marketplace – is governed by dynamics that systematically favour women, who are perceived to select partners exclusively on the basis of aesthetic, economic, and social capital (O’Donnell, 2022). This results in a binary and economic interpretation of sexuality, where average men (normies or betas) are structurally excluded in favor of a narrow elite of alpha males (Chads), fueling a discourse of victimhood and resentment.

This interpretation draws heavily on neoliberal market logics, which assume that most individuals are “irrational” and therefore require the discipline of the market to make optimal choices (Srnicek, 2017). Within this framework, competition, meritocracy, and individual responsibility are promoted as the dominant social values. In incel discourse, these principles are applied to the sexual domain: attractiveness becomes a form of capital, relationships are framed as outcomes of market competition, and personal failure in romantic or sexual life is interpreted either as a deficit in value (e.g. lacking looks, money, or status) or as proof of systemic injustice (Scarcelli, 2021).

<sup>1</sup>The article is grounded in Fragalà’s dissertation, in which she contributed substantially to the development of the content analysis and the final interpretation of the results. Anastasi is responsible for the composition of the article, the literature review, and the methodology related to corpus-based analysis.



This ideological structure is further reinforced by a conspiratorial worldview that identifies women and feminist agendas as directly responsible for incel exclusion, alongside a lexicon saturated with dehumanising language. As (Baele et al., 2021) argue, incel discourse exhibits many hallmarks of extremist rhetoric: essentialist social categorisations, a polarising ingroup/outgroup logic, and conspiratorial narratives in which incels perceive themselves as victims of systemic injustice. However, (Heritage, 2023) notes a degree of complexity in evaluative patterns, observing that in-group members may also be negatively portrayed by other users. Moreover, the use of metaphors – particularly in reference to women and marginalised groups (notably BIPOC) – frequently draws on animalistic, objectifying, and food-related imagery (e.g. *foid*, *roastie*, *vermin*), contributing to the normalisation and perpetuation of symbolic violence.

Similar ideological configurations emerge within Italian forums, though these contexts have received less attention in terms of discourse analysis. Dordoni and Magaraggia (Dordoni and Magaraggia, 2021) identify two key dimensions of Italian incel discourses: 1. An obsessive focus on physical appearance and a narrative of self-victimisation; 2. The reification of women, which legitimises violent imaginaries and hostile discursive practices. Within this framework, incel rhetoric extends beyond individual frustration and becomes a vehicle for advocating the restoration of patriarchal social orders, in which both explicit and latent forms of violence are positioned as necessary instruments to reinstate the “natural order” of gender relations.

## 1.2. Aesthetic Hierarchies and Biological Determinism

Within the Italian FDB, the hierarchical logics that dominate the sexual market are rigidly anchored to aesthetics, which is framed as an objective and immutable measure of individual worth. Terms such as *brutto* (ugly) and *brutto vero* (truly ugly) serve both identity and classificatory functions, demarcating a status of exclusion grounded in biological determinism – physical traits such as height, facial symmetry, or strong jawlines are typical markers of male sexual value. Moreover, aesthetics delineate the symbolic boundaries of the group, generating an internal value scale that ranges from *brutto vero* to *figo* (hot or handsome), the Italian equivalent of the Chad trope found in international-mainstream jargon.

This structure is informed by *lookism*, an instance of the redpill belief system asserting that beauty is measurable on numerical scales and should determine reproductive worth. Such views often draw on eugenic logics, supporting the idea that only attractive individuals should reproduce to avoid perpetuating the suffering condition associated with unattractiveness. Within this deterministic and nihilistic framework, users engage in collective evaluation of physical appearance and adopt strategies aimed at self-improvement, including *looksmaxxing* (cosmetic enhancement) and *gymmaxxing* (intensive physical training). These practices are regarded as the only viable routes to sexual recognition in what is otherwise seen as an inaccessible space. Moreover, these strategies are frequently accom-

panied by calls for the reinstatement of patriarchal control mechanisms, such as enforced monogamy, reflecting the broader restorative logic that underpins incel ideology (Sparks et al., 2024).

## 1.3. Incels and Dating apps

Recent scholarship has highlighted the central role that dating apps play in both the lived experience and identity construction of incel users. In a comparative quantitative study involving 38 self-identified incel men and 107 non-incel men, (Sparks et al., 2024) found that incels engage more frequently with dating apps – for example, by swiping right at significantly higher rates – but receive far fewer matches. These patterns are accompanied by elevated levels of anxiety, depression, and low self-esteem. The same study, supported by a review of prior literature, confirms a recurring constellation of relational difficulties among incels, including heightened sensitivity to rejection, fear of being alone, and insecure attachment styles.

At the same time, qualitative research such as that conducted by (Preston et al., 2021), based on a corpus of over 9,000 comments from Incels.co, reveals how dating apps are discursively constructed by incel users as amplifiers of hypergamous dynamics and exclusionary logics. Digital platforms and their algorithms are perceived as systems that systematically favour the most physically attractive men (Chads) while increasing women’s selective power.

Furthermore, the literature on dating app algorithms highlights a structural lack of transparency. As noted by (Finkel et al., 2012) and, more recently, by (Paul and Ahmed, 2024), the criteria by which compatibility and visibility are determined remain opaque, allowing for processes that reinforce prevailing norms of desirability and deepen patterns of exclusion.

## 2. Data and samples

Our analysis draws on the Italian subcorpus of the Incel Data Archive (IDA) (?), compiled from data collected on incels.is and the Forum dei Brutti. In the initial phase, we tracked the frequency of references to four dating apps commonly used in Italy: *Badoo*, *Tinder*, *Meetic*, and *Lovoo* over the time range 2010–2023. The data revealed a shift in discursive centrality from Badoo to Tinder beginning in 2016, with a peak in overall attention occurring in 2019. Based on these findings, we extracted a dedicated 2019 subcorpus (FDBDATING-2019, counting 2.4 million tokens), which was explored through keyword and multi-word term extraction to identify the most frequently mentioned apps and examine how they are discursively framed.

To complement this quantitative phase, we conducted a close reading of 200 posts (14,514 tokens) to further investigate user attitudes towards dating platforms and their role in shaping perceptions of social exclusion, as described in the Methodological section.

## 3. Methodology

### 3.1. Quantitative analysis

The quantitative analysis was conducted on the FDBDATING-2019 subcorpus, comprising approximately

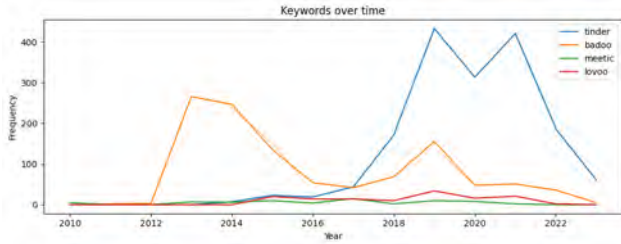


Figure 1: Keyword frequency related to dating apps in the Italian incel forum.

2.4 million tokens. Methodologically, the subcorpus was examined in two main phases. In the first phase, keywords and multiword terms were extracted using the SketchEngine platform (Kilgariff, 2009), with itTenTen20 (a large web-based corpus of Italian web language) serving as the reference corpus. The objective was to identify the discursive salience of major dating apps, as well as associated multiwords. Tables X and Y present, respectively, the list of keywords and multiwords, including their rankings, relative frequencies, and keyness scores calculated using the simple math method.

Table 1: Keywords related to dating apps

Rank	Item	RF Focus	RF Reference	Score
8	tinder	505	6410	145.2
67	badoo	114	2812	40.2
106	dating	113	11338	26.7
179	lovoo	43	447	18.2
450	meetic	25	3419	9.2

An inspection of the top 500 terms enabled the identification of key lexical items linked to dating platforms, with Tinder ranking among the top ten keywords in the entire subcorpus.

Table 2: Multiword terms related to dating apps

Rank	Item	RF Focus	RF Reference	Score
13	mercato sessuale	57	138	24.3
21	potere sessuale	45	826	18.5
22	vita sessuale	121	26606	18.0
26	app di incontri	42	1278	16.8
41	dating app	31	180	13.6
61	app di dating	26	326	11.5
120	sito di incontri	37	13864	8.3
126	match su tinder	17	14	8.0
414	dating apps	9	26	4.7

In corpus linguistics applied to Italian, analysing multiword constructions is essential because meaning is often distributed across word combinations rather than isolated tokens. Italian frequently relies on fixed or semi-fixed expressions (e.g., *mercato sessuale*, *app di incontri*), which carry specific socio-discursive functions that single words

alone may not reveal. In the second phase, keywords and multiword terms were analysed in close reading through concordance lines to explore their contextual usage and detect meaning in context.

### 3.2. Qualitative analysis

The qualitative analysis focused on how users within the community perceive and evaluate dating apps. It was based on a close reading of 200 posts (14,514 tokens), selected using the keywords identified during the quantitative phase. The analysis followed three stages: (1) classification of user sentiment (Positive, Negative, Neutral); (2) identification of core discursive themes (*Inutilità*, *Inganno*, *Costo*, *Affidabilità*); and (3) detection of prevailing emotional tones (*Frustrazione*, *Speranza*, *Scetticismo*).

A comment was coded as *Positive* when the user expressed satisfaction with the platform or a willingness to keep using it (i.e., *Not all of them – I managed to snag a few dates.*). Comments expressing disappointment, anger, or failure were classified as *Negative* (i.e., *I tried it a while back but didn't manage to score any dates.*). A *Neutral* comment was either descriptive or lacking strong value judgements (i.e., *Well, abroad it's a whole different story. I see that too with Cupid.*).

The theme *Reliability* was associated with posts describing successful outcomes following deliberate or mindful platform use (i.e., *Hinge is currently the best app as far as I'm concerned*). The theme *Futility* included posts that frame dating apps as ineffective – particularly for those not meeting conventional beauty standards (i.e., *Worse than in real life, on these apps the power dynamics are way too skewed in favor of the vagina-equipped.*). Posts tagged as *Deception* referred to fake profiles, bots, or manipulative strategies adopted by users or others to engage in conversation with potential partners (i.e., *Yes, all scams; there have even been official media reports.*). Finally, the theme *Cost* captured references to the *pay-to-win* logics of certain dating apps (i.e., *On Tinder you need to have Gold or Platinum, but you absolutely have to have BOOSTS*).

For what concerns the emotional labels, *Frustration* was applied to posts expressing anger due to repeated failure in gaining match (i.e., *I've deleted all the apps... they're a scam and women are a trap.*). The label *Hope* was applied when users acknowledged challenges but remained optimistic about future success through effort or strategic improvements (i.e., *On Tinder it's a whole different story: I snagged a month of Platinum on promo and the results were plentiful. Tinder Platinum definitely gets my seal of approval*). Finally, the label *Skepticism* applied to critical but emotionally detached comments (i.e., *In my opinion, all apps have one thing in common: they're a waste of time*).

All categories were developed through a bottom-up process grounded in close textual reading. Each post was manually annotated by the two authors to minimise subjective bias. Annotator agreement was discussed and negotiated case-by-case until full consensus was achieved. The use of mutually exclusive categories facilitated systematic coding and ensured analytical clarity.

## 4. Interpretation and findings

### 4.1. Key lexical items

Among the keywords extracted from the analysed sample, *Tinder* emerged as the most discussed dating app within the Italian incel forum. The analysis of multi-word expressions revealed recurring lexical patterns that enabled the identification of broader discursive formulas. Notably, multiword terms such as *mercato sessuale* and *potere sessuale* frame sexuality as a mechanism of social power and control. According to O'Neill, this economisation of sexuality is symptomatic of neoliberalism, which “disseminates the model of the market to all domains and activities” (O'Neill, 2018). In many comments, women – particularly those identified with feminists – are portrayed as dominant actors in both online and offline sexual economies. Examples include:

1. *Non ci sono soluzioni, specialmente ora che le np<sup>2</sup> sanno quanto valgono sul mercato sessuale.* [There are no solutions, especially now that the nps know exactly how much they're worth on the **sexual marketplace**].
2. *Le femministe non vogliono ammettere la loro differenza biologica che svantaggia l'uomo all'interno di un mercato sessuale libero.* [Feminists refuse to acknowledge their biological difference that puts men at a disadvantage within an open **sexual marketplace**].

Such excerpts illustrate the perceived asymmetry of power in sexual relations. Women's sexual agency is construed as a force that disadvantages less attractive men, excluding them from competition and relegating them to *involuntary celibacy*:

3. *Il potere sessuale delle donne è imbattibile.* [Women's sexual power is unbeatable].
4. *Le femministe lavorano per aumentare il potere sessuale della donna, se ci tolgono anche le prostitute, è proprio finita del tutto.* [Feminists are working to boost women's sexual power; if they even take away prostitutes, it's truly game over].

In this context, feminists are portrayed as the primary agents of the power imbalance, having “made women aware of their sexual power” and enabled them to wield it freely (see concordance 3). This framing sharply contrasts with institutional and academic reports that document the increasing vulnerability of women in online spaces (Jane, 2017). In the discourse of the forum, however, digital platforms, and dating apps in particular, are described as privileged arenas for the exercise of female power and agency. The economic framing also extends to the construction of the self within dating platforms. Profile photos are perceived as commodified assets. Indeed, in this digital market, women are seen as selecting partners based on

ephemeral criteria such as physical appearance. The constant exposure to idealised images and competitive self-presentation exacerbates anxiety around physical appearance and social worth. Within this highly competitive environment, expressions of misogyny are framed as legitimate responses to perceived sexual and social disenfranchisement.

Other recurring expressions, such as *app di dating* and *app di incontri*, are used interchangeably and signal spaces governed by distinct rules, in which market dynamics are believed to favour women. Moreover, a further metaphor of the economisation of sexuality is the reference to “free” male prostitution:

5. *Le app di incontri sono siti di prostituzione free a uso delle donne (o per uomini disposti all'ipogamia forte).* [Dating apps are free-prostitution sites at women's disposal (or for men willing to embrace strong hypogamy)].
6. *Quanta discriminazione nei confronti dell'uomo, tanto la donna ha i prostituti free, legali, dalle app di incontri.* [What discrimination against men – women get free, legal prostitutes via dating apps].

The centrality of *Tinder* is further underscored by the recurring construction *match su Tinder*, which becomes emblematic of an individual's sexual success. Concordance lines suggest that the number of matches functions in two primary ways: first, as a quantifiable external validation of one's perceived market value, reinforcing belief in redpill ideology; second, as a boundary-making device that distinguishes incels from *fake* incels users:

7. *9 match su Tinder in una settimana sono davvero buoni come feedback.* [9 Tinder matches in a week are really solid feedback.].
8. *Il solo fatto che qui ci siano utenti in LTR<sup>3</sup>, bellocci dichiarati che fanno svariati match su tinder, è una presa per il culo verso i veri incel.* [The mere fact that there are users in LTRs, self-proclaimed Chads pulling in multiple Tinder matches, is a total mockery of real incels.].

Another key finding that emerged from the concordance analysis concerns the cross-platform use of multiple digital environments for relational purposes, including *Badoo* and – significantly – *Instagram*. Although not formally designed as a dating app, *Instagram* is consistently assimilated into the same functional category, serving as a space in which sexual selection dynamics are activated. It thus operates as a parallel and complementary device within the logic of the sexual market, structured around visibility and display.

Many users report experiences spread across multiple platforms, comparing outcomes, responses, and interactions in order to measure their perceived sexual or social value. In this context, we also observe the emergence of low-selectivity engagement strategies – described by the users

<sup>2</sup>Acronym for *non-person*, Italian functional adaptation of the international disparaging jargon *femoid*.

<sup>3</sup>Acronym for *Long Term Relationships*.

themselves in almost mechanical or opportunistic terms. At the same time, the same users complain that the only interactions they receive on dating app are from women they consider aesthetically (or morally) undesirable.

This discursive pattern reveals a clear *double standard*: while women are criticised for selecting partners based on superficial aesthetic criteria, the users apply those very same criteria when evaluating the sexual worth of their female counterparts. As (Connell, 1987) suggests, this tension reflects the broader crisis of contemporary masculinities, wherein the desire to reassert male dominance coexists with an inability to recognise one’s own complicity in the exclusionary logics of the system.

The following comments illustrate these dynamics:

14. *Ho ottenuto un solo appuntamento iniziando io a mettere like/ scrivere su insta, badoo e lovoo e ne ho ottenuti 2.* [I scored just one date by starting to like/message on **Insta**, **Badoo**, and **Lovoo** – and I ended up getting two].
15. *Non ho instagram perché non ho nulla da condividere, mantengo il profilo solo su badoo e lovoo, magari qualcuna ci casca.* [I don’t have **Instagram** because I’ve got nothing to share; I keep profiles only on **Badoo** and **Lovoo**, maybe some girl will take the bait].

These statements confirm that users do not experience platforms solely as spaces for relational pursuit, but also as performative arenas in which they assess and negotiate their own position within a broader economy of attractiveness.

## 4.2. Content analysis

The content analysis uncovered discursive patterns through which incel users frame their experiences with dating apps. Communicative cues such as unmatched swipes, ignored messages, or superficial exchanges are interpreted not as isolated incidents, but as systemic confirmations of their perceived sexual and social undesirability. Sparks et al. (Sparks et al., 2024) describe this dynamic as a self-fulfilling cycle of failure, in which each negative experience reinforces users’ beliefs in their marginality. This narrative often neglects structural and emotional factors such as low self-esteem, deepening the sense of perceived injustice. Consequently, dating apps are not viewed as tools for relational empowerment but as *simulacra* of a society that reproduces and intensifies masculine hierarchies. Sentiment analysis (Figure 2) confirms this pattern, with 67% of comments expressing a negative stance. Indeed, users describe the platforms as “not made for people like us” and “only for the good-looking”, reinforcing a discourse of algorithmic exclusion.

Thematic coding (Figure 3) confirms this perception: the most recurrent label is *Futility*, accounting for 66% of the comments. Users describe dating platforms as fundamentally ineffective for *average* men, stating that “there’s no point trying when you already know you won’t get any matches”.

Finally, a close reading allowed us to identify the most prominent emotional tones, which include: Frustration,

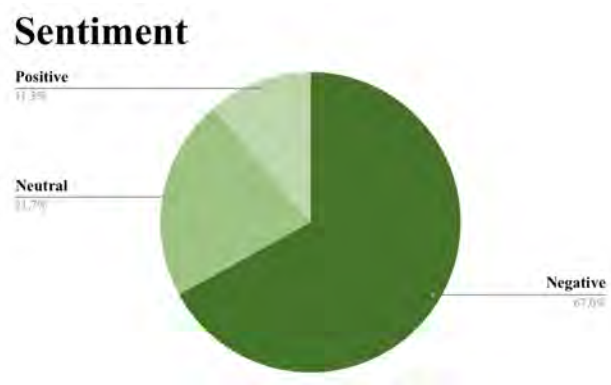


Figure 2: Percentage of Sentiments relating to dating apps

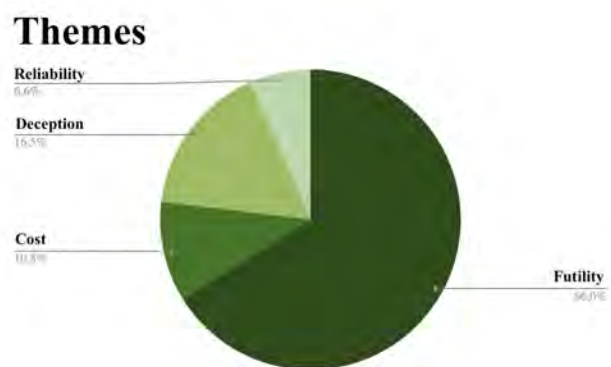


Figure 3: Percentage of Themes relating to dating apps

Skepticism, and Hope. Figure 4 illustrates the distribution of these categories in our sample.

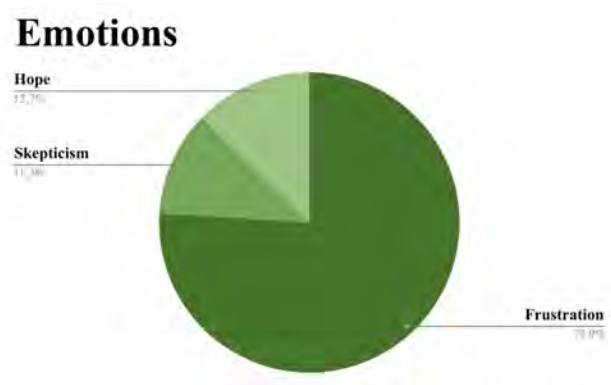


Figure 4: Percentage of Emotions relating to dating apps

According to our results, Frustration constitutes the predominant emotional response within the dataset, accounting for 75.9% of all coded emotions. Participants express this frustration in statements such as “every time I open the app, I feel even lonelier” and “it’s frustrating to see that even when you pay, nothing changes”.

The relationship between the three analytical dimensions is particularly salient: the overall negative sentiment (67%) stems from the perceived futility of dating platforms (66%), which in turn gives rise to overwhelming Frustration (75.9%). Themes further reinforce this interpretation: ref-

erences to Deception (16.5%) reflect a belief that dating apps offer misleading promises of romantic opportunity, while the theme of Cost (10.8%) underscores the disillusionment with having invested financially in systems perceived as rigged or futile.

Notably, a residual sense of Hope persists in 12.7% of comments, suggesting that some users retain expectations of success through dating apps. Paradoxically, this hope contributes to the ongoing cycle of disappointment that defines their engagement with dating technologies.

## 5. Conclusion and Future Works

This study explored how dating apps are discursively constructed within the Italian incel community, through a combined quantitative and qualitative analysis. Keyword extraction confirmed the centrality of Tinder in user discourse, where they function not merely as a tool but as a cultural symbol. The qualitative content analysis revealed that interactions, or their absence, are interpreted through the red-pill ideological framework, reinforcing narratives of hate towards women, systemic exclusion and self-victimisation. Theoretically, the findings intersect with literature on *neoliberal affective economies* (Illouz, 2007), revealing how incel users internalise competitive, market-oriented logics of desirability and worth. The platforms are perceived to algorithmically reward aesthetic and social capital while producing a self-fulfilling cycle of exclusion. At the same time, users replicate the very aesthetic hierarchies they condemn, demonstrating the contradictions at the heart of the contemporary masculinity crisis.

This study also contributes to critical perspectives on platform capitalism (Srnicek, 2017), positioning dating apps as Digitally-mediated infrastructures that commodify intimacy and reconfigure social relations through data-driven economies of attention and visibility. Far from being neutral, these platforms operate as algorithmic marketplaces that are not experienced as gateways to connection, but as systems that intensify normative hierarchies of gender.

### 5.1. References

- Anastasi, S., Fischer, T., Schneider, F., and Biemann, C. (2025). IDA – Incel Data Archive: A multimodal comparable corpus for exploring extremist dynamics in online interaction. In Louis Cotgrove, et al., editors, *Exploring digitally-mediated communication with corpora: Methods, analyses, and corpus construction*, pages 275–304. De Gruyter.
- Baele, S. J., Brace, L., and Coan, T. G. (2021). From “incel” to “saint”: Analyzing the violent worldview behind the 2018 toronto attack. *Terrorism and political violence*, 33(8):1667–1691.
- Banet-Weiser, S., Gill, R., and Rottenberg, C. (2020). Postfeminism, popular feminism and neoliberal feminism? sarah banet-weiser, rosalind gill and catherine rottenberg in conversation. *Feminist theory*, 21(1):3–24.
- Connell, R. (1987). Hegemonic masculinity and emphasized femininity. *Gender and power: Society, the person, and sexual politics*, 1(1):183–88.
- 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).
- Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., and Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public interest*, 13(1):3–66.
- Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and masculinities*, 22(4):638–657.
- Heritage, F. (2023). *Incels and Ideologies: Exploring How Incels Use Language to Construct Gender and Race*. Springer Nature.
- Illouz, E. (2007). *Cold intimacies: The making of emotional capitalism*. Polity.
- Jane, E. A. (2017). *Misogyny online: A short (and brutish) history*. SAGE.
- Kilgariff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, volume 6, Liverpool, July. University of Liverpool.
- O’Neill, R. (2018). *Seduction: Men, masculinity and mediated intimacy*. John Wiley & Sons.
- O’Donnell, J. (2022). Men’s rights activism and the manosphere. In *Gamergate and Anti-Feminism in the Digital Age*, pages 9–61. Springer.
- Paul, A. and Ahmed, S. (2024). Computed compatibility: examining user perceptions of ai and matchmaking algorithms. *Behaviour & Information Technology*, 43(5):1002–1015.
- Preston, K., Halpin, M., and Maguire, F. (2021). The black pill: New technology and the male supremacy of involuntarily celibate men. *Men and masculinities*, 24(5):823–841.
- Scarcelli, C. M. (2021). Manosphere periferiche. ragazzi, omosocialità e pratiche digitali. *AG About Gender-International Journal of Gender Studies*, 10(19).
- Sparks, B., Zidenberg, A. M., and Olver, M. E. (2022). Involuntary celibacy: A review of incel ideology and experiences with dating, rejection, and associated mental health and emotional sequelae. *Current psychiatry reports*, 24(12):731–740.
- Sparks, B., Zidenberg, A. M., and Olver, M. E. (2024). An exploratory study of incels’ dating app experiences, mental health, and relational well-being. *The Journal of Sex Research*, 61(7):1001–1012.
- Srnicek, N. (2017). *Platform capitalism*. John Wiley & Sons.

# Modelling the Interaction Space of Twitch: A Multimodal Framework for Corpus Structuring and Analysis

Ariane ROBERT

University of Salerno, DIPSUM

E-mail: arobert@unisa.it

## Abstract

This paper proposes a model of the interaction space of Twitch, based on and adapted from the CoMeRe framework for computer-mediated communication. The model addresses three core challenges of Twitch as a communicative environment: its multimodality, the asymmetric access to communicative modes based on participant roles, and the influence of non-verbal events such as gameplay or interface actions. We describe the different roles involved (broadcaster, viewer, moderator, bot) and map their access to six communicative modes. This model supports both theoretical reflection on role-based discourse production and the technical development of multimodal corpora. As a concrete application, we extended S. Coats's alignment tool by implementing role-based pseudonymisation and identifying areas for future extension, including the integration of non-verbal data layers. This work lays the foundation for corpus-based linguistic and discursive studies of live-streaming platforms and invites further inquiry into role-asymmetric and intermodal communication environments.

**Keywords:** Twitch, modelling, corpus, multimodality, streaming

## 1. Introduction

In recent years, new environments for computer-mediated communication (CMC) have emerged, marked by increasingly complex forms of interaction. Among them are streaming platforms such as Twitch<sup>1</sup> where video, audio, written text, video games, and other signs coexist and contribute to the construction of shared meaning. Although Twitch is relatively recent (2011), it has grown into a highly popular social network<sup>2</sup> that, while still centered on gaming, has expanded to other media forms, thereby reaching a wider audience and normalizing hybrid modes of communication.

There is a clear social interest in studying such platforms, to better understand their influence on social behaviors and associated excesses (Alklid, 2015; Nakandala et al., 2017; Robert & Pietrandrea, 2024), but also a linguistic interest in assessing how these complex communicative environments shape language use (Olejniczak, 2015; Recktenwald, 2017).

Twitch has been fairly well documented in the scientific literature in terms of the communicative modes it offers (Recktenwald, 2017). Sociological and ethnographic studies have also examined interactional dynamics on the platform to explore the formation of communities and their impacts on language practices (Ford et al., 2017; Hamilton et al., 2014; Olejniczak, 2015). However, the interaction space (IS) of Twitch has yet to be formally modelled. Developing such a model can serve both as a theoretical framework and as a technical tool. Modelling constitutes a crucial first step toward standardization and the creation of tools for processing multimodal data.

Steven Coats (2024) developed an open-access tool for extracting, transcribing, and temporally aligning oral (stream) and written (chat) data. As such, it provides a valuable foundation for future empirical, corpus-based research into

the multimodal nature of streaming communication. However, the tool does not yet account for other relevant modes such as gaze, gesture, or in-game events, despite their recognized importance in Twitch interactions (Recktenwald, 2017).

Modelling Twitch's IS is a way of clarifying how its data can be represented within a corpus and thus improve such tools. This paper thus proposes a model of Twitch's IS with two main goals: (i) to support theoretical and descriptive linguistic inquiries on Twitch (to identify what linguistic phenomena can be studied on the platform); and (ii) to develop tools for corpus creation. This is a challenging task due to (i) the platform's multimodality, (ii) the asymmetrical access to communicative modes depending on user roles, and (iii) the role of non-linguistic events such as gameplay actions.

Our model builds on the CMC IS developed within the CoMeRe project (Chanier & Jin, 2013). As a concrete application, we extend the tool proposed by Steven Coats.

The following section reviews prior work describing Twitch as an environment, along with the CoMeRe model of CMC. Section 3 presents our proposed model of Twitch, and Section 4 offers an application, particularly focusing on role-based data pseudonymisation.

## 2. Previous research

### 2.1 Describing the Twitch Environment

Several studies have described the Twitch environment. Notably, Recktenwald (2017) provides a detailed account, describing Twitch as "interactive television" because viewers actively participate in the interaction<sup>3</sup>, distinguishing between broadcaster and audience.

He outlines the audience's interface as including the video stream (which they receive but do not control), the chat (in which they participate as both senders and recipients), and various secondary informational or interactive elements.

<sup>1</sup> A social network considered *gaming-adjacent* as it facilitates interaction around, rather than within, gameplay.

<sup>2</sup> In 2024, Twitch recorded 2.4 million viewers and 21 billion

hours watched [source: twitchtracker.com].

<sup>3</sup> It may also be seen as a "vodcast" (Guerra, 2024).



The broadcaster’s setup typically involves two monitors: one displaying the game and webcam feed (shared with the audience), and another showing the live chat. This configuration requires the broadcaster to shift attention and body orientation depending on game events. During pauses, they turn toward the chat; otherwise, their focus remains on the game screen. Figure 1 summarizes the asymmetry of communicative access between broadcasters and viewers. Recktenwald also notes that in highly populated channels, messages appear in overwhelming volume and often disappear within seconds, limiting readability.

Participant	Visual Field	Relation to Activity	Communicative Resources	Number
Broadcaster	Game stream and chat	Player or performer	Speech, non-verbal cues	Individual
Audience	Game stream and chat	Spectator	Chat messages, emojis/emotes	Unlimited

Figure 1. Summary table (adapted from Recktenwald, 2017)

Beyond this, interactions on Twitch are mostly synchronous but can also be pre-structured. Subscriptions or donations made before a live stream, for instance, can influence the conversation (e.g., streamers begin by thanking donors). Twitch is also connected to external platforms such as Twitter or Discord, enabling communication before or beyond the stream, with ongoing exchanges brought back into the live session. Another notable feature is “shared viewership,” where two or more streamers co-broadcast and share their chats, effectively merging their communities.

## 2.2 Modelling CMC

The French project CoMeRe<sup>4</sup> developed a model of the IS for CMC. One of its key objectives was to establish a standard format for corpus representation to facilitate interoperability. This was achieved in two steps: first, modelling the IS; and second, extending the TEI standard (2019) to describe that IS<sup>5</sup>.

The IS model was designed to identify the key parameters that define a communicative situation in CMC. Though simplified, it remains relevant across a wide variety of CMC genres and environments, from the most basic to the most complex, whether synchronous or asynchronous, mono- or multimodal, simple or layered.

Chanier et al. (2014: 5) define the IS as “an abstract concept, located in time [...] where interactions between a set of participants occur within an online location”. An *online location* refers to “the properties of the set of environments used by the set of participants,” where *online* means “that interactions have been transmitted through networks” (*ibid.*: 5). Participants may be individuals or groups, and participation can be discontinuous (users may enter, exit, and return at any point). If participants cannot see, hear, or read one another, they are not considered to share the same IS. Within a single IS, interactions can vary widely and are “related to the environment(s) participants use and their corresponding modes and modalities” (*ibid.*: 5).

In this framework, modes are defined as “semiotic resources which support the simultaneous genesis of discourse and interaction” (*ibid.*: 6), while modalities are

“specific ways of realizing communication” (*ibid.*: 6). Figure 2 illustrates this model.



Figure 2. Interaction Space of CMC (from Whigham & Poudat, 2020: 7)

Although the IS model was initially designed to support TEI encoding, its abstraction and versatility make it a strong theoretical framework for describing online communicative environments. We chose to build on this model because it provides a standardized and interoperable framework, offering a robust foundation for adapting corpus design to complex platforms such as Twitch. However, the model had not previously been applied to streaming platforms, where role asymmetry, multimodality, and dynamic interface elements play a key role. Applying it to Twitch thus allows us to test its scope and reveal the kinds of adaptations such environments require.

## 3. Modelling Twitch’s Interaction Space

Our model draws on both Recktenwald’s empirical description of Twitch’s communicative setup and the IS framework’s theoretical scope. While Recktenwald (2017) offers a detailed analysis of interface features and role-based access to communicative modes, his description remains informal and lacks a standardized structure for corpus design. Conversely, the CoMeRe model provides an interoperable schema but does not fully account for layered roles or the multimodal complexity of streaming platforms.

To bridge these gaps, we extended the IS model with two key additions: (i) a finer classification of participant roles, and (ii) a finer differentiation of non-verbal modes, rather than treating them as a single category. This hybrid adaptation enhances the model’s relevance for both corpus annotation and theoretical inquiry.

Building on the IS model presented earlier, we propose an adaptation for Twitch that addresses three key issues: (i) the platform’s multimodality; (ii) asymmetric access to communicative modes by role; and (iii) the influence of non-linguistic events, such as gameplay actions. Following the format used by Whigham and Poudat (2020: 9), we developed Table 1 to represent these features.

To guide this modelling, we draw on questions posed by Whigham and Poudat (2020: 7) and add several others. First, we introduce a row specifying participant roles, which extends the original “participants” category that did not differentiate between functions. As shown by Recktenwald (2017), distinguishing between broadcaster and audience is essential in streaming contexts, as they engage in different communicative activities and modalities. We further distinguish two additional roles that are equally relevant but have

<sup>4</sup> Aimed to create a corpus of French CMC data, grounded in three principles: diversity, standard, open access (Chanier et al., 2014).

<sup>5</sup> Via the CMC-core schema (Beisswenger & Lungen, 2020), which became part of official TEI documentation in 2024.

Participant Roles	Broadcaster	Audience	Moderator	Bot
<b>Synchronicity</b>	Synchronous	Synchronous	Synchronous	Synchronous
<b>Communication Modes</b>	Oral	-	-	-
	Textual	Textual	Textual	Textual
	Non-verbal visual (proxemics, kinesics, appearance)	-	-	-
	Non-verbal textual (emojis, emotes, GIFs)	Non-verbal textual (emojis, emotes, GIFs)	Non-verbal textual (emojis, emotes, GIFs)	Non-verbal textual (emojis, emotes, GIFs)
	Non-verbal interface (subs, raids, donations, etc.)	Non-verbal interface (reporting, subs, donations, commands, etc.)	Non-verbal interface (deletions, warnings, subs, donations, commands, etc.)	Non-verbal interface (deletions, warnings, commands, etc.)
	Non-verbal game (victory, shooting, building, etc.)	-	-	-
<b>Communication Modalities</b>	Audio (voice, music, game sounds)	-	-	-
	Chat	Chat	Chat	Automated chat messages
	Video (webcam, game screen)	-	-	-
	Interactive screen events	Interactive screen events	Interactive screen events	-
	Game events	-	-	-
<b>Combination of Modes or Modalities</b>	Multimodal	Multimodal	Multimodal	Multimodal
<b>Communication Space</b>	One space, public or private	One space, public or private	One space, public or private	One space, public or private
<b>Interaction Types</b>	<i>En bloc</i> (chat)	<i>En bloc</i> (chat)	<i>En bloc</i> (chat)	<i>En bloc</i> (chat)
	<i>En continu</i> (oral)	-	-	-
	Proxemic and kinesic acts (gestures, gaze, facial expressions, posture, appearance)	-	-	-
	Screen-based activities (deletions, warnings, subs, donations, commands, gameplay actions, etc.)	Screen-based activities (reporting, subs, donations, commands, etc.)	Screen-based activities (deletions, warnings, subs, donations, commands, etc.)	Screen-based activities (deletions, warnings, commands, etc.)
<b>Temporal Frame</b>	Start and end of the live session	Start and end of the live session	Start and end of the live session	Start and end of the live session

Table 1: Properties of Twitch’s Interaction Space

varying levels of access:

- **Broadcaster:** the person streaming the content.
- **Audience:** viewers of the live stream.
- **Moderator:** designated by the broadcaster to manage chat interactions.
- **Bot:** automated program sending reminders, announcements, or launching mini-games.

Next, we address the temporal dimension. Twitch live streams are inherently synchronous. Interaction must occur in real time: older chat messages are quickly overwritten, making delayed participation difficult. Our focus is exclusively on live sessions, not VODs, which are considered archival and fall outside the IS framework, defined by a clear beginning and end.

We then examine the available modes and modalities. A mode is understood as a general type of communication, while a modality refers to a specific form it takes in a given environment. We assume all participants have access to all modes as recipients, but not as producers. Each mode corresponds to a communicative action (either verbal or non-verbal). These acts can be initiated by participants or the system, with either direct (e.g., speech) or indirect communicative function (e.g., alerts). The following modes and their associated modalities are identified:

- **Oral mode**, via **audio** (voice, music, game sounds): broadcaster only.
- **Textual mode**, via **chat**: all participants (bots use pre-programmed messages).
- **Non-verbal visual mode**, via **video** (webcam, screen;

proxemics, kinesics, appearance): broadcaster only.

- **Non-verbal textual mode**, via **chat** (emojis, emotes, GIFs): all participants except bots (non-interactive use).
- **Non-verbal interface mode**, via **interactive screen events** (subs, raids, donations, commands, warnings): all participants, but with varying permissions.
- **Non-verbal game mode**, via **game events** (victory, shooting, building): broadcaster only.

This analysis highlights the multimodal and role-dependent nature of communication on Twitch. Users interact through combinations of modes and modalities that vary depending on their function in the stream.

Regarding communication space, Twitch provides a single live interaction space that can be public or private depending on the broadcaster's settings. Other publication modes (VODs and stories) exist, but are not modelled here. For our purposes, the temporal frame of the IS corresponds strictly to the live stream’s start and end.

Finally, Twitch involves various interaction types (which played a key role in the CoMeRe project’s use of TEI):

- *En bloc* interaction (chat, emotes, emojis, GIFs): accessible to all.
- *En continu* oral interaction: broadcaster only.
- Proxemic and kinesic interaction (gestures, gaze, posture): broadcaster only.
- Screen-based activities (interface or gameplay events): accessible to all, but in distinct ways depending on role.



This modelling of Twitch’s IS highlights both its complexity and its relevance for corpus design and discourse analysis. The complexity lies in the range of communicative modes and the unequal access tied to participant roles, a dimension not explicitly addressed in the original CoMeRe framework.

## 4. Uses case

## 4.1 Technical considerations

Steven Coats (2024) developed a highly valuable pipeline tool for corpus linguistics, available as open-access code. It enables the extraction of audio and chat data from streaming platforms (YouTube and Twitch), automatic audio transcription, and temporal alignment of audio and chat into an HTML file with four columns: timestamp, audio transcription, usernames, and chat messages. Coats (2024: 19) identifies three future development paths: (i) adapting the tool to other streaming platforms; (ii) integrating video analysis models for automated descriptions of visual content, especially in-game; and (iii) incorporating gesture and facial expression recognition.

This tool aligns well with our proposed IS model, as it supports the synchronous structuring of multiple communicative modes. One of the main challenges in working with Twitch data is that oral, written, and visual information all unfold on the same temporal axis. The alignment of speech (streamer’s voice) and chat (written text) enables simultaneous analysis of these data streams.

However, as Coats notes, two additional columns would be beneficial: one for screen activity and another for non-verbal visual behavior. In our model, these correspond to two distinct non-verbal modes: interface and visual. We also distinguish a third non-verbal mode: game, which appears independently within the stream. Recktenwald (2017) used four columns in his corpus: timestamp, game events, streamer transcript, and chat (with usernames). In the output developed by Coats, we propose adding a fifth column to specifically capture gameplay events, using a transcription scheme inspired by Recktenwald’s work.

One mode remains to be addressed: non-verbal textual, referring to emojis and emotes (graphical signs encoded in text, often unique to Twitch). For example, typing “LUL” in chat displays a specific emote<sup>6</sup>, but these can also be selected from the emote bar. Thanks to Coats’s HTML format, emotes are transcribed in two ways: both as the original text and as the corresponding image. This dual representation preserves both the textual trace and the visual fidelity of the chat, an approach also supported by Recktenwald (2017).

As it stands, the tool supports transcription and alignment of oral, textual, and non-verbal textual modes. Three others remain to be fully integrated: non-verbal visual, interface, and gameplay, each requiring a dedicated column in the HTML output.

The inclusion of a username column also proves useful for

identifying participant roles. This feature enables automatic pseudonymisation while supporting role-based distinctions. We developed a Python script that pseudonymises both the username column and any “@mentions” in the chat. Each pseudonym receives a unique role-based label:

- Broadcaster → [Broadcaster]
- Audience → [User + unique ID]
- Moderators and Bots → [Modo + unique ID]

Role identification relies on badges (icon images) found in the HTML, which indicate whether a user is a broadcaster, moderator/bot (they have the same one). While Twitch uses many badge types<sup>7</sup>, we focus on those most relevant to our model. Future improvements could include: (i) distinguishing between moderators and bots; and (ii) pseudonymisation user mentions without the “@” symbol. Figure 3 illustrates role-based pseudonymisation: the user column displays each role with a unique ID, while the fourth column shows pseudonymised mentions.

time	text	author	message
1953.000		User19	@User12 Surtout pour une question de sécurité avamiendWTF 🤪
1955.000		Moderator3	POGGERS
1957.000		User67	also ?
1958.000		User88	banjour Madame Mind @brownsator comment ça va ?
902.000		User37	@Moderator3 avamiendBOU avamiend. 🤪👁
904.000		User19	@User38 Tu serais le premier sur ma liste Kappa 🤪
905.218	Ça craque, oui.		
908.000		User30	Rosh 1 poncoJAM poncoJAM poncoJAM 🤪👁👁
909.000		User6	@User39 @Moderator3 avamiendLx CursuL avamiendLx CursuL3 avamiendLx CursuL8 avamiendLx CursuL8 avamiendLx CursuL8 avamiendLx CursuL8 👁👁👁👁👁👁👁👁👁👁👁👁👁👁👁👁👁👁👁👁

Figure 3: Role-based pseudonymisation

We would like to briefly comment on the example shown in Figure 3. These data were collected before Coats’s tool was available, using a different method, and are no longer accessible on Twitch. Following guidance from S. Coats, we modified the Python code to reproduce the same output format using pre-downloaded data. Specifically, he suggested extracting audio from archived videos using `ffmpeg`, transcribing the audio with Whisper, and converting chat data from `.json` to `.html` using the TwitchDownloader CLI<sup>8</sup>, before running the rest of the pipeline. We followed this approach and successfully generated a comparable output. We consider this an important advancement, given that Twitch data is often ephemeral and not always available for direct extraction from the platform.

This technical differentiation of modes also raises important theoretical questions, as discussed in the next section.

## 4.2 Theoretical considerations

Differentiating user roles within the proposed model opens important theoretical perspectives for analysing discourse on Twitch. As Recktenwald (2017: 78) notes, audience members “mostly produce single-turn messages that are highly context dependent,” while broadcasters “tend to elaborate and respond with several utterances.” This asymmetry, he explains, reflects the broadcaster’s varying cognitive and physical engagement depending on on-screen

<sup>6</sup> <https://twitchemotes.com/>

<sup>7</sup> <https://help.twitch.tv/s/article/twitch-chat-badges-guide?language=fr>

<sup>8</sup> <https://github.com/lav295/TwitchDownloader>

activity, whereas audience participation is more stable. This linguistic imbalance appears closely linked to users' unequal access to communicative modalities. An annotated corpus could help explore how these asymmetries shape interactional dynamics on the platform.

Another key issue concerns how meaning is constructed in an environment where multiple channels operate simultaneously. Twitch interactions are shaped by visual stimuli (gameplay, interface events, and streamer gestures) which structure the timing and nature of exchanges, as discussed in Section 2.1.

Recktenwald (2017) describes this as *intermodality*. A corpus structured according to our model would allow for analysis of correlations between screen events and linguistic acts, offering insight into how these elements trigger or influence discourse.

Finally, Twitch is organized around communities, and ethnographic research (e.g., Hamilton et al., 2014) has shown that these shape the platform's discursive norms. Comparing data from different Twitch communities could reveal how social context affects language use.

Together, these theoretical avenues demonstrate the value of our model in supporting corpus-based research on Twitch, particularly for investigating how interactional structures and social dynamics shape multimodal discourse.

## 5. Summary and outlook

This paper proposed a model of the IS of Twitch, based on the CoMeRe framework, adapted to account for the platform's multimodal, asymmetric, and dynamic communicative environment. The model illustrates how participants access and use different communicative modes depending on their roles, and opens new avenues for corpus annotation and discourse analysis.

We demonstrated how this model can inform the technical development of multimodal corpus tools, particularly by extending S. Coats's alignment tool to include role-based pseudonymisation and the potential integration of interface, gameplay, and kinesic/proxemic layers.

Future work will focus on two directions: (i) extending S. Coats's tool to incorporate the three non-verbal components; and (ii) releasing and annotating a version 1.0 of the Twitch corpus within an open science framework. This corpus is currently being developed as part of the CMC corpus-building initiative led by the OLiNDiNUM project<sup>9</sup> (Robert & Pietrandrea, 2024). The proposed model supports this effort and serves as a framework for both annotation and structural analysis. It also invites further theoretical inquiry into the co-construction of meaning in role-asymmetric, intermodal communication spaces.

## 6. References

Aklid, J. (2015). Twitch, a Breath of Fresh Air? An Analysis of Sexism on Twitch.tv. Dissertation on Language and Literature. Linnaeus University.  
 Beißwenger, M. and Lungen, H. (2020). CMC-core: a schema for the representation of CMC corpora in TEI.

Corpus, (20).  
 Chanier, T. and Jin, K. (2013). *Defining the online interaction space and the TEI structure for CoMeRe corpora*. Projet CoMeRe (Communication Média par les Réseaux), IR Corpus-écrits.  
 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 of Language Technology and Computational Linguistics*, 29(2), pp. 1–31.  
 Coats, S. (2024). A framework for analysis of speech and chat content in YouTube and Twitch streams. In *Proceedings of the 11th Conference on CMC and Social Media Corpora for the Humanities*. Nice, France: CORLI; Université Côte d'Azur, pp. 16–19.  
 Ford, C.; Gardner, D.; Horgan, L. E.; Liu, C.; Tsaasan, A. M., Nardi, B. and Rickman, J. (2017). Chat speed OP PogChamp: Practices of coherence in massive Twitch chat. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 858–871.  
 Guerra, A. (2024). Podcasts & vodcasts: the “alternative radios” of the digital world. *Journal of Inclusive Methodology and Technology in Learning and Teaching*, 4(4).  
 Hamilton, W. A.; Garretson, O. and Kerne, A. (2014). Streaming on Twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1315–1324.  
 Nakandala, S.; Ciampagliai, G. L.; Suz, N. M. and Ahnyz Y.-Y. (2017). Gendered Conversation in a Social Game-Streaming Platform. In *Proceedings of the international AAAI conference on web and social media*, 11(1), pp. 162–171.  
 Olejniczak, J. (2015). A Linguistic Study of Language Variety Used on Twitch.Tv: Descriptive and Corpus-Based Approaches. *Redefining Community in Inter-cultural Context*, 4, pp. 329–334.  
 Recktenwald, D. (2017). Toward a transcription and analysis of live streaming on Twitch. *Journal of Pragmatics*, 115, pp. 68–81.  
 Robert, A. and Pietrandrea, P. (2024). The 3DSeTwitch corpus – A three-dimensional corpus annotated for sexist phenomena. In *Proceedings of the 11th Conference on CMC and Social Media Corpora for the Humanities*. Nice, France: CORLI; Université Côte d'Azur, pp. 110–112.  
 Text Encoding Initiative (2019). “Text Encoding Initiative Consortium” [https://teic.org/].  
 Wigham, C. R.; and Poudat, C. (2020). Corpus complexes et standards : un retour sur le projet CoMeRe. *Corpus*, (20).

<sup>9</sup> https://olindinum.huma-num.fr/

# Strategic Transparency or Deliberate Ambiguity? A Corpus-Assisted Multimodal Analysis of Airline CSR Communication on LinkedIn

Fabiola Notari

University of Modena and Reggio Emilia, Italy

E-mail: fnotari@unimore.it

## Abstract

Airline companies face intense scrutiny concerning their societal and environmental impacts. As such, they increasingly rely on social media for Corporate Social Responsibility (CSR) communication. However, the inherently multimodal nature of these platforms complicates objective assessments of transparency. This paper introduces and empirically tests an integrated analytical framework for classifying multimodal CSR signals (*soft*, *semi-hard*, *hard*), enabling a systematic examination of how transparency is strategically constructed online. Drawing on a purpose-built corpus of LinkedIn posts from four major international airlines (Delta Air Lines, British Airways, ITA Airways, and China Southern Airlines), representing distinct US, Italian (EU), and Chinese communicative contexts, the analysis combines Signalling Theory, Systemic Functional Linguistics (SFL), and Multimodal Discourse Analysis (MDA). Unlike previous approaches, this framework integrates process-driven linguistic annotation and multimodal coding, enabling robust, replicable comparison across institutional contexts and highlighting practices that either foster or obscure transparency. The findings reveal that, while all companies provide clear self-presentation and some accessible data, most favour *soft* and *semi-hard* signals that limit verifiable, externally validated information, with *hard* signals remaining rare. Ambiguity and partial disclosure are thus strategically preferred over full transparency—underscoring how CSR communication, in practice, serves primarily to enhance corporate image rather than maximise accountability. By bridging discourse analysis with transparency metrics, the study demonstrates how digital CSR signals are classified and how strategic ambiguity and selective disclosure affect stakeholder perceptions and the credibility of corporate discourse in digital environments.

**Keywords:** Transparency, Multimodal Discourse Analysis, Social Media Communication, Signalling Theory, Corporate Social Responsibility

## 1. Introduction

Every year, airline companies publish thousands of social media posts dedicated to sustainability and Corporate Social Responsibility (CSR). On LinkedIn alone, over 10,000 CSR-related posts appeared in 2024, underlining the intense reputational scrutiny and the centrality of digital self-presentation for the sector. Yet, the key question remains: How transparent are these digital claims, and how is transparency itself discursively and visually constructed? The challenge is intensified by the inherently multimodal character of platforms such as LinkedIn, where text, visuals, and interactive features blend to produce subtle (and often ambiguous) communicative effects. While recent research has analysed aspects of corporate CSR discourse, few studies have adopted a truly integrated multimodal and corpus-assisted approach, nor have they systematically interrogated the *continuum* between disclosure and omission in digital CSR signalling. This study addresses this gap by introducing and empirically applying an original, replicable framework that classifies CSR signals as *soft*, *semi-hard*, or *hard*—capturing not only their accessibility and informativity, but also the strategic ways in which omission and ambiguity are mobilised. Drawing on Signalling Theory, Systemic Functional Linguistics (SFL), and Multimodal Discourse Analysis (MDA), the framework enables robust, cross-context comparison of digital CSR performance across linguistic and visual dimensions. The investigation is guided by three research questions: *RQ1*: What are the most salient CSR themes (e.g., environmental sustainability, community engagement, governance) in the digital multimodal CSR discourse of airline companies on LinkedIn? *RQ2*: How do these

companies strategically construct transparency, as evidenced by their use of *soft*, *semi-hard*, and *hard* signals? *RQ3*: How are these transparency signals linguistically, visually, and semiotically constructed across the corpus?

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature, Section 3 introduces the theoretical framework, and Section 4 details the methodology. Section 5 presents the core findings, illustrated through both quantitative mapping and an in-depth qualitative case analysis. Concluding remarks and avenues for further research are outlined in Section 6.

## 2. Literature review

Recent scholarship underscores transparency as a central strategic element in Corporate Social Responsibility (CSR) communication, especially amidst public skepticism and greenwashing concerns (Kim & Lee, 2018; Lee & Comello, 2018). It is recognized as essential for building trust and mitigating skepticism through accountability and stakeholder participation (Baraibar-Diez & Luna-Sotorrio, 2018; Kashyap et al., 2020). Indeed, transparency is conceptualized beyond mere factual disclosure; it requires substantial information and active stakeholder involvement to effectively reduce consumer skepticism, particularly in stigmatized industries. More recently, the concept has been expanded to “performance transparency”—the intentional and objective provision of information on corporate actions to signal sincerity (Liu et al., 2023). This approach, advocating for consistent and verifiable practices, aligns closely with the notion of *hard* evidential signals central to this study. However, social media platforms significantly complicate these efforts. Described as the “Wild West” of CSR communication (Tench & Jones, 2015), their chaotic

nature risks message “co-destruction” if transparency is inadequate. Empirical studies confirm that firms often resort to one-way “stakeholder information strategies” (Vitellaro et al., 2022) or engage in “parallel talking” on non-core issues (Gómez-Carrasco et al., 2021), leaving the platforms’ potential for authentic, credibility-building engagement largely untapped (Yang et al., 2018). Taken together, these studies underscore the complexity of transparency in CSR communication. While recent work—such as Poppi’s (2025) study of airline emission disclosures—has shown how textual choices and omissions can subtly mask strategic ambiguity, a critical gap remains in systematically analysing how these processes unfold within the inherently multimodal environment of social media. In particular, there is still limited research on how omission, vagueness, and semiotic design interact to shape transparency or ambiguity in real-world digital CSR communication. The present study addresses this need by employing an integrated framework to analyze real-world social media content, thus advancing ecological validity and addressing the research gaps identified by previous literature.

### 3. Theoretical framework

This study introduces an innovative, integrated framework to analyze the complex construction of transparency in CSR communication on social media. Drawing on Signalling Theory from economics (Spence, 1973), this framework operationalizes the concept of communicative signals. The central contribution of this framework is a novel typology that classifies CSR signals as *soft*, *semi-hard*, or *hard* based on their strategic balance of accessibility and informativity.

### 4. Methodology

#### 4.1 The LinkedIn Corpus

This study draws its empirical basis from a purpose-built corpus compiled from the official English-language LinkedIn pages of four major international airlines: Delta Air Lines (US), British Airways (UK), ITA Airways (Italy/EU), and China Southern Airlines (China). To ensure a data-driven and unbiased sample, the 55 most recent posts from each company were collected in reverse chronological order (May 2023–July 2025), resulting in a total corpus of 220 posts. This approach—deliberately avoiding pre-selection via CSR-related hashtags—provides a more accurate estimation of the prominence and nature of CSR topics within each company’s overall communication strategy. Only posts with English-language captions were included, and reshares without original commentary were excluded. The final dataset thus comprises 220 posts, each containing both textual and visual content.

#### 4.2 Analytical framework

The analytical process was structured into three sequential phases. In the first phase, all 220 LinkedIn posts (55 per airline) were manually screened and annotated using INCEpTION (Klie et al., 2018) to identify CSR-related

content. Relevant posts were then thematically categorised into inductively derived micro-themes, which were further grouped into broader macro-domains. Basic descriptive and comparative statistics were used to explore thematic distribution across airlines. The second phase focused on how transparency was signalled through three dimensions: internal informativity (self-reported data), external accessibility (hyperlinks), and external informativity (third-party validation). Each post was coded along these dimensions using a binary scheme (1 = present, 0 = absent), and subsequently classified as a *soft*, *semi-hard*, or *hard* signal depending on its degree of evidentiary support. This classification enabled cross-airline comparison of transparency strategies. The final phase examined how these signals were discursively and visually constructed. Linguistic annotation followed the principles of Systemic Functional Linguistics and Appraisal Theory (Halliday, 1994; Martin & White, 2005), while visual and symbolic resources were interpreted through corpus-assisted multimodal analysis. Key semiotic processes—such as iconisation, recursivity, and erasure (Irvine & Gal, 2000; Notari, 2024)—were used to identify how meaning and credibility were constructed across modalities.

### 5. Findings

#### 5.1 Illustrative Example: China Southern Airlines’ ‘Green Leadership’ Signal

A May 2025 LinkedIn post by China Southern Airlines (see Fig. 1) announced the release of what it called “the industry’s first White Paper on Green Development,” asserting its commitment with the phrase “China Southern Airlines prioritizes climate action” and referencing compliance with IFRS S2 standards. Linguistically, the use of declarative clauses and policy-oriented relational processes (e.g., ‘We follow IFRS S2 standards in governance, strategy, risk, and metrics’) indicates high internal informativity, grounded in formal positioning. However, the post does not provide specific quantitative indicators (e.g., emission metrics or timelines), and no third-party validation is mentioned—thus limiting external informativity. Multimodally, the image of a 787 Dreamliner iconises *green innovation*, visually reinforcing the airline’s sustainability narrative. The visual layer is emotionally resonant and symbolic, but lacks embedded data or external logos. Despite referring to a detailed White Paper, the post provides no hyperlink or attachment, thereby reducing external accessibility. This combination—textual policy assertion without supporting metrics, emotionally symbolic visuals, and the absence of direct access or third-party validation—positions the message as a *semi-hard* signal. It constructs credibility through strategic self-presentation while avoiding full verifiability. As such, the post illustrates how transparency can be performed through curated omission and selective accessibility, particularly within highly visible environmental claims.



Figure 1: China Southern Airlines Post

## 5.2 Quantitative Overview

### 5.2.1 Thematic Distribution of CSR Posts (RQ1)

The quantitative analysis of 220 LinkedIn posts (55 per airline) reveals clear cross-airline variation in CSR visibility and thematic focus. ITA Airways published the highest proportion of CSR-related posts (60.0%; 33/55), followed by Delta Air Lines (52.7%; 29/55), China Southern Airlines (41.8%; 23/55), and British Airways (34.5%; 19/55). Table 1 presents the internal distribution of CSR content across the three macro-themes. Community & Social Engagement is dominant across all carriers, particularly for Delta (75.9%) and ITA (51.5%). Governance & Partnerships is especially salient for ITA (39.4%), while Environmental Sustainability remains marginal throughout, never exceeding 9.1% of CSR posts. These differences were statistically significant ( $\chi^2 = 46.12$ ,  $p < 0.00001$ ), indicating distinct strategic priorities in airline CSR communication.

Theme	Delta Air Lines	China Southern	British Airways	ITA Airways
Environmental Sustainability	6.9%	9.1%	5.3%	6.1%
Community & Social Engagement	75.9%	52.2%	63.2%	51.5%
Governance & Institutional Partnerships	13.8%	17.4%	31.6%	39.4%

Table 1: Distribution of CSR Themes by Airline (% of CSR-related posts)

### 5.2.2 Strategic construction of transparency (RQ2)

CSR-related posts were classified into three transparency signal categories—*Soft*, *Semi-Hard*, and *Hard*—based on

the presence of self-reported quantitative data, hyperlinks, and third-party validation (see Methodology). The results, summarised in Table 2, reveal highly significant cross-airline variation in their strategic communication of transparency ( $\chi^2 = 81.82$ ,  $p < .00001$ ). ITA Airways relies predominantly on *Semi-Hard* signals (93.9%), systematically pairing quantitative claims with accessible references, but only rarely providing full external validation. In contrast, *Delta Air Lines* (58.6%) and *China Southern Airlines* (69.6%) primarily favour *Soft* signals, characterised by symbolic or emotionally resonant narratives that typically lack robust substantiation. British Airways adopts a balanced strategy, distributing its CSR communication relatively evenly between *Soft* (36.8%) and *Semi-Hard* (63.2%) signals. Across all airlines, *Hard* signals—representing comprehensive transparency with complete external validation—remain exceptionally rare, appearing minimally in ITA Airways (3.0%) and Delta Air Lines (10.3%), and completely absent in British Airways and China Southern Airlines. Collectively, these findings underscore a clear industry preference for strategically managed ambiguity and selective disclosure rather than fully verifiable transparency, highlighting the reputational and communicative risks airlines perceive in providing fully externally validated claims.

Signal Category	Delta Air Lines (n=29)	China Southern Airlines (n=23)	British Airways (n=19)	ITA Airways (n=33)
Soft	58.6%	69.6%	36.8%	3.0%
Semi-Hard	31.0%	30.4%	63.2%	93.9%
Hard	10.3%	0.0%	0.0%	3.0%

Table 2: Transparency Signals by Airline (% of CSR-related posts)

### 5.2.3 Accessibility vs. Informativity strategic profiles (RQ3)

The corpus-assisted multimodal analysis reveals clear qualitative distinctions in how *soft*, *semi-hard*, and *hard* transparency signals are constructed linguistically, visually, and semiotically. Linguistically, soft signals frequently rely on emotionally resonant language, realised mainly through affective and material processes (e.g., ‘support,’ ‘connect’) paired with abstract or affectively charged objects (e.g., ‘hope,’ ‘community’), without offering quantifiable details or external validation. *Semi-hard* signals, while still predominantly narrative, integrate internally reported data using more specific material and relational processes—for instance, claims of ‘reduced CO<sub>2</sub> emissions by 12%’ or ‘implemented new accessibility features,’ albeit without external verification. *Hard* signals, though exceptionally rare, explicitly provide externally validated claims or third-party recognition, most often realised through dense declarative mood and institutionalised material processes (e.g., ‘scored 100% on the Disability Equality Index’). Visually and semiotically, these linguistic patterns align closely with the strategic use of imagery and design. *Soft* signals typically employ emotionally connoted visuals

(children, volunteers, or smiling staff), with low informational density and a prevalence of iconisation of abstract CSR values—mirroring the use of symbolic material processes in the text. *Semi-hard* signals shift towards professional and institutional imagery—branded events, uniforms, procedural scenes—sometimes including numeric overlays but usually remaining cautious in visual detail, often reinforcing relational processes and procedural credibility. *Hard* signals differ markedly, foregrounding third-party logos, visible badges, certificates, and numeric data explicitly embedded in the visual field, thus visually anchoring external credibility and accountability. Semiotic strategies—iconisation, recursivity, and erasure—further support this continuum. *Soft* signals rely heavily on the iconisation of abstract values, while semi-hard signals often use institutional motifs to reinforce professionalism and trust, with recursivity ensuring thematic continuity through repeated verbal and visual cues. *Hard* signals explicitly integrate external evidence and data through consistent repetition across modes, minimising erasure. Conversely, erasure—or the strategic omission of expected details—dominates soft and *semi-hard* signals, preserving interpretative openness and limiting direct scrutiny, while *hard* signals minimise erasure by fully integrating evidence across textual and visual modes, ensuring complete external verifiability. Overall, these qualitative insights—framed through the lens of systemic functional linguistics and multimodal discourse analysis—highlight strategic corporate communication patterns. Airlines predominantly favour partial transparency through *soft* and *semi-hard* signals, strategically balancing narrative appeal, material and relational process selection, and reputational risk, while full transparency (*hard* signals) remains rare and carefully managed.

## 6. Conclusions and contribution

This study demonstrates that within the multimodal CSR discourse of major airlines on LinkedIn, companies consistently foreground Community & Social Engagement, although strategic divergence clearly emerges in how transparency is communicated. Most airlines opt predominantly for *soft* and semi-hard signals, balancing narrative persuasion and partial disclosure, thereby preserving significant zones of ambiguity. *Hard* signals, characterised by externally validated claims, remain notably scarce across all carriers. From a theoretical perspective, this research contributes a novel analytical framework that systematically integrates Signalling Theory, Systemic Functional Linguistics, and Multimodal Discourse Analysis, allowing for nuanced, replicable assessments of transparency within digital corporate communication. Empirically, the study provides insights into the discursive mechanisms by which airlines strategically leverage ambiguity and selective disclosure to balance reputational control and stakeholder accountability. Practically, the framework offers robust tools for stakeholders and communicators in the field of CMC to critically evaluate and enhance the credibility of

CSR messaging in digital environments, thereby promoting more transparent, accountable, and effective digital corporate communication practices.

## 7. References

- Bibliographical references should be listed in alphabetical order at the end of the article. The title of the section, "References", should be a level 1 heading. The first line of each bibliographical reference should be justified to the left of the column, and the rest of the entry should be indented by 0.35 cm.
- Baraibar-Diez, E., & Luna-Sotorrío, L. (2018). The mediating effect of transparency in the relationship between corporate social responsibility and corporate reputation. *Review of Business Management*, 20(1), 5–21.
- Gómez-Carrasco, P. P., Guillamón-Saorín, E., & García Osma, B. (2021). Stakeholders versus firm communication in social media: The case of Twitter and corporate social responsibility information. *Business Ethics: A European Review*, 30(4), 483–498.
- Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). London: Edward Arnold.
- Irvine, J. T., & Gal, S. (2000). Language ideology and linguistic differentiation. In P. V. Kroskrity (Ed.), *Regimes of Language: Ideologies, Politics, and Identities* (pp. 35–84). Santa Fe, NM: School of American Research Press.
- Kashyap, R., Menisy, M., Caiazza, P., & Samuel, J. (2020). Transparency versus performance in financial markets: The role of CSR communications. *arXiv preprint, arXiv:2008.03443*.
- Kim, H., & Lee, T. H. (2018). Strategic CSR communication: A moderating role of transparency in trust building. *International Journal of Strategic Communication*, 12(2), 107–124.
- Klie, J.-C., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9.
- Kress, G., & van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design* (2nd ed.). London: Routledge.
- Lee, T. H., & Comello, M. L. N. G. (2019). Transparency and industry stigmatization in strategic CSR communication. *Management Communication Quarterly*, 33(1), 68–85.
- Liu, Y., Heinberg, M., Huang, X., & Eisingerich, A. B. (2023). Building a competitive advantage based on transparency: When and why does transparency matter for corporate social responsibility? *Business Horizons*, 66(4), 517–527.
- Martin, J. R., & White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. New York: Palgrave Macmillan.
- Notari, F. (2024). Cracking the code of change in EU legal discourse: signifying practices shaping inclusion for the vulnerable in the digital age. *Comparative*



- Legilinguistics*, 60, 342–383.
- Poppi, F. (2025). Airlines' emission disclosures: The fine line between opportunity and environmental inaction. *Iperstoria*, (25).
- Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.
- Tench, R., & Jones, B. (2015). Social media: The Wild West of CSR communications. *Social Responsibility Journal*, 11(2), 290–305.
- Vitellaro, F., Satta, G., Parola, F., & Buratti, N. (2022). Social media and CSR communication in European ports: The case of Twitter at the Port of Rotterdam. *Maritime Business Review*, 7(1), 1–21.
- Yang, J., Basile, K., & Letourneau, O. (2018). The impact of social media platform selection on effectively communicating about corporate social responsibility. *Journal of Marketing Communications*, 26(1), 65–87.
- .

# Emerging digital discourse traditions: A contrastive analysis of the r/todayilearned subreddit and its German and French counterparts

Dominique Dias

Sorbonne University, France

E-mail: Dominique.Dias@sorbonne-universite.fr

## Abstract

This paper explores the development of discourse traditions in parallel online communities. Focusing on the r/todayilearned phenomenon, the study presents a contrastive analysis of language practices observed in 1,500 posts and their associated comments from the English subreddit r/todayilearned, the German r/HeuteLernTelch, and the French r/Aujourd'huiJ'aiAppris. By analyzing the verbs *learn*, *know*, and *understand* and their cooccurrences across English, German, and French, the study identifies shared linguistic practices in articulating knowledge acquisition and transmission. The findings reveal that *learn* is primarily used to affirm personal knowledge within a defined temporal framework, while *know* and *understand* play a more interactive role, negotiating shared understanding between users. This comparative approach aims to uncover the shared and unique discursive norms shaping these online spaces, providing insights into the cross-cultural adaptation and innovation of digital communication practices within comparable virtual communities.

**Keywords:** discourse traditions, reddit, knowledge

## 1. Introduction

The notion of discourse traditions (henceforth DT), originally conceived by German Romanists (Koch 1997) and now enjoying a new revival (Kabatek 2014; Winter-Froemel 2020; Winter-Froemel et al. 2022), refers to conventionalized ways of speaking and writing that go beyond the grammatical rules of a language. DT suggest that communication is not just about forming sentences based on linguistic rules but also involves adhering to established pragmatic conventions. They influence textual genres, and evolve through processes like differentiation, mixing, convergence and disappearance (Koch 1997; Eckkrämmer 2019a). The notion of DT has primarily been used in historical linguistics in the study of medieval genres. However, many contemporary approaches in text and media linguistics apply similar principles to current issues (Müller-Lancé 2022; Eckkrämmer 2019b). Here, we use the term digital DT to refer to online language practices whose social and interactional dynamics shape discursive habits that may, over time, contribute to the emergence of new language patterns or genres. These habits and patterns are part of a broader knowledge system (Fix 2006) that encompasses text genres and their defining characteristics, including structural, functional, and stylistic conventions. This knowledge can be intuitively internalized (as in everyday communication genres), explicitly learned (such as academic genres), acquired within a community of practice (Mercieca 2017), and/or imported from other languages. In what follows, we propose to look at the last case: the r/todayilearned subreddit serves as an excellent example for studying the emergence of digital DT within the English-speaking context before they expand to other linguistic and cultural areas. This subreddit serves as a crowdsourced knowledge-sharing platform, where users post concise, engaging facts that spark curiosity and discussion. These communities adopt the same core principles—sharing fact-based, verifiable content—while adapting to their respective linguistic and cultural contexts.

## 2. Literature review

### 2.1 Exploring Reddit and subreddits

Reddit.com, founded in 2005 by Alexis Ohanian and Steve Huffman (Anderson 2015: 4) is a platform where registered users, known as “Redditors”, post text or link-based content. These posts, along with user-generated comments, are subject to a system of upvotes and downvotes. Positive votes contribute to a user’s karma points, while negative votes detract from them (Anderson 2015: 3). The platform also features a publicly accessible wiki containing essential information and the community’s etiquette guidelines, known as “reddiquette” (Anderson 2015: 3). Each subreddit functions as a specialized discussion forum centered around a specific topic. These communities are overseen by volunteer moderators, who manage discussions, enforce community guidelines, and ensure that interactions remain relevant and constructive (Choi et al. 2015: 234). Because of its structural and community-driven features, Reddit has served as a valuable corpus for numerous academic studies across various disciplines interested in studying relational dynamics within online communities (Proferes et al. 2021). The diverse subreddits provide valuable insights into digital interactions or social roles (Buntain & Goldbeck 2014). Datta and Adar (2019) explored, for instance, the phenomenon of community-on-community conflict on the platform. Their work aimed to identify mechanisms for determining users’ social and anti-social affiliations based on their commenting behavior. Botzer et al. (2023) examined the process of moral judgment within the context of online social systems. Furthermore, some studies have explored Reddit’s potential as a platform for bridging the gap between science and the public with subreddits like r/ELI5 (Explain me like I’m 5) (Fan et al. 2019) or r/todayilearned. These subreddits are supposed to play a crucial role in stimulating public curiosity and fostering a continuous desire to learn.



## 2.2 Studying digital discourse traditions

Subreddits have predominantly been investigated within the field of computational linguistics, primarily focusing on social interaction analysis and the construction of discursive roles. There is a notable scarcity of research from the perspectives of text- and media linguistics. The existing studies almost exclusively concentrate on English-language subreddits, highlighting a significant gap in contrastive analyses. Considering subreddits through the lens of DT allows for the exploration of how established conventions can be adopted and adapted across different linguistic and cultural contexts. Furthermore, the notion of DT offers a new perspective on how languages influence one another, shifting the focus from lexical units to specific communicative contexts. Rather than viewing for instance English and German as abstract linguistic entities, Kabatek (2015) emphasizes that it is the English used in a computer manual, a scientific article, or a late-night show that interacts with its German counterpart in the same contexts (2015: 56).

The identification and study of DT present methodological challenges, as highlighted by Kabatek (2015). He argues that understanding “true traditionality” in discourse should not rely on rigid scientific hypotheses but rather on expert intuition, empathy, and a deep familiarity with textual production (2015: 60). Kabatek advocates for a corpus-driven approach, where the search for traditions within texts or discourses takes precedence over predefined categories. He suggests that a symbiotic collaboration between expert intuition and automated analysis holds promise for advancing this research (2015: 60). DTs approaches are difficult to implement in quantitative analyses of corpus data primarily because of their lack of clear boundaries (Rosemeyer 2022: 661). Rosemeyer argues that DT must be derived inductively, allowing linguistic patterns to emerge from corpora in a bottom-up fashion, rather than being imposed through predefined categories. The contrastive study of DT has sparked significant debate regarding how national traditions evolve under external influences. As Kaiser (2003: 184) highlights, shifts in terminology, syntax, style, and macrostructure of texts raise questions about the forces driving discursive change. Empirical studies illustrate how DT cross linguistic boundaries (see Kaiser 2003 on the exploration of Anglo-Saxon influences on academic language in Hispano-America or more recently in a digital context Dias 2024 or Werle 2025 about digital phenomena reflecting both the digital adaptation of traditional DT and their migration across linguistic and cultural spheres).

Despite the increasing globalization of digital discourse, research on DT in online environments remains scarce. By investigating how digital DTs evolve and interact across languages, this study aims to provide an explorative corpus-driven analysis and contribute to the development of a methodological framework for better understanding DT in digital contexts.

## 3. Corpus and method

The name Reddit is a pun derived from the phrase “I read it on Reddit”. The clever naming strategy enhances Reddit’s identity as a community-driven source of information, fostering an ecosystem where collective knowledge is exchanged and expanded. On Reddit, the

majority of discussions and subreddits operate in English. However, some communities exist in other languages, often mirroring popular English-language subreddits.

Acronym	TIL	HLI	AJA
Language	English	German	French
Created in	2008	2014	2022
Members	40M	271k	242k
Tokens	443,255	84,084	402,617
Posts	500	500	500
Comments	15,473	2,256	16,796
Timeframe	2023-2025		

Table 1: General Data of the corpus

The corpus analyzed in this study is made up of posts and comments extracted from three comparable subreddits. Comparability is based on the name of the community, the operating principle, the use of a similar acronym in the three languages and equivalent number of posts (500 in each language). However, the general data of the corpus (Table 1) already shows certain differences in terms of the roots and dynamism of the respective communities: the r/todayilearned subreddit acts as a reference model. Created in 2008, it is by far the most active, with 40M members and a large number of comments that testify to its dynamism. The German equivalent, created in 2014, has 271k members and a much lower number of comments (many posts are not commented on). Finally, the corresponding French subreddit is much more recent, with slightly fewer members than the German community, but a very large number of comments nonetheless. This high number of posts is due in particular to the introduction in May 2023 of an automatic moderation system that stimulates user participation. The AJA-Bot is an automated bot used in r/Aujourd’huiJ’aiAppris. Its main function is to track and report engagement with posts using the AJA-mètre system: users comment “AJA” (Aujourd’hui j’ai appris / Today I learned) if they actually learned something new or JLS (Je le savais / I knew it) if they already knew the information. Every six days, the AJA-Bot scans the comments, counts the votes, and updates a table showing the distribution of responses (Jahver et al. 2019). The three subreddits share a similar reddiquette that enforces strict rules to maintain the quality of shared knowledge. Posts must be accurate, verifiable, and supported by sources, avoiding opinions, anecdotes, or misleading claims. Each title must begin with TIL (or HLI/AJA), be descriptive, concise, and specific. These guidelines help maintain a clear and structured format, making it easier for users to browse and absorb interesting facts efficiently. For this

analysis, we compiled only “hot posts” from Reddit – those with the highest engagement, meaning they are getting upvotes and comments at a high rate.

This study employs a mixed-methods approach, combining qualitative and quantitative analyses to investigate the emergence of DT on Reddit. We focus on the explicit transmission of knowledge, examining the most frequent cooccurrences of the verbs *learn*, *know* and *understand* – alongside their German equivalents (*lernen*, *wissen*, *verstehen*) and French counterparts (*apprendre*, *savoir*, *comprendre*). These verbs were selected due to their high frequency and cross-linguistic comparability in terms of syntax and semantics. The analysis made with the textometry software TXM (Heiden 2010) focuses on several key linguistic elements: the verb forms used, particularly the tenses employed; the subject pronouns, which indicate who is positioned as the learner or knower; the presence of negation; the prepositions associated with verbs, shedding light on semantic functions; and finally, the nouns that collocate with these verbs, illustrating how knowledge-related concepts are framed.

## 4. Results

For the sake of clarity, we present the results illustrated by verb profile. The numbers in the tables indicate the total occurrences, while the percentages are calculated based on the number of cooccurrences within the dataset. Extremely low-frequency values are excluded from the tables, as they lack analytical significance.

### 4.1 Learning on Reddit

	Learn (294)	Lernen (72)	Apprendre (361)
<b>Forms</b>	5	7	13
<b>Tenses</b>	past (57.8%)	past (83.3%)	past (62.3%)
<b>Subject pronouns</b>	I (70.4%) you (28.2%)	Ich (72.2%)	je (71.4%) tu (14.1%)
<b>Negation</b>	5.1%	-	-
<b>Time adv.</b>	44.5%	62.5%	20.7%
<b>Prepositions</b>	about (26.8%) from (14.2%)	über (23.6%)	sur (13.5%)
<b>Other co-occurrences</b>	school (8.5%)	-	école (3.6%)

Table 2: Profile of *learn*, *lernen*, *apprendre*

The analysis reveals that the verb *learn* is predominantly

used in the past tense, with the first-person pronoun being the most frequent subject. This pattern aligns well with the nature of these Reddit posts, which are supposed to recount personal experiences of acquiring knowledge (“I learned about these in Count of Monte Cristo!”). The tendency to use past-tense forms highlights the retrospective nature of knowledge sharing on the platform. Additionally, negation of *learn* is relatively rare, suggesting that users primarily employ the verb to affirm the acquisition of new information rather than to express lack of knowledge. This reinforces the positive framing of learning experiences within the corpus. Another notable trend is the frequent presence of time adverbials, indicating that users often contextualize their learning experiences within specific temporal references (“Wow you only learned this today? I learned it 3 days ago when it was first posted in this sub”; “Als ich acht war hab ich gelernt<sup>1</sup>...”). This suggests an emphasis not only on what was learned but also on when and under what circumstances the knowledge was acquired, providing a richer narrative structure. The most frequent prepositions co-occurring with the verb *learn* primarily indicate the object of learning, specifying what knowledge is acquired. In English, there is a notable use of the preposition *from* to indicate the source of information (“I first learned about this from an episode of JAG”). A particularly interesting contrast lies in references to school as a place of learning, which appear frequently in English, somewhat less in French, and are absent in German. This connection to formal education is expressed through terms such as *grade*, *school*, *student* (“I thought this was common knowledge? I learned this in sixth grade”; “Rien d’incroyable lorsque l’on a étudié les astres en collège<sup>2</sup>”).

### 4.2 Knowing on Reddit

	Know (1,053)	Wissen (97)	Savoir (822)
<b>Forms</b>	5	7	24
<b>Tenses</b>	past (30.3%)	past (14.4%)	past (25.5%)
<b>Subject pronouns</b>	I (56.8%) you (25.8%)	Ich (70.1%) du (16.4%)	je (19.9%) tu (17.2%)
<b>Other subjects</b>	people (11.8%)	-	on (12.5%)
<b>Negation</b>	27.8%	48.4%	53.8%

Table 3: Profile of *know*, *wissen*, *savoir*

The analysis indicates that the verb *know* is predominantly used in the present tense, emphasizing its role in expressing immediate knowledge or awareness rather than past learning experiences. In English and German, the first- and second-person singular pronouns are overwhelmingly the

<sup>1</sup> When I was eight, I learned... (All translations are mine)

<sup>2</sup> Nothing incredible when you’ve studied the stars in school.

most frequent subjects, highlighting how this verb is central to interactive exchanges where users explicitly articulate knowledge (“I did not know this, but I posted that article”; “Das wusstest du nicht? Echt? Ist doch Allgemeinbildung!<sup>3</sup>”). In contrast, French exhibits a broader distribution of subjects, suggesting a more varied use of *savoir* in different contexts. A distinctive pattern in French is the use of the impersonal pronoun *on*, which echoes English indefinite forms such as *people*, *folks*, *someone*, *anyone*, referring to shared knowledge or common beliefs (“on sait souvent pas qu’une pseudoscience en est une !<sup>4</sup>”; “More folks should know about Tom Lehrer”). This collective framing contrasts with the more individualized expression of knowledge observed in German. Regarding negation, approximately half of the occurrences in German and French involve negated forms, marking a strong presence of epistemic uncertainty of contradiction within discussions. In English, however, negation appears less frequently, suggesting a more affirmative use of *know* in discourse. Finally, interrogative constructions incorporating *know* appear in both English (18%) and French (12%), indicating its function in questioning the boundaries of knowledge. This pattern suggests that, beyond its declarative use, *know* plays a significant role in negotiating understanding within digital interactions (“mais savais tu que l’anxiété les crises d’angoisses de panique peuvent ressembler à un infarctus ?<sup>5</sup>”; “How are there still people who don’t know this?”).

### 4.3 Understanding on Reddit

	Understand (143)	Verstehen (32)	Comprendre (355)
<b>Forms</b>	4	5	15
<b>Tenses</b>	past (9%)	past (31.2%)	past (32.6%)
<b>Subject pronouns</b>	I (66.4%) you (19.5%)	ich (31.2%) du (21.8%)	je (64.2%) tu (17.4%)
<b>Other subjects</b>	people (15.38%)	man (25%)	gens (4.78%)
<b>Negation</b>	48.9%	46.8%	55.4%

Table 4: Profile of *understand*, *verstehen*, *comprendre*

The analysis shows that the verb *understand* is predominantly used in the present tense, emphasizing its role in expressing immediate comprehension within discussions. Unlike *learn*, which is mainly used in past-tense narratives, *understand* is more dynamic, appearing frequently in interactive exchanges where users affirm or

question their grasp of a topic. This aligns with the broader diversity of subjects, with both first and second-person singular pronouns playing a significant role, reflecting direct engagement between users as they discuss who has understood a given concept (“Once you understand the power of this idea...”). A particularly notable trend in English and especially German is the use of subjects referring to unspecified groups (“Schon klar, man kann nur gegen Bitcoin sein, wenn man es nicht verstanden hat<sup>6</sup>”; “People never understand why”). These constructions point to the presence of shared assumptions within the community, where understanding is framed as collective knowledge rather than an individual realization. Across all three languages, nearly half of the occurrences include negation, demonstrating a clear interplay between understanding and misunderstanding (“I’ve read the method for determining the Easter date 10 times over and I still don’t understand it”). In English, the preposition *about* (13.2%) is the only one that consistently co-occurs with *understand*, marking the specific object of comprehension (“This is what I don’t understand about the whole ‘nepo baby’ thing of demonizing people”). This contrasts with German and French, where similar structures appear less frequently, indicating potential cross-linguistic differences in how understanding is framed grammatically.

## 5. Conclusion

The *r/todayilearned* subreddit offers a compelling example of the emergence of digital DT and the influence of the English-speaking sphere on other cultural and linguistic communities. This influence is reflected in the structural similarities of interactions, particularly through the adoption of a similar reddiquette. By analyzing statements that explicitly articulate knowledge acquisition and transmission, focusing on the cooccurrences of the verbs *learn*, *know*, and *understand*, the study has established a discursive profile that confirms the development of shared linguistic practices. The verb *learn* (and its counterparts) is primarily used to affirm knowledge acquisition by the speaker within a clearly defined temporal framework, reinforcing its narrative function in digital exchanges. In contrast, *know* and *understand* are more central to direct interactions between users, often referring to shared knowledge or collective comprehension. Additionally, the analysis highlights linguistic variations, suggesting that while global communicative norms are being formed, local linguistic distinctions still shape the expression of knowledge. A valuable next step in refining these findings would be to examine the modality of these verbs, assessing how users hedge, emphasize, or nuance knowledge claims within digital discourse.

<sup>3</sup> You did ’nt know that? Really? It’s general knowledge.

<sup>4</sup> We often don’t know that pseudoscience is a science.

<sup>5</sup> But did you know that anxiety and panic attacks can resemble a

heart attack?

<sup>6</sup> Of course, people can only be against bitcoin, if they don’t understand it.

## 6. References

- Anderson, K.E. (2015). Ask Me Anything: What Is Reddit? *Library Hi Tech News*, 32(5), pp. 8–11.
- Botzer, N.; Shawn, G. and Weninger, T. (2023). Analysis of Moral Judgment on Reddit. *IEEE Transactions on Computational Social Systems*, 10(3), pp. 947–957. <https://doi.org/10.1109/TCSS.2022.3160677>.
- Buntain, C. and Golbeck, J. (2014). Identifying social roles in reddit using network structure. *Proceedings of the 23rd International Conference on World Wide Web*. New York: Association for Computing Machinery, pp. 615–620. <https://doi.org/10.1145/2567948.2579231>.
- Choi, D.; Jinyoung, H.; Taejoong, C.; Yong-Yeol, A., Byung-Gon, C. and Taekyoung Kwon, T. (2015). Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors. *Proceedings of the 2015 ACM on Conference on Online Social Networks*. New York: Association for Computing Machinery, pp. 233–243. <https://doi.org/10.1145/2817946.2817959>.
- Datta, S. and Adar, E. (2019). Extracting Inter-Community Conflicts in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, pp. 146–157.
- Dias, D. (2024). Sprachkontakte in den sozialen Medien: Übernahme von Diskursmustern. In S. Schwerter, N. Rentel & B. Meisnitzer (Eds.), *Mehrsprachigkeit. Herausforderungen, Sprechereinstellungen und mediale Erscheinungsformen*. Hannover: Ibidem, pp. 95–117.
- Eckkrammer, E.M. (2019a). Textlinguistik und Digitalität: eine Diskussion. In N. Janich (Ed.), *Textlinguistik: 15 Einführungen und eine Diskussion*. Tübingen: Narr Francke Attempto, pp. 34–66.
- Eckkrammer, E.M. (2019b). Genre Theory and the Digital Revolution: Towards a Multidimensional Modal of Genre Emergence, Classification and Analysis. In A. Brock, J. Pflaeging, & P. Schildhauer (Eds.), *Genre Emergence*. Berlin: Peter Lang, pp. 163–189.
- Fan, A.; Jernite, Y., Perez, E., Grangier, D., Weston, J. and Auli, M. 2019. ELI5: Long Form Question Answering. *arXiv*. <https://doi.org/10.48550/arXiv.1907.09190>.
- Fix, U. (2006). Was heißt Texte kulturell verstehen? Ein- und Zuordnungsprozesse beim Verstehen von Texten als kulturellen Entitäten. In H. Blühdorn, E. Breindl, & U.H. Waßner (Eds.), *Text – Verstehen: Grammatik und darüber hinaus*. Berlin: De Gruyter, pp. 254–276.
- Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*. Sendai, Japan, pp. 389–398. Retrieved from [http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24\\_sheiden.pdf](http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf)
- Jhaver, S.; Birman, I.; Gilbert, E. and Bruckman, A. (2019). Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5): <https://doi.org/10.1145/3338243>.
- Kabatek, J. (2015). Wie kann man Diskurstraditionen kategorisieren? In E. Winter-Froemel, A. Lopez Serena, A. Octavio de Toledo & B. Franck-Job (Eds.), *Diskurstraditionelles und Einzelsprachliches im Sprachwandel*. Tübingen: Narr, pp. 51–65.
- Kabatek, J. (2014). Genre textuel et traditions discursives. In C. Gérard, & R. Missire (Eds.), *Coseriu aujourd’hui : linguistique et philosophie du langage*. Limoges: Lambert Lucas, pp. 195–206.
- Kaiser, D. (2003). Zum Einfluss angelsächsischer Diskurstraditionen auf die Wissenschaftssprache in Hispanoamerika. In H. Aschenberg & R. Wilhelm (Eds.), *Romanische Sprachgeschichte und Diskurstraditionen*. Tübinger Beiträge zur Linguistik 464. Tübingen: G. Narr, pp. 183–201.
- Koch, P. (1997). Diskurstraditionen: zu ihrem sprachtheoretischen Status und ihrer Dynamik. In B. Frank, T. Haye, & D. Tophinke (Eds.), *Gattungen mittelalterlicher Schriftlichkeit*. Tübingen: Günter Narr, pp. 43–79.
- Mercieca, B. (2017). What Is a Community of Practice? In J. McDonald & A. Cater-Steel (Eds.), *Communities of Practice: Facilitating Social Learning in Higher Education*. Singapore: Springer, pp. 3–25.
- Müller-Lancé, J. (2022). Discourse Traditions, Multimodality and Media Studies. In E. Winter-Froemel & Á.S. Octavio de Toledo y Huerta (Eds.), *Manual of Discourse Traditions in Romance*. Berlin/Boston: De Gruyter, pp. 767–780.
- Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C. and Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2): <https://doi.org/10.1177/205630512111019004>.
- Rosemeyer, M. (2022). Discourse Traditions and Corpus Linguistics. E. Winter-Froemel & Á.S. Octavio de Toledo y Huerta (Eds.), *Manual of Discourse Traditions in Romance*. Berlin/Boston: De Gruyter, pp. 647–668. <https://doi.org/10.1515/9783110668636-032>.
- Werle, L. M. (2025). Textsorten auf französischsprachigen virtuellen Gedenkstätten: Zwischen Transfer und Hybridisierung. *Neuphilologische Mitteilungen*, 126(1): 92–119. <https://doi.org/10.51814/nm.145497>.
- Winter-Froemel, E. (2020). Digitale Kommunikationsformen und Diskurstraditionen zwischen Nähe und Distanz. In B. Kluge, W. Mihatsch, & B. Schaller (Eds.), *Kommunikationsdynamiken zwischen Mündlichkeit und Schriftlichkeit*. Tübingen: Narr, pp. 81–102.
- Winter-Froemel, E.; Octavio de Toledo y Huerta, Á.S.; G. Holtus, and Sánchez-Miret, F. (Eds.) (2022). *Manual of Discourse Traditions in Romance*. Berlin/Boston: De Gruyter.

# Evaluating Different Methods for Building Specialized Corpora: A Case Study on the German Discourse on AI

Bruno Brocai, Janine Dengler

University of Heidelberg, Department of German Language and Literature  
E-mail: bruno.brocai@gs.uni-heidelberg.de, janine.dengler@gs.uni-heidelberg.de

## Abstract

Specialized thematic corpora play an important role in discourse linguistics. Yet methods for constructing them remain underexplored. Typically, queries (lists of topical words) are used to find suitable documents, but suboptimal queries can retrieve irrelevant documents or omit relevant ones. In this paper, we evaluate several approaches for query generation to construct specialized corpora: (1) subjective, manual query generation; (2) corpus-driven methods utilizing RQTR, keyness, and word collocation analyses; and (3) prompting large language models. We apply these methods to a categorized German pilot corpus to find texts about artificial intelligence. Our results demonstrate that corpus-driven and LLM-based methods consistently outperform subjective approaches. Additionally, employing stricter inclusion thresholds — rather than the common practice of including all documents with at least one query match — significantly enhances corpus relevance. These findings advance the understanding of query formulation strategies, providing practical guidance for researchers aiming to construct replicable, high-quality corpora.

**Keywords:** Corpus linguistics, corpus creation, discourse linguistics, query expansion, information retrieval

## 1. Introduction

The creation of specialized thematic corpora is particularly relevant for discourse-linguistic approaches (Koester, 2022; Baker, 2023, 56-62), which involve collecting texts that address the discourse in question. However, the problem is how to identify relevant texts to include in the corpus. In corpus linguistics, researchers usually formulate a list of topically related terms (search queries) to search for in a database. Many different approaches for creating a query exist. Often, researchers select terms introspectively, through a literature review, or using corpus-driven methods, such as keywords or collocation analysis. These methods are often integrated in restricted-access databases, for which Gabrielatos (2007) has outlined a more robust approach for building queries by combining several methods (RQTR). But so far, these approaches have neither been systematically compared nor sufficiently evaluated. Furthermore, many corpus resources, especially for computer-mediated communication (CMC), are non-restricted, with full corpora accessible as public datasets or via web crawling, which makes more methods applicable.

The challenge in creating a specialized corpus is to obtain as many thematically relevant texts as possible. Certain methods may result in the inclusion of many texts that fall outside the thematic focus and thus the research interest (resulting in low precision). Conversely, other methods may retrieve only a fraction of all relevant texts in the corpus (resulting in low recall).<sup>1</sup>

Therefore, we evaluate various methods and metrics for formulating a search query to select relevant written texts from a database and build a thematic corpus, and assess how different approaches influence precision and recall. Besides

optimizing query methods, we aim to demonstrate how various parameters and decisions impact corpus construction. We are thus positioning our work in the field of corpus-assisted discourse studies (Mautner, 2022, 250) and also query expansion, which, as a part of information retrieval, is often used to improve searches on the internet (Azad and Deepak, 2019; Gauch et al., 1999). The field of information retrieval offers many more methods and metrics than we are able to discuss here, including manual query expansion, automatic expansion with neural networks (Word Embeddings) (Roy et al., 2016; Imani et al., 2018), prompting LLMs (Jagerman et al., 2023; Lei et al., 2024), and interactive expansion, which combines both approaches (Azad and Deepak, 2019, 4).

## 2. Methodological Approach

To construct a specialized thematic corpus, our procedure is as follows:

1. **Identify a topic** of interest along with a set of **core terms** representing its key concepts.
2. **Create a fully accessible, non-thematic corpus**, but already specialized in other aspects, e.g., time period, if desired.
3. **Develop a query** that represents the topic of interest by applying corpus-linguistic methods to the non-thematic corpus.
4. **Create the thematic corpus** by using the query to search for thematically relevant documents in the non-thematic corpus.

To evaluate different methods to formulate a query, we construct an example corpus on the topic of artificial intelligence (AI). For this purpose, we first create a full-access non-thematic corpus by crawling articles from three German websites: Informatik Aktuell (INFO), Spektrum.de (SPEKTRUM), and ZEIT ONLINE (ZEIT), which have a high, medium, and low level of specialization on IT topics, respectively. INFO is a specialist journal for programmers and other professionals in the IT industry, SPEKTRUM is

<sup>1</sup>Precision (Prec) measures the proportion of all retrieved relevant items to the number of all retrieved items, indicating the accuracy of the results. Recall (Rec) measures the ratio of all relevant items retrieved to all relevant items, reflecting the completeness of the results (Buckland and Gey, 1994, 12f.).

a science magazine that addresses specialists and informed lay people, and ZEIT is an online newspaper that reports on current events and targets the broader public. All three websites report on news about AI. We retrieved all available articles that were published in February in each of the years from 2020 to 2025<sup>2</sup>, a total of 6737 texts.

We use several high-recall methods (cf. section 3.) to develop queries and search through our non-thematic corpus to find as many texts as possible that have AI as a topic. We retrieved 424 texts and employed pooling (Sparck-Jones and Rijshergen, 1975) without fixed depth, which means that the 6313 documents that were not retrieved by any method are automatically categorized as not discussing AI. The other 424 texts were then categorized by the authors and two students with a linguistic background. The categories are: (1) AI is the main topic; (2) AI is a side topic, i.e., an interesting aspect of the main topic; (3) AI is not discussed. The latter includes texts in which AI-related terms appear, but no substantial statement about the topic is made.<sup>3</sup> The categorization and training were improved until our best agreement levels, an ordinal agreement of 0.870 and nominal agreement of 0.690 (Krippendorff’s alpha), were reached. Table 1 provides an overview of the size of the corpora and the documents in which AI was either categorized as the main topic or side topic.

Source	Docs.	(1) Main Top.	(2) Side Top.
INFO	64	5	6
SPEKTRUM	892	16	10
ZEIT	5781	53	25
Sum	6737	74	41

Table 1: Size of the Non-Thematic Corpus and Number of AI-Related Articles it Contains

### 3. Methods

#### 3.1. Subjective Query Generation

The first approach we use to generate a search query, which also serves as the **baseline**, is to simply use the term(s) that most accurately describe the topic. In our case, these are the German terms *künstliche Intelligenz* (= artificial intelligence) and the abbreviation *KI* (= AI).

Another common query creation method is **subjective introspection**. The researchers familiarize themselves with a topic and create a query based on their knowledge. To emulate this procedure and avoid introducing our bias into the query, we asked a linguistic expert whose research focuses on AI discourses to write us a query.

<sup>2</sup>All resources are made available on GitHub.

<sup>3</sup>E.g., *Premierminister Justin Trudeau sagte am Rande eines Treffens zu künstlicher Intelligenz in Paris, seine Regierung werde kanadische Arbeiter schützen.* (Translation by the authors: Prime Minister Justin Trudeau said on the sidelines of a meeting on artificial intelligence in Paris that his government would protect Canadian workers). Source accessed on Feb 11, 2025.

#### 3.2. Corpus Linguistic Metrics

**Word collocation** metrics indicate how strongly a word is associated with another. The association can be analyzed at different windows around the study term. We use both the paragraph level and a window of 5 words before and after *künstliche Intelligenz* or *KI* (our baseline) to calculate collocations. Specifically, normalized point-wise mutual information (**nPMI**) (Bouma, 2009), which ranks words highly when they cooccur exclusively, and **LogDice** (Rychlý et al., 2008), which ranks words highly when they cooccur both exclusively and frequently, are used.

We use **keywords** as the second quantitative procedure to generate query terms. Our procedure is as follows: First, we generate a subcorpus with the baseline method. Then, the keyness of each term in the subcorpus is calculated, using the entire corpus as the reference corpus. We test an effect size keyness metric — Odds Ratio (**OR**) — and a more common statistical significance metric, namely log likelihood (**LL**) (Pojanapunya and Watson Todd, 2018). **OR** indicates the amount of overrepresentation in the subcorpus, while **LL** is affected by overrepresentation and frequency.

The **RQTR** score identifies words that strongly cooccur with a set of base terms (**BT**) while occurring rarely without them. At least two base terms are needed to calculate the baseline value that indicates significant cooccurrence. Terms with a higher cooccurrence with the base terms than the baseline are considered good query candidates. To better validate the method and its different trade-offs, we used different base terms, namely *künstliche Intelligenz* and *KI* (**BT1**); *künstliche Intelligenz*, *Chatbot* and *Roboter* (**BT2**). The **RQTR** values are then filtered by keyness (**keyn.**), which is what Gabrielatos (2007) recommends as the best procedure, dropping terms with low statistical significance ( $LL < 15.13$ ) (Rayson et al., 2004).

For all corpus linguistic metrics, these settings were used: Laplace smoothing ( $\alpha = 0.0001$ ) to avoid division by zero; terms must appear at least in 5 different documents in the AI subcorpus; stop word filtering. We collect metrics for unigrams and bigrams (except collocations). From the resulting list, the 50 highest-ranking terms are taken as a query. Using value-based cutoffs (e.g.  $RQTR_n > 0$ ) resulted in queries with such low precision that they were not productive; furthermore, top 50 is good for comparability.

#### 3.3. Language Models

Generative language models are state-of-the-art for many use cases, and we hypothesize that they can perform well on corpus building. Two approaches were tested with Claude Sonnet 3.7 (Anthropic, 2025): First, **the LLM was instructed to generate a** (regular expression) **query** with high precision and recall.

We also use **LLM categorization** to get a rank for each document, similar to our classification process. This is not a query-based approach. Instead, the LLM is prompted with our categorization guidelines together with one document at a time and asked to give a classification. This is repeated for all texts in the corpus. In addition to Claude, we also tested GPT-4o (OpenAI, 2024) for classification.

	Threshold: 1			Threshold: 5			Threshold: 9,000 PMW		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Baseline	0.448	0.932	0.605	0.946	0.473	0.631	<b>1.000</b>	0.473	0.642
Subjective (1)	0.078	<b>1.000</b>	0.145	0.547	0.784	0.644	0.548	0.689	0.611
Subjective (2)	0.294	<b>1.000</b>	0.454	0.824	0.757	0.789	0.925	0.662	0.772
LogDice (Window: 5)	0.018	0.986	0.035	0.090	0.932	0.164	0.143	0.919	0.248
nPMI (Window: 5)	0.131	0.946	0.230	0.723	0.635	0.676	0.884	0.514	0.650
LogDice (Paragraph)	0.018	<b>1.000</b>	0.036	0.075	0.932	0.138	0.097	0.932	0.176
nPMI (Paragraph)	0.210	0.946	0.344	0.838	0.838	0.838	0.902	0.743	0.815
Keyn. LL (BT1)	0.024	0.986	0.047	0.111	0.932	0.198	0.137	0.946	0.239
Keyn. OR (BT1)	0.335	0.946	0.495	0.840	0.851	0.846	0.892	0.784	0.835
Keyn. OR (BT2)	0.216	0.973	0.354	0.724	0.851	0.783	0.817	0.784	0.800
RQTR LL (BT1)	0.347	0.946	0.507	0.855	0.878	<b>0.867</b>	0.935	0.784	0.853
RQTR LL (BT2)	0.232	0.973	0.375	0.767	0.892	0.825	0.855	0.797	0.825
Claude Sonnet 3.7 Query	0.241	0.946	0.384	0.812	0.878	0.844	0.934	0.770	0.844

Table 2: Binary Evaluation Scores for Main Topic Retrieval

## 4. Evaluation

Table 2 reports the results of the evaluation regarding precision, recall, and F1. We pay particular attention to the F1 score, the harmonic mean of precision and recall (Sasaki, 2007), as it provides a single metric that balances the trade-off between including the highest number of texts (recall) and ensuring that the texts are relevant (precision). This balance is crucial for constructing a corpus that is both comprehensive and focused.

The results are given for the absolute/aggregate thresholds of 1 and 5 (meaning that a document needed at least 1 or 5 hits to be included in the thematic corpus), and for a relative frequency threshold of 9,000 per million words (PMW), a metric that corrects for article length.

### 4.1. Thresholds

Figure 1 shows the relationship between the minimum number of times a search term appears in a document to be included in the thematic corpus and the F1 score.

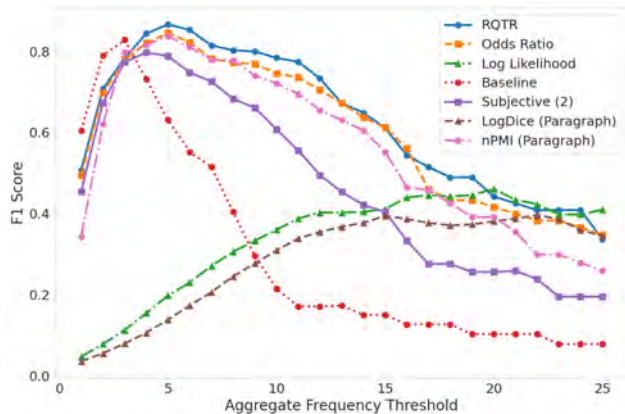


Figure 1: F1 Scores by Aggregate Frequency Threshold (Main Topic)

One observation is that a minimum of 1 is never the optimum — for any method. This is a key observation because many tools (e.g., Sketch Engine’s Kilgarrieff et al. (2004)

concordance subcorpus builder) default to returning all documents with at least one hit. But our data suggests that a more careful threshold selection should be considered an essential step in corpus construction.

As the graph shows, there are two types of methods: those peaking around lower thresholds of around 5, and those peaking at higher thresholds. However, these latter peaks are still significantly lower and are typical for poorly performing frequency-sensitive methods (cf. 4.2.).

Of course, our specific threshold optima are not generalizable. Longer text genres probably need higher absolute thresholds and vice versa. A relative threshold could solve this, but surprisingly, we find that using a PMW threshold does not improve the results. The reported threshold of 9,000 PMW (Table 2) is an optimum similar to the threshold of 5 for many methods, but the peak F1 scores are slightly lower and are more skewed toward precision.

### 4.2. Quantitative Evaluation

The **baseline** approach has a relatively high F1 score because of the method’s high precision, as it includes only terms that are central to the discourse. However, this comes at the cost of lower recall. The **subjective** queries perform better in terms of recall. Nevertheless, the precision was problematic for **subjective 1**. The reason is that the query included *\*bot*, which not only matched words like *Chatbot*, but also German words that are not related to AI, such as *Verbot* (= ban). To address this issue, a second run (**subjective 2**) was performed and explicitly excluded this wildcard, but the resulting F1 is still below baseline.

For the corpus linguistic metrics, **LogDice** and **LL** underperform. Both are affected by the frequency of the observed phenomenon.

But exclusivity with the core terms seems to be far more important than general frequency when compiling a query, as shown by **RQTR**, which proves to be the best method, balancing precision and recall. **OR**, with the second highest F1, could also be an objective way to build a corpus. It is a shorter pipeline, included in several corpus tools, and does not require two base terms (especially useful for discourses centering around one term). Given that these two



methods were the most promising, we also tried varying the base terms. This only resulted in worse F1 scores, albeit recall was slightly higher. Overall, sticking to only the most pertinent terms seems best.

The **LLM-generated query** performs well, but a bit worse than the highest performing methods. Given this result, and its lack of transparency, the corpus-driven methods prove superior.

**Using LLMs to categorize** (cf. Table 3) shows that the models have a strong bias towards precision, at least with our prompt. This procedure does not outperform corpus-driven methods, despite being more resource-intensive. Given the high compute and/or monetary cost of prompting LLMs with thousands of documents, a faster and cheaper method is to first reduce the corpus to a set of candidate documents with a high recall method (e.g. with OR at a threshold of 1) and then prompt only with those documents. With our data, this does not impact performance.

	Prec	Rec	F1	Cost
Claude Sonnet 3.7	0.982	0.723	0.837	30.89€
OR → Claude	0.982	0.723	0.837	0.93€
OR → GPT-4o	<b>0.983</b>	<b>0.770</b>	<b>0.864</b>	0.55€

Table 3: Evaluation of LLMs in Main Topic Classification

Prompting should also work better if a categorized test set (such as the one we have created) is used to refine the prompt. However, since we set out to evaluate methods that work ‘out of the box’, we refrained from doing so. But if such categorizations are available, prompting is likely to perform better than indicated here. Given the differing performance of different models, such a test set is also important for model selection.

### 4.3. Qualitative Evaluation

The frequency-sensitive metrics **LL** and **LogDice** prioritize words such as *Software* or *Text*, which are associated with the general topic but also appear frequently outside of it. Exclusivity metrics, namely **PMI**, **OR** and **RQTR**, do not rank those words highly and include more low-frequency, high-pertinence terms such as *OpenAI-Chef Sam* or *ausspucken* (= to spit out). This distinction largely explains why exclusivity metrics outperform other queries.

The **subjective query** and the **LLM query** also include terms that seem highly pertinent but turn out to be more ambiguous than intended. The problematic *\*bot* was already mentioned. Another example is *Algorithmus*, which causes too many false positives. Similarly, *Claude* and *Gemini* are language model families but also match a French last name and astronomical phenomena, respectively.

A notable observation is that some low-frequency terms are highly pertinent but not easily identified through introspection, such as *ausspucken*. Additionally, other terms may seem topic-specific but prove too broad in practice, such as *Algorithmus*. This again underscores the value of corpus-driven methods.

## 5. Excursus: Including Side Topic

The methods perform similarly when also looking for documents where AI is a (2) side topic (Figure 2). A threshold of 1 is still not the optimal threshold. But because side-topic documents less frequently mention AI terms, a lower threshold, namely 2 for the baseline and 2-3 for other methods performs better. Again, OR and RQTR are the best methods to create a query, with OR slightly ahead in terms of F1. Overall, the lexical indicators for main and side topics seem to be the same.

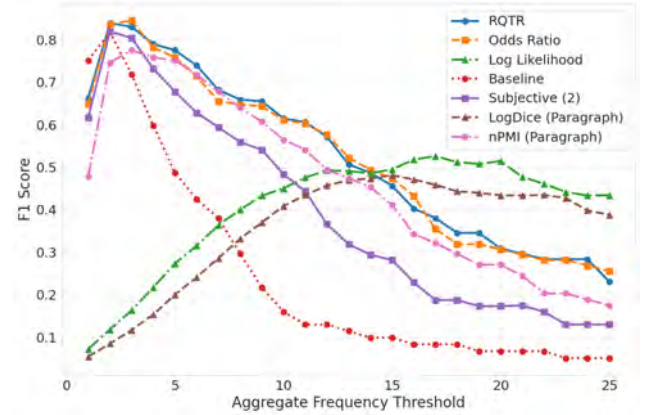


Figure 2: F1 Scores by Aggregate Frequency Threshold (Main & Side Topic)

## 6. Conclusion & Limitations

Our evaluation of methods for creating queries for corpus creation indicates that the best queries contain words that are highly pertinent to the topic. This is already somewhat achieved with the high-precision baseline of picking only the discourse-central terms, but methods such as OR and especially RQTR can enhance a query and improve recall. Notably, many good search terms are not introspectively salient, and subjective approaches can also result in lower precision, i.e., the inclusion of texts that do not address the topic under investigation. Metrics sensitive to word frequency, such as LL and LogDice, perform even worse in terms of precision. LLMs do not currently outperform corpus-linguistic metrics but could be a tool for increasing precision. However, there are other ways of improving precision, especially increasing the threshold. Overall, thresholds emerge as an important consideration, as low frequency thresholds significantly lower precision and high ones reduce recall.

It should be noted that our results are based on a pilot study on the discourse on AI, meaning that the effectiveness of the evaluated methods may vary in other discourses, especially for discourses based on more abstract concepts, such as “safety” or “love”. Also, there are methods that were not examined here, e.g., more complex query expansion or vector-based similarity search. For these reasons, more research needs to be done in this area. We invite future work that reflects on and evaluates different methods for corpus creation and thus contributes to more objective and replicable research practices.



## 7. Acknowledgements

We would like to thank Marcel Kückelhaus, Marie-Sophie Faust and Melissa Brezina for their help during the research process and Mayumi Ohta, Maria Nuralieva, Julia Goetz, Maria Becker and the anonymous reviewers for their valuable feedback on our paper.

## 8. References

- Anthropic. (2025). *Claude 3.7 Sonnet and Claude Code*. <https://www.anthropic.com/news/claude-3-7-sonnet>.
- Azad, H. K. and Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5), pp. 1698--1735.
- Baker, P. (2023). *Using corpora in discourse analysis*. London, New York, Oxford, New Delhi, Sydney: Bloomsbury Academic.
- Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the Biennial GSCS Conference*. Tübingen: Gunter Narr Verlag, pp. 31--40.
- Buckland, M. and Gey, F. (1994). The Relationship between Recall and Precision. *Journal of the American Society for Information Science*, 45(1), pp. 12--19.
- Gabrielatos, C. (2007). Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME Journal*, 31, pp. 5--43.
- Gauch, S., Wang, J., and Rachakonda, S. M. (1999). A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems*, 17(3), pp. 250--269.
- Imani, A., Vakili, A., Montazer, A., and Shakery, A. (2018). *Deep Neural Networks for Query Expansion using Word Embeddings*. arXiv:1811.03514 [cs].
- Jagerman, R., Zhuang, H., Qin, Z., Wang, X., and Bender-sky, M. (2023). *Query Expansion by Prompting Large Language Models*, May. arXiv:2305.03653 [cs].
- Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EU-RALEX International Congress*. Lorient, France, pp. 105--116. URL: <http://www.sketchengine.eu>.
- Koester, A. (2022). Building small specialised corpora. In O’Keeffe, A. and McCarthy, M. J. (Eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2nd edition. pp. 48--61.
- Lei, Y., Cao, Y., Zhou, T., Shen, T., and Yates, A. (2024). *Corpus-Steered Query Expansion with Large Language Models*, February. arXiv:2402.18031 [cs].
- Mautner, G. (2022). What can a corpus tell us about discourse? In O’Keeffe, A. and McCarthy, M. J. (Eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2nd edition. pp. 250--262.
- OpenAI. (2024). *Hello GPT-4o*. <https://openai.com/index/hello-gpt-4o/>.
- Pojanapunya, P. and Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), pp. 133--167.
- Rayson, P., Berridge, D., and Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *JADT 2004: 7es Journées internationales d’Analyse statistique des Données Textuelles*, volume 2. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 926--936.
- Roy, D., Paul, D., Mitra, M., and Garain, U. (2016). *Using Word Embeddings for Automatic Query Expansion*. arXiv:1606.07608 [cs].
- Rychlý, P., Sojka, P., and Horák, A. (2008). A Lexicographer-Friendly Association Score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*. Brno, Czech Republic, pp. 6--9.
- Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, Vol. 1, no 5.
- Sparck-Jones, K. and Rijshergen, C. J. V. (1975). *Report on the Need for and Provision of an ‘Ideal’ Information Retrieval Test Collection*. Number 5266 in British Library Research and Development Report. Cambridge: University Computer Laboratory.

## 9. Appendix

In the following, we provide a selection of our search queries. All search queries can be found in our GitHub repository: <https://github.com/GS-Uni-Heidelberg/Paper-SpecializedCorpora>.

Whitespace is added to the raw queries for readability.

### 9.1. Subjective Queries

Matches lemmas with a (Python) regular expression.

künstlich Intelligenz | KI | Algorithm[a-z]{0,2} | Roboter | Robotik | autonom[a-z]{0,2} | Waffensystem[a-z]{0,2} | machine learning | maschinell[a-z]{0,2} | Lernen | Large Language Model | LLM | Artificial Intelligence | AI | Chat.?GPT[0-9a-zA-Z]{0,4} | Gemini | llama | deep fake | chatbot | griefbot | deathbot | digital afterlife | unsupervis[a-z]{0,2} | learning | artificial general intelligence | AGI | [A-Za-z]\*bot

### 9.2. nPMI (Paragraph)

Matches lemmas

Gemini | Datenmeng | AI | disruptiv | intrinsisch | abgekürzt | Techkonzern | nachbilden | KI-System | KI-Forscher | Nachvollziehbarkeit | KI-Gipfel | akzeptiert | KI-Anwendung | ChatGPT | OpenAI | KI-Programm | OpenAI-Chef | Tarnung | keimen | Potential | Pessimismus | Expertenwissen | Anfangszeit | verlernen | KI-Funktion | generativ | erreicht | Automatisierte | Nachverfolgung | Nvidia | Software-Entwicklung | ungeahnt | Bildgenerator | generiert | s. | selbstfahrend | Illustration | Doktorande | Altman | krönen | Chatbot | KI-Chatbot | KI-Modelle | Deep | zurücklehnen | Engineering | Softwarekonzern | Rechenleistung | KI-Entwicklung | künstlich Intelligenz | KI

### 9.3. LogDice (Paragraph)

Matches lemmas

ChatGPT | OpenAI | Google | Technologie | AI | Software | Microsoft | Chatbot | Datenmeng | neuronal | Nachhaltigkeit | trainieren | Deep | lernen | nachhaltig | Anwendung | entwickeln | Entwicklung | bzw. | Entwickler | [ ] | Kunst | erkennen | mithilfe | Computer | KI-System | Digitalisierung | Wirklichkeit | Bereich | verarbeiten | Einsatz | Netz | Text | Gemini | Informatik | Künstler | Arbeitsweise | Meta | programmieren | Unternehmen | dank | generiert | Datum | Bing | verbessern | Training | System | Möglichkeit | Nutzer | künstlich Intelligenz | KI

### 9.4. Log Likelihood (LL)

Matches lemmas

KI | künstlich Intelligenz | Intelligenz | künstlich | ChatGPT | OpenAI | Google | Microsoft | Chatbot | Nutzer | Bot | digital | Technologie | Unternehmen | Anwender | Bing | Text | System | Amazon | AI | Apple | Computer | Sprachmodelle | Reddit | neuronal Netz | User | Modell | Software | Intelligenz KI | Maschine | trainieren | Antwort | DeepSeek | Digitalisierung | Sprachmodell | KI-System | entwickeln | bzw. | Trainingsdat | Suchmaschine | Entwickler | maschinell | Künstler | neuronal | Ding | Artikel | IT | menschlich | Informatiker | Business | App | Tool

### 9.5. Odds Ratio (OR) – BT1

Matches lemmas

KI | künstlich Intelligenz | OpenAI | Sprachmodelle | neuronal Netz | Intelligenz KI | Sprachmodell | KI-System | Trainingsdat | groß Sprachmodelle | KI-Modelle | Chatbot ChatGPT | KI-Gipfel | KI-Forscher | KI- | Sam Altman | Websuche | KI generiert | generativ | Bildgenerator | OpenAI-Chef | KI-Modell | KI-Chatbot | mithilfe künstlich | Künstlich Intelligenz | Firma OpenAI | Robotik | KI-Funktion | KI-Entwicklung | OpenAI-Chef Sam | Hilfe künstlich | ChatGPT entwickeln | ChatGPT | Chatbot | Bot | Altman | Bing | Intelligenz | Spracherkennung | KI-generiert | Sundar | darauf trainieren | Sundar Pichai | AI | Gemini | ausspucken | Künstlich | Suchmaschine Bing | KI-Anwendung | Reddit | Anwender | Pichai

### 9.6. RQTR (+LL) – BT 1

Matches lemmas

neuronal Netz | mithilfe künstlich | großSprachmodelle | generativ | Websuche | Trainingsdat | Sprachmodelle | Sprachmodell | Sam Altman | Robotik | OpenAI-Chef Sam | OpenAI-Chef | OpenAI | Künstlich Intelligenz | KI-System | KI-Modelle | KI-Modell | KI-Gipfel | KI-Funktion | KI-Forscher | KI-Entwicklung | KI-Chatbot | KI- | KI generiert | Intelligenz KI | Hilfe künstlich | Firma OpenAI | Chatbot ChatGPT | ChatGPT entwickeln | Bildgenerator | ChatGPT | DeepSeek | Altman | KI-Anwendung | Sundar Pichai | Sundar | Pichai | Intelligenz | ausspucken | Chatbot | AI | darauf trainieren | Spracherkennung | Künstlich | Suchmaschine Bing | KI-generiert | Softwarekonzern | Bot | Bing | maschinell | künstlich Intelligenz | KI

### 9.7. LLM-Generated Query (Excerpt)

Too long to be given in full.

Matches raw text with a regular expression (Python)

```
(?ix)
\b(?:KI|AI)\b| \bkünstlich(?:e|er|en|es)?\s+Intelligenz\b|
\bartificial\s+intelligence\b|
\bmaschinell(?:e|er|en|es)?\s+Intelligenz\b|
\bKI-[a-zäöüß]+\b|
\b(?:machine|maschinell(?:e[snm]?|er))\s+learning\b|
\bdeep\s+learning\b| \btief(?:e[snm]?|er|es)\s+lernen\b|
\b(?:neural(?:e[snm]?|er|es)?|neuronal[snm]?)\s+
s+(?:netz(?:e|werk(?:e)?))\b|
\breinforcement\s+learning\b|
\b(?:un)?überwacht(?:e[snm]?|er|es)?\s+lernen\b|
\b(?:un)?supervised\s+learning\b|
\b(?:sprach|language)\s+*modell(?:e[s]?)\b|
\b(?:foundation|fundament(?:al)?)[-s]*modell(?:e[s]?)\b|
\bCNN\b|\bRNN\b|\bLSTM\b|\bGAN\b|\bNLP\b|
|\bLLM(?:s)?\b|\bAGI\b|\bASI\b|
\bnatural\s+language\s+(?:processing|understanding)\b|
\bcomputer\s+vision\b| \btransformer(?:-
?architektur)?\b| \btextgenerierung\b|
\bbilderzeugung\b| \b(?:text|sprach)[-s]+(?:zu|to)[-
\s]+bild\b| \bsprach(?:erkennung|verstehen
|verarbeitung)\b
```

# The most common features of the Albanian language used in computer-mediated communication – an overview based on corpus data

Besim Kabashi

Computational Linguistics, University of Tübingen, Germany

E-Mail: [besim.kabashi@uni-tuebingen.de](mailto:besim.kabashi@uni-tuebingen.de)

## Abstract

The language used in computer-mediated communication (CMC) is to a large extent based on the spoken language and has many of its characteristics. In addition, posts or texts in formally written standard language as well as official announcements up to commercial advertising etc. can be observed often. In some cases, it is possible to organize the texts using meta data so that they can be analyzed separately from other texts, i.e. only a certain amount can be examined concerning their certain properties. I have collected Albanian data for corpus building over several years. One of these corpora contains social media contributions from selected users as well as randomly selected contributions on Twitter (current X) as part of an Albanian corpus between 2019-2022 collected from Twitter. This contribution deals with the usage of CMC in Albanian and its sub-variants, an underrepresented language resource. The examination also includes an analysis of standard language requirements or rules and phenomena in Albanian CMC including the use of emojis.

**Keywords:** The Albanian language, Computer-mediated Communication, Usage properties, Automatic data processing

## 1. Introduction

Some characteristics of the Albanian language typical for Computer-Mediated Communication have already been discussed (e.g. Kabashi 2024). These studies concern lexical variations but also show other appearances in CMC. In contrast to this, a broader discussion of phenomena is covered in this contribution. When one reads computer-mediated-texts in Albanian, one of the interesting questions is, whether these texts align with language standards or not. This can be in different ways as well as in degree or strength, shown in the section 2.

## 2. The usage of language variants

An example of the use of language variants is *dmth qetash isha nis per ndeti edhe dej nshtator mo sisha kthy*. engl. *I would have left for the sea immediately/now and wouldn't have returned until september*. In the standard language version, the sentence would be like this *D.m.th. që tash isha nisur për në det dhe deri në shtator më s'isha kthyer*. This sentence is close to original, one could have written it a little more standard, like this *D.m.th. menjëherë isha nisur për në det dhe deri në shtator s'isha kthyer më*. Anyone familiar with the Kosovar version of Albanian can quickly recognize that it is this version / or a subversion. This can also be quickly determined using statistic or corpus-based methods. The example demonstrates that almost every word or word-form is written differently from the standard language and even the word order could be changed. Compared to this easily identifiable example, other samples in the corpus are more difficult to recognize.

### 2.1 Horizontal language variants

Horizontal language variants (geographic factors) in the corpus usually refer to geographical and dialectal characteristics. A user can be assigned to a specific place, region or geographical dialect. This means that it can be identified whether users are native and from which part of Albania (North or South) or Kosovo they come from.

### 2.2 Vertical language variants

Vertical language variants (person-related factors) in the corpus enable conclusions concerning the users' social and educational background as well as also whether a particular user has a migratory background or not. Horizontal and vertical variants can occur also combined with one another. The example above shows that the author is a native speaker and could certainly write or speak in standard without mistakes but deliberately used a sub-standard variant.

## 3. Consideration of standard language spelling and other rules

While writing in dialect or in language variants, language standards, rules, and regulations are usually not considered. Even if some users try to align with rules and regulations of standard Albanian, their language use can contain also regional or social characteristics. One example for this observation is the use of particular grammatical constructions, e. g. the Geg infinitive with *me*+Infinitiv instead of *për të*+Infinitive. Both variants differ in terms of their appearance including different grammatical rules and forms (diacritical marks in Geg, e. g. *â*, *Â*, or the non-use of the prescribed, obligatory diacritical marks *ç*, *Ç*, and *ë*, *Ë* in the standard language variant). Following samples underpin this hypothesis:

(1) Someone writes correctly in standard language but does not use diacritical marks, e. g. *Per hir te se vertetes jam shume e inatosur karshi tyre. Nuk e fsheh dot kete fakt.* should have been *Për hir të së vërtetës jam shumë e inatosur karshi tyre. Nuk e fsheh dot këtë fakt.*, engl. *To be honest, I am very angry with them. I can't hide this fact.* The author does not use diacritical marks for technical reasons, i. e. she or he may have lacked an Albanian keyboard.

(2) Following sample demonstrates that the authors do not master the standard language variant and make mistakes, e. g. *ka pas edhe tjer*, engl. *there were also others*, that, should have been *ka pasur edhe (të) tjerë*. These types of users can be further subdivided according to various criteria, mainly based on regional language characteristics and

characteristics of migratory language use. The latter shows symptoms of code mixing and code switching. Those who have not mastered the standard language tend to make spelling mistakes more often, especially concerning the rules of sounds/letters

- *ë/Ë*, e. g. *tjer* vs. *tjerë*, engl. *others*.
- *ë/Ë* vs. *e/E*, e. g. *eshtë* vs. *është*, engl. *she/he/it is*.
- *ç/Ç* (the confusion with *q/Q*), e. g. *qka* vs. *çka*, engl. *what*.
- There are also differences in the use of *xh/Xh* vs. *gj/Gj*, e. g. *rexhistro* vs. *regjistro[je]*, engl. *(to) register / register it*. Variation also occurs in terms of the use *dh* vs. *th*, *l* vs. *ll*, *r* vs. *rr*, *s* vs. *sh*, *z* vs. *zh*. Letters or sounds that sound similar are often used interchangeably in the sub-standard variation.
- The correct use of the apostrophe is also a source of differences between standard usage and dialect: e. g. *çka*, engl. *what*, vs. *ç'ka* (*me vete*), engl. *what does she/he have (with her/him)?*.

The following statistics (Albanian Tweets 2019–2022 Corpus, 3,415,860 words [213 texts], randomly selected tweets) show how often each form was used:

---

*çka* and *Çka* returned 325 matches in 146 different texts.  
*cka* and *Cka* returned 268 matches in 133 different texts.  
*qka* and *Qka* returned 709 matches in 189 different texts.  
*tjerë* and *Tjerë* returned 152 matches in 103 different texts.  
*tjer* and *Tjer* returned 98 matches in 76 different texts.

---

Table 1: Some of spelling mistakes (*çka* and *tjerë*) in Albanian Tweets (2019–2022) Corpus.

## 4. The most common features

Kabashi (2024) lists some of the most important phenomena of the CMC in the Albanian language, namely abbreviation, e. g. *flm* (short form for *ju/të/... falemnderit*, engl. *thank you*), e. g. contraction *ti* (engl. *you*) instead of *t'i* (i.e. subjunctive particle, accusative or dative object clitic), and creative spelling. I take the types as a basis.

### 4.1. Abbreviation

Abbreviations occur very often, logically because some words are used quickly and repeatedly. The most well-known and commonly used abbreviation appears apart from standardized abbreviations such as *etj.*, i. e. engl. *etc.* or *znj.*, i. e. engl. *mrs.*) or *flm* (202 matches in 117 different texts vs. *falemnderit* 375 matches in 158 different texts). There are two types of abbreviations: standardized (lexicalized, which can also be found in dictionaries) and those not yet included in (definition) dictionaries. The abbreviation *nqs.*, i. e. alban. *nëqoftëse*, engl. *in case that*, is a standardized abbreviation, while *flm.* is not. For some standardized abbreviations, different (parallel) versions have been created. For example, for *nqs.*, the version *nqse* is also used.

Another form of abbreviation is the shortening of particle forms, especially those that are special characters, such as

*të* in *t*, *që* in *q*, or *një* in *nj*, e. g. *t dy* instead of *të dy*, engl. *both*. In dialectal forms, the word *nji*, i. e. in standard *një*, engl. *one*, is shortened to *i*. Since *i* has multiple functions, this abbreviation leads to many or multiple ambiguities. The abbreviation *sh*, i. e. *shumë*, engl. *much*, *a lot* is used very frequently (128 matches in 89 different texts).

### 4.2. Contractions

When dealing with empirical material from CMC, contractions are unavoidable. They are present and make reading, understanding, and processing the text more difficult. First of all, a distinction is made between contractions that are standardized, i. e. that correspond to the standard language variant, e. g. *s'e ke ...*, engl. *you do not have (it) ...* where *s'* is the negation *not*, and those that correspond to dialectal language variants, e. g. *se ke ...*, i. e. variant spelling and without apostrophe, where *s'* is negation and *e* is the dative object marker / pronominal clitic. Without the apostrophe, *se* has a different semantics, i. e., a conjunction, which can often lead to syntactic ambiguities. Negations *nuk* and *s'* and interrogative particles *ç'* very often lead to contractions with the corresponding (negated or questioned) words.

Some dialectal contractions are so strong that they cannot be easily segmented and analyzed, e. g. *je te menu* i. e. *je duke e menduar*, engl. *you're thinking about it right now*. The *t* stands for geg *tue* (= *duke* in standard version, engl. [I am] *doing it right now*), while *e* for accusative object / clitic. Similarly, shortening forms such as *pi*, which is an abbreviation of the preposition *prej*, engl. *from*, led to ambiguities. The word *pi* means (*I*) *drink* in standard language, so it is a verb.

### 4.3. Creative spelling

I were able to observe more than 100 forms of creative spelling in the corpus. In particular, these are numbers whose pronunciation or sequence of letters can be used as part/segment for shortening the spelling of a word. Many words are spelled creatively in a dialect. Especially, when some people are unsure how exactly certain dialect words are spelled since there are no spelling rules/standards. So, a number creates leeway.

Some of the most commonly used and well-known creative spellings that I observed in the corpus are listed as follows: *1* for *një*, engl. *one*, *Iher*, *Iherë*, engl. *once*, *2shim* for *dysim*, engl. *doubt*, *i 2ti* for *i dyti*, engl. *the second one*, *3g* for *treg*, engl. *market*, *3go/m* for *trego/më*, engl. *tell (me)*, *3t/je* for *tret/je ...*, engl. *dissolved* or *digested* etc.

## 5. Handling of non-standard language data

Dealing with data, first on the formal side, lexis, morphology, morpho-syntax and syntax can turn out difficult. The fact that everyone writes, even those who do not know the language and language rules, such as spelling, makes decoding computer-mediated data challenging. According to the data, however, it is not uncommon for a highly detailed specialist knowledge to be required in order to decode the data, i. e. to normalize it in terms of language variants, e. g. the language variant of a migrant child (in the 3rd generation) who lives in the German-speaking part of Switzerland and communicates in an Albanian language dialectic variant combined with a German dialectic variant (i.e. idiolect with code mixing, code switching). In

addition, there are also other levels that are coded in computer-mediated data requiring the coder social, cultural, and professional knowledge of the register of a particular language and its' variational characteristics. Coders can also decode communicator-related factors such as the communicator's intention (Pragmatics) and apply a specific codebook for this purpose. This level can also encode many characteristics, i. e. from such positive complimentary communication to hate speech.

## 6. Automatic processing of the data

Due to the large amounts of data, they can first be processed automatically. Some preliminary work has already been

done or tools have been developed for this purpose. We are building on the work of Kabashi & Proisl (2018) and Proisl & Uhrig (2016) among others to tokenize and annotate the data. We also follow the handling and working methods with CMC Data, presented in Proisl et al. (2019) and Proisl et al. (2020).

A gold standard is very important for the successful processing of data with NLP methods. We have done preliminary work for data collection from the standard language, which we are gradually adapting or building up for the CMC data. The following table demonstrates the automatic processing of the data:

→ correct encoding of the source texts/data: UTF8, XML, CWB/CQP ...

→ normalization of the data

code mixing: Well e kam akoma dhe tendin q teknikisht ishte / esht i imi so ☺ ☺ Im thriving kam m shm content per t par  
 engl. engl. engl. engl. engl.  
 spelling: Well e kam akoma dhe tendin q teknikisht ishte / esht i imi so ☺ ☺ Im thriving kam m shm content per t par  
 normalizing: tëndin që është I'm më shumë për të parë

→ Lemmatization

spelling: Well e kam akoma dhe tendin q teknikisht ishte / esht i imi so ☺ ☺ I'm thriving kam m shm content per t par  
 normalized: [Mirë] e kam akoma dhe tëndin që teknikisht ishte / është i imi [kështu] ☺ ☺ [jam duke lulëzuar] kam më shumë [përmbytje] për të parë  
 normed-linked: [Well=Mirë] ... [so=kështu] ... ["I'm thriving"="jam duke lulëzuar"] ... [content=përmbytje] ...  
 lemmatization: [Mirë] e kam akoma dhe tënd q teknikisht jam / jam i imi [kështu] ☺ ☺ [jam duke lulëzoi] kam më shumë [përmbytje] për të shoh

→ the POS-tagging

normalized text: [Mirë] e kam akoma dhe tëndin që teknikisht ishte / është i imi [kështu]. ☺ ☺ [Jam duke lulëzuar]. Kam më shumë [përmbytje] për të parë.  
 tagged text: Pt , Art V Adv ConIS PPoss ConIS Adv V / V Art PPossA Adv . EM EM VAux PtGer VPart . V PtComp Adv N PtInf PtSubi Vpart .

→ Universal Dependencies (UD)

parsing: gold standard, in progress

Table 2: The processing steps for texts in our CMC corpora of Albanian.

## 7. Conclusion

In contrast to standard data, the computer-mediated data discussed, especially those that deviate from the standard, represent a major scientific challenge while pursuing an automatic identification of linguistic characteristics of sub-standards. One of the main problems is the normalization of the data, including solving for contractions, etc., to create a gold standard for training a model that can then be used to tag the data (in large quantities).

I believe that this contribution will help to get an insight into the main phenomena of computer-mediated data from social media in the Albanian language. In addition, I hope that the paper contributes to the development of automatized processes aiming at the identification of the recognition of variants in large computer-mediated data in Albanian (as far as it is possible so far, due to many limitations, e. g. lack of data collections, lack of gold standards and processing tools).

## 8. References

Kabashi, B. and Proisl, T. (2018). Albanian part-of-speech tagging: Gold standard and evaluation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European

Language Resources Association (ELRA).

Kabashi, B. (2024). Lexical variation of the Albanian language used in computer-mediated communication and the challenge for processing. In *The 11th Conference on computer-mediated communication and social media corpora*, pp. 106–107, Nice, France. Université Côte d'Azur.

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

Proisl, T., Dykes, N., Heinrich, P., Kabashi, B., and Evert, S. (2019). *Lematisierungsrichtlinien*. Guideline document.

Proisl, T., Dykes, N., Heinrich, P., Kabashi, B., Blombach, A., and Evert, S. (2020). EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatization and Semantic Tagging to a German Web and CMC Corpus. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, Marseille, France. European Language Resources Association.



## **V. Poster Abstracts**





# Augmenting the CoWoYTP1Att Corpus with Emotion and Hate Speech Annotations: A Study on the Relationship with Appraisal Theory

Valentina Tretti-Beckles, Adrian Vergara-Heidke

Potsdam University

E-mail: vale3t@gmail.com, adrian.vergara@ucr.ac.cr

## Abstract

This study presents a methodology for enriching the Corpus on Women in YouTube on Performance with Attitude Annotations (CoWoYTP1Att) (Tretti-Beckles, Vergara-Heidke and Molina-Valverde, 2025). Originally developed based on Appraisal Theory (Martin and White, 2005), this corpus has been expanded with two additional layers of annotation: emotion and hate speech. CoWoYTP1Att consists of 1,521 Spanish-language YouTube comments related to the performance *Un violador en tu camino* (A Rapist in Your Path) by the feminist collective LasTesis. The decision to analyze these comments stems from the video's engagement with a pressing social issue—structural violence against women—which elicits diverse opinions. Moreover, the comments originate from various Spanish-speaking countries, offering potential variation in both evaluative stances and discursive constructions surrounding gender-based violence.

The original annotation of the corpus includes the attitude subdomains of Appraisal Theory (affect, judgement, appreciation), as well as polarity, target, fragment, and the explicitness and implicitness of the attitude expressed.

To augment this resource, we fine-tune a transformer-based language model using the Spanish subset of the EmoEvent dataset (Plaza-del-Arco, Strapparava, Ureña-López, and Martín-Valdivia, 2020), which contains manually labeled instances of emotion (anger, sadness, joy, disgust, fear, surprise, offensive, other) and hate speech (OFF/NO). The resulting model is then used to automatically annotate the CoWoYTP1Att corpus with predicted emotion and hate speech labels.

This approach seeks to facilitate the analysis of the relationships between evaluative language and affective or hateful content. The study is guided by four hypotheses: (1) affective segments are likely to co-occur with emotion labels; (2) implicit affect is underrepresented in predicted labels; (3) explicit evaluations are more frequently associated with emotion; and (4) negative judgments correlate with hate speech.

The expected outcome is a multilayered corpus that enables integrated analysis of appraisal, emotion, and hate speech in Spanish-language online discourse. Ultimately, the study aims to shed light on how patterns of emotional and evaluative language contribute to toxic or polarized interactions on social media platforms.

**Keywords:** Appraisal Theory, emotion, hate speech, gender, data augmentation

## References

- Tretti-Beckles, V., Vergara-Heidke, A. & Molina-Valverde, N. (2025). CoWoYTP1Att: A Social Media Comment Dataset on Gender Discourse with Appraisal Theory Annotations. In Proceedings of the Fifth Conference on Language, Data and Knowledge. [To Appear]
- J.R. Martin and P. White. 2005. The Language of Evaluation: Appraisal in English. Palgrave Macmillan.
- Plaza-del-Arco, F.; Strapparava, C.; Ureña-López, L.A.; Martín-Valdivia, M.T. (2020). EmoEvent: A Multilingual Emotion Corpus based on different Events. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, pp. 1492–1498.

# Methodology for Developing a Fact-Checked News Dataset in Norwegian Bokmål for Fake News Detection (The Fakespeak-NOR Corpus)

Aleena Thomas<sup>1</sup>, Silje Susanne Alvestad<sup>2</sup>

SINTEF AS, Oslo, Norway<sup>1</sup>, University of Oslo<sup>2</sup>

E-mail: aleena.thomas@sintef.no<sup>1</sup>, s.s.alvestad@ilos.uio.no<sup>2</sup>

## Abstract

This work presents the methodology for constructing a novel dataset of fact-checked news articles in Norwegian bokmål, a language with relatively limited publicly available resources for natural language processing. To the best of our knowledge, this is the first dataset of its kind that combines the text of the news article and its veracity label. The source of the data is Faktisk.no, the only Norwegian fact-checking organization. Each of their fact-checks is published with detailed assessment of a claim, including a link to the original article in which the claim first appeared along with a verdict (5 categories from completely true, partially true, not sure, partially false and completely false) and a justification based on factual evidence. The dataset creation process involves several filtering steps. Firstly, all the links to the articles with the original claim were validated. Articles that had been deleted, often due to the claim being flagged as false, were excluded. Non textual content, such as video and audio, were identified using keywords in the url of the link and removed. Articles that were behind hard paywalls were also removed. From the initial pool of 423 articles, approximately 200 valid instances were retained. Each article was manually reviewed to ensure that the claim being assessed was still present in the current version of the source article. A key challenge in compiling such datasets is that false claims are frequently deleted or edited after being fact-checked, resulting in many articles being unusable. The final dataset includes, for each instance, the claim under evaluation, the corresponding article text, its title, and its veracity label. This collection is intended to support future research on the language of fake news as well as mis- and disinformation detection in low-resource languages.

**Keywords:** Fake news detection, Norwegian bokmål, dataset creation, fact-checking, misinformation

# Building and querying Wikipedia discussion corpora using KorAP

Eliza Margaretha, Harald Lungen, Nils Diewald, Marc Kupietz, Rameela Yaddehige

Leibniz-Institut für Deutsche Sprache

Mannheim

(margaretha|luengen|diewald|kupietz|yaddehige)@ids-mannheim.de

## Abstract

We introduce the new German Wikipedia talk page corpus with 1.14 billion tokens and multiple linguistic annotation layers, available via the corpus analysis platform KorAP.

**Keywords:** Wikipedia, talk pages, wikitext, CMC, corpus construction, KorAP

The 349 language versions (as of August 2024) of the online encyclopedia Wikipedia are consulted by hundreds of thousands of users daily and serve as the foundation of major LLMs. They are a huge collaborative effort with thousands of authors who contribute on a voluntary basis. Besides the encyclopaedic articles, Wikipedia contains talk pages (a.k.a. discussions), i.e. a namespace where the authors discuss and negotiate the composition of articles (article talk) or discuss more freely (user talk). Unlike the articles, talk pages are organised in a dialogue structure with postings and threads, hence they form a very large and interesting linguistic archive of CMC (Lungen and Kupietz, 2017; Ho-Dac, 2024).

The poster presents the latest article and user discussion corpora of the German Wikipedia we built in 2024, comprising 1.14 billion tokens in 1.5 million documents. The corpus is accessible through KorAP (Bański et al., 2012), a web-based corpus analysis platform. In addition to a graphical user interface, KorAP provides API web-services that allow users to access corpora using client applications such as RKorAP-Client (Kupietz et al., 2020) to extract and visualize the results in these applications.

## Enhanced Wikipedia corpus builder

KorAP takes I5, the TEI customisation for the German Reference Corpus DEREKO (Lungen and Sperberg-McQueen, 2012) as import format. The Wikipedia sources come as database dumps from <https://dumps.wikimedia.org/> and contain all pages in the wikitext format. Figure 1 illustrates our conversion pipeline comprising five modules: pre-processing, parsing, XML rendering, transformation, and post-processing. In pre-processing, HTML tags are escaped to retain structure, missing tags are corrected using TagSoup<sup>1</sup>, and posting segmentation is applied. Subsequently, the wikitext is parsed using Sweble<sup>2</sup> into an abstract syntax tree (AST), which is eventually rendered as XML, producing a WikiXML corpus. An

I5 Wikipedia corpus is finally produced through XSL transformation followed by post-processing to handle categories and cross-language links.

## Querying emojis and emoticons in KorAP

Talk pages exhibit orthographic and lexical features typical of CMC including emoticons and emojis. Plain text emoticons (e.g. :-)) are searchable in KorAP. KorAP also supports searching emojis by using Unicode (see Figure 2), however, only a small number of emojis are encoded in Unicode (e.g. 😊) in wikitext. Most emojis are encoded as templates (e.g. {{S}}), which requires special processing in tokenization to enable their searchability, and a normalization to Unicode to improve visualization.

The KorAP instance that contains the latest wiki corpora is located at <https://korap.ids-mannheim.de/instance/wiki>.<sup>3</sup>

## Extracting and visualising results using RKorAPClient

The poster will feature analyses that compare linguistic properties of the article talk pages, the user talk pages, and a press-subcorpus from the German Reference Corpus DEREKO (Kupietz et al., 2010).

## References

- Bański, P., Fischer, P. M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O., and Witt, A. (2012). The new IDS corpus analysis platform: Challenges and prospects. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2905–2911. European Language Resources Association (ELRA).
- Dohrn, H. and Riehle, D. (2011). Design and implementation of the Sweble Wikitext parser: unlocking the structured data of Wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11*, pages 72–82. ACM.
- Ho-Dac, L.-M. (2024). Building a comparable corpus of online discussions on Wikipedia. In *Investigating*

<sup>1</sup><http://vrici.lojban.org/~cowan/tagsoup/>

<sup>2</sup>Sweble (Dohrn and Riehle, 2011) is a parser for wikitext documents. It has been updated recently for compatibility with Java 17, but is no longer under active development. <https://github.com/sweble>

<sup>3</sup>In addition, the corpora are available for download at <https://www.ids-mannheim.de/en/digspra/pb-s1/projects/corpus-development/verfuegbarkeit/>.

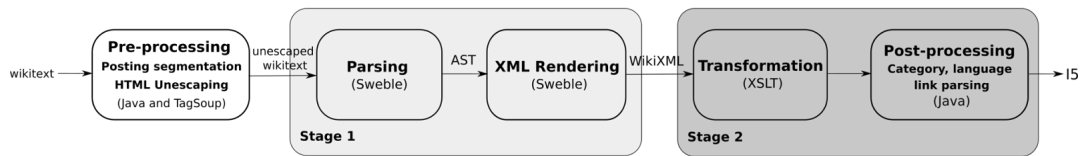


Figure 1: Wikitext to I5 conversion pipeline



Figure 2: Creating a new query to search for an emoji by using the *Query-By-Match* assistant in the annotation view in KorAP

*Wikipedia: Linguistic corpus building, exploration and analysis*, pages 12–44. Benjamins.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1848–1854. European Language Resources Association (ELRA).

Kupietz, M., Diewald, N., and Margaretha, E. (2020). RKorAPClient: An R package for accessing the German Reference Corpus DeReKo via KorAP. In

*Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7015–7021. European Language Resources Association.

Lüngen, H. and Kupietz, M. (2017). CMC corpora in DeReKo. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and natural Language Processing (CMLC-5+BigNLP)*.

Lüngen, H. and Sperberg-McQueen, M. (2012). A TEI P5 document grammar for the IDS text model. *Journal of the Text Encoding Initiative (jTEI)*, pages 1–18.

# “Prompt as Culture”: A Cross-linguistic Analysis of Prompt Engineering Discourse on Chinese and English Social Media

Xiaomin Zhang

University of Modena and Reggio Emilia

E-mail: xiaomin.zhang@unimore.it

## Abstract

With the increasing integration of generative AI systems such as ChatGPT and DeepSeek into everyday communication, “prompt engineering” has become a salient topic across global social media platforms. While much attention has been paid to the technical function of prompts, relatively little is known about how prompts are discussed, framed, and shared as cultural and linguistic resources across languages.

This study examines how prompt-writing practices are discursively constructed in Chinese and English online communities, analyzing the linguistic and cultural framing of prompts as communicative resources. A bilingual corpus of approximately 500 prompt-related posts and comments was compiled from Chinese platforms (e.g., Xiaohongshu, Zhihu) and English-speaking platforms (e.g., Reddit, Twitter), focusing on user-generated posts and comments discussing prompt strategies, effectiveness, and creative uses. Each platform contributed around 10 – 15 prompt-themed threads or posts and 200 – 250 top-level or threaded comments. Sampling was based on relevant keywords (e.g., “prompt engineering”, “提示词”, “AI 提效”) and stratified to include different genres (e.g., tutorials, reviews, exploratory discussions) and user orientations. All data were manually collected from publicly accessible sources following ethical guidelines. The analysis applies a genre-based discourse approach (Swales, 1990) and integrates pragmatic and rhetorical annotation focusing on moves, and metadiscursive framing (Hyland, 2005; Martin & White, 2005).

Preliminary findings indicate clear contrasts in the discursive orientation and rhetorical practices surrounding prompt engineering. Chinese-language discourse frequently presents prompts in the form of reusable templates, strategic guides, and “foolproof formulas,” reflecting a utilitarian and instructional approach. Posts often emphasize practical success, optimization, and standardization, aligning with a “knowledge-sharing” and tutorial-driven style. In contrast, English-language discourse tends to frame prompt writing as an experimental and iterative process, marked by terms such as “hack,” “jailbreak,” and “reverse-engineer.” This reflects a more exploratory, self-reflexive, and technologically playful orientation.

By comparing how prompt engineering is discursively constructed in two linguistic and cultural contexts, this study sheds light on the culturally embedded nature of emerging AI literacies. It contributes to the growing body of research in cross-linguistic computer-mediated communication (CMC), discourse pragmatics, and the sociolinguistics of digital expertise (Androutsopoulos, 2013; Lee & Barton, 2013). The findings also call attention to the role of language and culture in shaping how users conceptualize agency, creativity, and control in AI-mediated environments.

**Keywords:** Prompt Engineering Discourse, Cross-linguistic CMC Analysis, Human-AI Interaction in CMC

## References

- Androutsopoulos, J. (2013). Computer - mediated Communication and Linguistic Landscapes. *Research methods in sociolinguistics: A practical guide*, 74-90.
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.
- Lee, C., & Barton, D. (2013). *Language online: Investigating digital texts and practices*. Routledge.
- Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.
- Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

# Towards a Biased Language Taxonomy: first steps

Marini Costanza, Elisabetta Jezek

University of Pavia

E-mail: costanza.marini@unipv.it, elisabetta.jezek@unipv.it

## Abstract

The Biased Language Taxonomy (BLT) is a linguistically sound and theoretically informed taxonomy of biased language phenomena which is being developed at the University of Pavia to provide Higher Education actors – i.e., students, teachers and researchers – with an analytical tool to critically assess and counteract the pervasiveness of biases in CMC, which can strengthen inequality even in traditionally progressive fields such as HE (Beukeboom & Burgers 2019). Given that, from a social psychology perspective, stereotypes, prejudices and discrimination can be considered forms of cognitive, emotional and behavioural biases (Fiske 2024), with the term “biased language” we refer to all those instances in which language is used to reflect shared cognitive and affective associations that may induce discriminatory behaviour. Moreover, since the target groups towards which biased language is aimed at are as many as the ways in which we naturally categorize individuals (e.g., age, gender, nationality), the BLT is primarily focused on the linguistic dimension of biased language. Informed by interdisciplinary literature (Spinde et al. 2024) and empirical observations on an ad-hoc dataset of Telegram posts from a group of Trump supporting students (Students for Trump), this initial version of the BLT includes different categories of biased language phenomena, such as: epistemological bias, syntactic bias, and semantic bias. Epistemological bias is linked to linguistic cues that impact the believability of a statement (often via presupposition), such as factive and assertive verbs, hedges and boosters (Recasens et al. 2013). On the other hand, passive constructions used without a by-phrase and anticausative verbs (Greene & Resnik 2009) are clear-cut cases of syntactic bias. Finally, all framing effects caused by lexical choices involving subjective, one-sided or polarized words (Entman 2007) fall under the umbrella of semantic bias. The BLT was initially pilot tested on CMC data in Pavia in May 2025 on a group of international students from the EC2U (European Campus of City Universities) Alliance.

**Keywords:** biased language, taxonomy, media bias, stereotypes, Higher Education

## References

- Beukeboom, C. J., Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7, pp. 1–37. <https://doi.org/10.12840/issn.2255-4165.017>
- Entman, R.M. (2007). Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1), pp. 163–173.
- Fiske, S. T. (2024). Prejudice, discrimination, and stereotyping. In R. Biswas-Diener, & E. Diener (Eds.), *Noba textbook series: Psychology*. Champaign, IL: DEF publishers. <http://noba.to/jfkx7nrd>
- Greene, S., Resnik, P. (2009). More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, CO: Association for Computational Linguistics, pp. 503–511. <https://aclanthology.org/N09-1057>
- Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D. (2013). Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Sofia, Bulgaria: Association for Computational Linguistics, pp. 1650–1659. <https://aclanthology.org/P13-1162>
- Spinde, T., Hinterreiter, S., Haak, F., Ruas, T., Giese, H., Meuschke, N., Gipp, B. (2024). The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias, *arXiv*. <https://arxiv.org/abs/2312.16148>

# Diversifying Meaning in a Viral Age: The Case of 'Demure' on Social Media

Haruka Nishiyama

Keio University

E-mail: haruka.nishiyama.3@gmail.com

## Abstract

This presentation explores how the meaning of a trending word extends and diversifies through widespread social media use. Given that social media involves context collapse among diverse users (Androutsopoulos, 2014) and facilitates the rapid spread of novel expressions (Bailey & Durham, 2020), it provides a compelling environment for examining how language meaning becomes unstable, reframed, or experimentally extended.

In discursive semantics, meaning construction is examined through situational meaning and context (Lecolle, Veniard & Guérin, 2018; Reboul-Touré, 2021). This study adopts that lens in a micro-diachronic and quantitative analysis of the word *demure*, tracing its semantic trajectory as it went viral on social media in 2024—an event that led to its selection as Dictionary.com’s Word of the Year. Traditionally associated with modesty in (especially female) behavior or appearance, *demure* was adapted in user posts to describe a wide range of behaviors and attitudes, often referencing the phrase “very demure, very mindful.” Based on approximately 4,000 X (Twitter) posts collected over several days, the study examines how the word was recontextualized and reimagined in everyday discourse.

Findings show that *demure* was initially used with photogenic images (e.g., people, celebrities, pets) or in simple declarations like “I’m very demure.” Over time, it was linked not only to polite or restrained actions (e.g., dieting, not smoking) but also to inanimate events such as mild earthquakes or small food portions. While some uses stemmed from the original collocation with *mindful*, others reflected playful, ironic, or abstract reinterpretations. These diverse uses suggest that users were not aiming for a unified meaning, but instead engaging in unpredictable, creative repurposing.

The presentation concludes by discussing how these shifts unfolded over time, revealing mechanisms of meaning variability, reuse, and mutation in viral online discourse.

**Keywords:** Meaning variation, Viral language, Social media linguistics

## References

- Androutsopoulos, Jannis. (2014). Languageing when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, 4-5, 62–73. <https://doi.org/10.1016/j.dcm.2014.08.006>
- Bailey, Laura R., & Mercedes Durham. (2020). A cheeky investigation: Tracking the semantic change of cheeky from monkeys to wines. *English Today*, 37(4), 214–223. <https://doi.org/10.1017/s0266078420000073>
- Lecolle, Michelle, Marie Veniard & Olivia Guérin. (2018). Pour une sémantique discursive: propositions et illustrations. *Langages*, 210(2), 35–54. <https://doi.org/10.3917/lang.210.0035>
- Reboul-Touré, S. (2021). The crisis in discourse: As an event, a discursive semantics, and a culture. *Z Literaturwissenschaft Linguist*, 51(3), 399–420. <https://doi.org/10.1007/s41244-021-00211-5>

# Discursive Polarisation and the (Non-)Binary Spectrum: Social Media Debate on Gender Diversity

Andressa Costa

Karlsruhe Institute of Technology

E-mail: andressa.costa@kit.edu

## Abstract

The role of social media platforms in amplifying debates around gender binary and non-binarity has become increasingly evident in recent years. Indeed, these platforms have been shown to have an effect on ideological and affective polarisation. The gender binary characterises gender as inherently dichotomous and perpetuates social hierarchies, including patriarchy and cisnormativity. Non-binary identities challenge this binary, exposing its constructed nature and reframing gender as a spectrum rather than a dichotomy. Through their technological structure and interaction dynamics, social media intensify psychological mechanisms of opinion formation, consolidating attitudes and increasing polarisation (Könneker, 2020). This study aims to investigate discursive polarisation in the gender debate in social media by addressing two fundamental questions: How do ideological and affective polarisation manifest in social media discussions on gender diversity? Which rhetorical strategies are used to create or reinforce polarisation? Combining discursive polarisation frameworks (Brüggemann & Meyer, 2023) with computational text analysis, this study analyses the KoKoKom corpus of German social media comments (YouTube, Reddit, Instagram, X, Facebook) on the Gender debate. This addresses the gaps in operationalising polarisation through both linguistic patterns and affective discourse. Biterm Topic Modelling (BTM) (Yan et al., 2013) was used to identify latent topics, with concordance analysis and manual coding employed to classify the topics as ideological framings and detect polarising rhetorical devices (Fortuna, 2019). Three central frames were identified through topic modelling and qualitative analysis: the binary gender model, biological versus social concepts of sex/gender, and epistemic frameworks and controversies in sex and gender discourse. Concordance analysis revealed rhetorical strategies, including derogatory language, references to moral or scientific authority, overgeneralisation and simplification through dichotomies. Subsequent analyses will concentrate on affective polarisation to examine how groups are communicatively constructed in the debate. The present study contributes methodologically to language-centred research on polarisation by employing computational and discourse-analytic methods.

**Keywords:** discursive polarisation, gender debate, social media corpus, topic modelling, rhetorical strategies

## References

- Brüggemann, M., & Meyer, H. (2023). When debates break apart: Discursive polarization as a multi-dimensional divergence emerging in and through communication. *Communication Theory*, 33(2–3), 132–142. <https://doi.org/10.1093/ct/qtad012>
- Fortuna, A. (2019). *Polarization: Rhetorical Strategies in the Tea Party Network*. De Gruyter.
- Könneker, C. (2020). Wissenschaftskommunikation und Social Media: Neue Akture, Polarisierung und Vertrauen. In *Wissenschaft und Gesellschaft: Ein vertrauensvoller Dialog. Positionen und Perspektiven der Wissenschaftskommunikation heute* (S. 25–46). Springer. <https://doi.org/10.1007/978-3-662-59466-7>
- Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A Biterm Topic Model for Short Texts. 1445–1456. <https://github.com/xiaohuiyan/xiaohuiyan.github.io/blob/master/paper/BTM-WWW13.pdf>



# Gender differences in Chinese sensory adjectives: A corpus-based study of food videos on Bilibili

Mingyu Liu

The Hong Kong Polytechnic University

E-mail: 19078695d@connect.polyu.hk

## Abstract

Sensory adjectives play a crucial role in articulating human experiences across five modalities: vision, hearing, touch, taste, and smell. These adjectives not only describe perceptual experiences but also exhibit patterns of linguistic synaesthesia, where meaning extends from one sensory domain to another (Zhao, Huang & Ahrens, 2019; Zhao, 2018). However, while previous studies have examined these patterns from a theoretical perspective using existing corpora (Zhao & Huang, 2016; Zhao, Huang & Long, 2018; Zhao, Long & Huang, 2020), limited attention has been given to domain-specific, real-world contexts. Moreover, although gender differences in sensory physiology (Halpern, 2012; Velle, 1987) and language use (Coates, 2004; Nemati & Bayer, 2007) are well-documented, few studies have focused on the difference between males and females in the use of sensory adjectives. This study addresses these gaps by investigating gender differences in the use of sensory adjectives and synaesthesia in Chinese food-related discourse. Specifically, it aims to answer the research questions: (1) Are there gender differences in the frequency of sensory adjective usage in food videos on Bilibili? (2) Are there gender differences in the frequency and directionality of synaesthesia in food videos on Bilibili?

To address these questions, this research adopts a corpus-based approach, compiling transcripts from food-related videos on the Bilibili platform, a leading Chinese video-sharing platform with a highly gender-balanced user base. The male and female corpora each contained approximately 50,000 characters, drawn from 51 male-authored and 50 female-authored videos respectively. To identify and classify sensory adjectives, a three-step process was adapted. First, a random 10% sample of the corpus was manually analyzed, extracting sensory adjectives. An independent coder re-evaluated this subset to ensure consistency in coding. Next, ambiguous cases were further examined through lexical and semantic analysis, including etymological tracing. The full corpus was then analyzed using the Sketch Engine, applying the validated sensory adjective list to retrieve and compare occurrences across male and female data.

Results show that females use sensory adjectives more frequently and with higher lexical diversity, particularly in the tactile domain. Females also employ synaesthesia more often than males. Regarding synaesthetic directionality, males are more likely to transfer to the olfactory-targeted domain, while females favor gustatory-based synaesthetic mappings. The findings suggest that gender differences exist in the use of sensory adjectives and synaesthesia. These differences may be attributed to variations in physical sensory sensitivity between genders (Brand & Millot, 2001; Halpern, 2012) and differences in linguistic abilities and preferences (Chouchane, 2016; Cryan et al., 2020).

**Keywords:** sensory adjectives, synaesthesia, gender differences, corpus-based, Chinese

## References

- Brand, G., & Millot, J.-L. (2001). Sex differences in human olfaction: Between evidence and enigma. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, 54(3), 259–270.
- Chouchane, A. M. (2016). Gender Language Differences Do men and women really speak differently. *Global English-Oriented Research Journal (GEORJ)*, 2(2), 182–200.
- Coates, J. (2004). *Women, men, and language: a sociolinguistic account of gender differences in language* (3rd ed.). Pearson Longman.
- Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2020). Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11.
- Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Psychology Press.
- Nemati, A., & Bayer, J. M. (2007). Gender Differences in the Use of Linguistic Forms in the Speech of Men and Women: A Comparative Study of Persian and English. *Language in India*, 7(9).
- Velle, W. (1987). Sex differences in sensory functions. *Perspectives in biology and medicine*, 30(4), 490–522.
- Zhao, Q. Q. (2018). Synaesthesia, metaphor, and cognition: a corpus-based study on synaesthetic adjectives in Mandarin Chinese. Hong Kong Polytechnic University.
- Zhao, Q. Q., & Huang, C. R. (2016). A Corpus-Based Study on Synaesthetic Adjectives in Modern Chinese. *Chinese Lexical Semantics (CLSW 2015)*, 9332, 535–542.
- Zhao, Q. Q., Huang, C. R., & Ahrens, K. (2019). Directionality of linguistic synesthesia in Mandarin: A corpus-based study. *Lingua: International Review of General Linguistics*, 232.
- Zhao, Q. Q., Huang, C. R., & Long, Y. (2018). Synaesthesia in Chinese: a corpus-based study on gustatory adjectives in Mandarin. *Mouton De Gruyter*.
- Zhao, Q. Q., Long, Y., & Huang, C. R. (2020). Linguistic synaesthesia of Mandarin sensory adjectives: corpus-based and experimental approaches. *Springer*.

# Emotional Expression in Text-Based Communication: An Analysis of Online Mentoring for Girls in STEM

Claudia Uebler<sup>1</sup>, Albert Ziegler<sup>2</sup>, Heidrun Stoeger<sup>1</sup>

<sup>1</sup>University of Regensburg, <sup>2</sup>University of Erlangen-Nuremberg  
E-mail: claudia.uebler@ur.de, albert.ziegler@fau.de, heidrun.stoeger@ur.de

## Abstract

Online mentoring is a promising measure to support girls in STEM (Stoeger et al., 2013; 2023). Because girls in such programs communicate asynchronously with female STEM professionals and peers by text message, these settings are well suited for analyzing participants' text-based communication. Previous research shows that STEM-focused communication typically enhances mentees' mentoring success (Stoeger et al., 2016; 2021), especially when the communication with the personal mentor is frequent and steady (Uebler et al., 2023). However, little is known about the role of the emotional tone of communication in informal online learning settings. Understanding the emotional expression in text-based communication is essential, especially when relationships and motivational processes are involved. In the context of online mentoring for girls, emotional tone might reflect mentees' engagement, interest, fear or difficulties related to STEM. Therefore, this study examines the emotional tone of mentees' communication in a one-year online mentoring program for girls (N = 859, M = 13.79 years, SD = 1.97 years) in STEM. Tracking mentees' text-based communication on the program's secure online platform results in a text corpus of 15,107 anonymized messages. A randomized and representative subset of the text corpus is annotated with emotional labels following Pennebaker et al.'s (2015) emotional categories. This annotated dataset serves as training data for a BERT model (Devlin et al., 2018), enabling automated classification of emotional tone in previously unlabeled messages. By linking the emotional tone of messages with the type of communication partners (i.e., individual mentors, peers) and indicators of mentoring success, we explore how emotional tone varies across different interaction contexts within the program. Preliminary results from zero-shot classification reveal distinct patterns in mentees' emotional and STEM-related communication with mentors compared to peers and demonstrate its predictive value for mentoring outcomes. Results of work in progress and implications for mentoring practice will be presented.

**Keywords:** emotional expression, text-based communication, online mentoring, girls in STEM, BERT model

## References

- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <http://arxiv.org/pdf/1810.04805>
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. University of Texas at Austin. <https://doi.org/10.15781/T29G6Z>
- Stoeger, H., Debatin, T., Heilemann, M., Schirner, S., & Ziegler, A. (2023). Online mentoring for girls in secondary education to increase participation rates of women in STEM: A long-term follow-up study on later university major and career choices. *Annals of the New York Academy of Sciences*, 1523(1), 62–73. <https://doi.org/10.1111/nyas.14989>
- Stoeger, H., Duan, X., Schirner, S., Greindl, T., & Ziegler, A. (2013). The effectiveness of a one-year online mentoring program for girls in STEM. *Computers & Education*, 69, 408–418. <https://doi.org/10.1016/j.compedu.2013.07.032>
- Stoeger, H., Heilemann, M., Debatin, T., Hopp, M., Schirner, S., & Ziegler, A. (2021). Nine years of online mentoring for secondary school girls in STEM: An empirical comparison of three mentoring formats. *Annals of the New York Academy of Sciences*, 1483(1), 153–173. <https://doi.org/10.1111/nyas.14476>
- Stoeger, H., Schirner, S., Laemmle, L., Obergriesser, S., Heilemann, M., & Ziegler, A. (2016). A contextual perspective on talented female participants and their development in extracurricular STEM programs. *Annals of the New York Academy of Sciences*, 1377(1), 53–66. <https://doi.org/10.1111/nyas.13116>
- Uebler, C., Emmerdinger, K. J., Ziegler, A., & Stoeger, H. (2023). Dropping out of an online mentoring program for girls in STEM: A longitudinal study on the dynamically changing risk for premature match closure. *Journal of Community Psychology*, 51(8), 3121–3151. <https://doi.org/10.1002/jcop.23039>

# Comparative Analysis of Comments on Feminism on Hupu and Xiaohongshu: A Text Mining Approach

Mingyu Liu

The University of Hong Kong  
E-mail: mingyuuu@connect.hku.hk

## Abstract

Feminist discourse on social media has been a topic of significant scholarly interest in recent years. However, existing research mainly focuses on analyzing feminist activities and movements on individual social media platforms within Western contexts (Jackson et al., 2020; Li, 2022; Molder et al., 2022; Suk et al., 2021). There is limited exploration of feminist comments themselves, particularly in terms of comparative analysis between different platforms. In the Chinese context, although platforms like Weibo have received academic attention (Bao, 2023; Huang, 2023), research remains scarce on how feminist discourse manifests on other major platforms.

This study seeks to address the gaps by examining comments about feminism on two distinct Chinese social media platforms: Hupu, a male-dominated platform with 95.4% male users, and Xiaohongshu, a female-oriented platform where 90.41% of users are female (Chi et al., 2022; Guo, 2022). It applies text-mining techniques, specifically BERTopic and sentiment analysis, to analyze how discussions around feminism vary in content and emotional tone. Accordingly, the study investigates the following research questions: (1) Is there any variation in topics between comments on feminism on Hupu and Xiaohongshu? (2) Is there any variation in sentiment between comments on feminism on Hupu and Xiaohongshu?

To address these questions, this study collected 28,909 comments (13,506 from Hupu and 15,403 from Xiaohongshu) based on eight selected feminism-related keywords. After text cleaning, the study applied BERTopic for topic modeling, using BERT-based sentence embeddings, UMAP for dimensionality reduction, HDBSCAN for clustering, and ClassTFIDF to enhance topic distinctiveness. Then, sentiment analysis was conducted using SnowNLP, which assigns sentiment scores ranging from 0 (negative) to 1 (positive) to evaluate the emotional tone of comments.

Results show that topic modeling generated 13 meaningful topics from Hupu and 14 from Xiaohongshu. Hupu's discourse showed a neutral to negative sentiment average ( $M=0.48$ ,  $SD=0.35$ ), with topics such as "Radical Feminism", "Misogyny and Misandry", and "Garbage" reflecting critical attitudes. In contrast, Xiaohongshu exhibited a positive sentiment average ( $M=0.67$ ,  $SD=0.31$ ), with topics like "Encouragement", "Kindness", and "Awareness" reflecting emotional support and alignment with feminist values. And some shared topics (e.g., Marriage, Work) showed platform-specific interpretations and sentiment trends. Taken together, these findings suggest clear differences in both topics and sentiment between the two platforms.

To further understand these patterns, the study tends to interpret these differences through the lens of ingroup and outgroup dynamics (Tajfel & Turner, 1979; Brewer, 1999). On Hupu, where users largely identify with conventional male perspectives, feminism is often perceived as an outgroup threat, resulting in critical, stereotyped, and polarized discourse (Sunstein, 2002; Judd & Park, 1988). In contrast, Xiaohongshu users, who tend to incorporate feminism into their ingroup identity, exhibit more supportive, emotionally resonant, and heterogeneous discussions (Park & Rothbart, 1982).

**Keywords:** feminism, social media, text-mining, Chinese

## References

- Bao, K. (2023). When feminists became 'extremists': A corpus-based study of representations of feminism on Weibo. *Discourse & Communication*, 17(5), 590-612.
- Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of Social Issues*, 55(3), 429-444.
- Chi, H., Liu, R., & Pan, J. (2022). Users' Behaviour under the Uneven Gender Ratio of Social Media Platforms: Taking Hupu and Xiaohongshu as Examples. *SHS Web of Conferences*, 148, 3003-3007.
- Guo, J. (2022). The postfeminist entrepreneurial self and the platformisation of labour: A case study of yesheng female lifestyle bloggers on Xiaohongshu. *Global Media and China*, 7(3), 303-318.
- Huang, Q. Q. (2023). Anti-Feminism: four strategies for the demonisation and depoliticisation of feminism on Chinese social media. *Feminist Media Studies*, 23(7), 3583-3598.
- Jackson, S. J., Bailey, M., & Welles, B. F. (2020). #HashtagActivism: Networks of race and gender justice. Mit Press.
- Judd, C. M., & Park, B. (1988). Out-Group Homogeneity: Judgments of Variability at the Individual and Group Levels. *Journal of Personality and Social Psychology*, 54(5), 778-788.
- Li, M. (2022). Visual social media and black activism: Exploring how using Instagram influences Black activism orientation and racial identity ideology among Black Americans. *Journalism & Mass Communication Quarterly*, 99(3), 718-741.
- Molder, A. L., Lakind, A., Clemmons, Z. E., & Chen, K. (2022). Framing the global youth climate movement: a qualitative content analysis of Greta Thunberg's moral, hopeful, and motivational framing on instagram. *The International Journal of Press/Politics*, 27(3), 668-695.
- Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, 42(6), 1051-1068.
- Suk, J., Abhishek, A., Zhang, Y., Ahn, S. Y., Correa, T., Garlough, C., & Shah, D. V. (2021). #MeToo, networked

- acknowledgment, and connective action: How “empowerment through empathy” launched a social movement. *Social Science Computer Review*, 39(2), 276-294.
- Sunstein, C. R. (2002). The Law of Group Polarization. *The Journal of Political Philosophy*, 10(2), 175–195.
- Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Monterey, CA: Brooks/Cole.

# Metapragmatic Perspectives on Autistic Digital Communication: A Corpus-Assisted Analysis of Self-Reported Practices

Nelya Koteyko

Queen Mary University of London

E-mail: n.koteyko@qmul.ac.uk

## Abstract

This study examines how autistic young people describe and evaluate their communication across diverse digital platforms, including Reddit, Snapchat, Facebook, and Instagram. Based on open-text responses from 144 UK-based participants aged 16–26, I examine how respondents navigate platform-specific interactional norms, focusing on the intersection of communicative expectations with sensory regulation, focused interests, and identity construction.

While thematic analysis (Braun and Clarke, 2006) may provide broad insights into communicative strategies, corpus analysis, employing Sketch Engine (Kilgariff et al, 2014), offers an empirical lens to substantiate and refine these findings. Keyword analysis identified statistically salient lexical items (e.g., "infodump", "hyperfocusing"), illuminating the centrality of interests and patterns characteristic of a highly focused communication style. Collocate analysis provided granular detail by revealing the typical linguistic environments of these keywords. For instance, analyzing collocates of "mask" (e.g., "have to", "constantly", "subconsciously") goes beyond identifying masking as a theme; it demonstrates the perceived involuntariness, enduring effort, and often unconscious nature of adaptive behaviours. Finally, grammatical pattern analysis offers insight into pragmatic intent and agency. The prevalence of 'I try to' constructions, for example, followed by verbs like 'limit', 'mask', 'avoid' or 'not to seem', illustrates participants' effortful self-regulation to avoid negative social perception and exclusion.

This systematic linguistic evidence enriches thematic categorization, revealing the linguistic patterning of self-regulation and pragmatic alignment in communicative choices. By foregrounding participants' own metapragmatic accounts of computer-mediated communication, this study demonstrates corpus-linguistic techniques' productive integration with thematic analysis to trace discourse-level regularities in self-report data. Future research will triangulate these findings with observational studies of actual online interactions to enhance generalizability and further explore diverse neurodivergent experiences.

**Keywords:** metapragmatics, autism, corpus-assisted analysis, social media

## References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Kilgariff, A., Baisa, V., Bušta, J. et al. (2014) The Sketch Engine: ten years on. *Lexicography ASIALEX* 1, 7–36.

# **(A)I Can Empathize with You: Analyses of Linguistically Performed Empathy and Human Identity of Large Language Models in Psychotherapeutic Settings**

**Florina Züllli**

University of Zurich

E-mail: florina.zuelli@uzh.ch

## **Abstract**

Against the backdrop of widespread shortages in mental health services and increasingly overstretched care infrastructures, artificial intelligence (AI) is gaining attention as a potential tool to support psychotherapeutic work. Empathic language use is, however, a fundamental prerequisite for such applications. This study examines the extent to which contemporary large language models can simulate empathic interaction in therapeutic dialogues. Based on 150 anonymized chat transcripts from a psycholinguistic study, the analysis compares human–human and human–machine interactions across three double-blind experimental conditions. Participants engaged in a written conversation about a psychologically distressing experience with either (i) a layperson, (ii) a trained psychology student, or (iii) a chatbot (ChatGPT-3.5) prompted to act as a psychotherapist. Importantly, participants did not interact with the chatbot directly via its standard interface. Instead, a human intermediary entered the chatbot’s responses into the online chat interface to ensure that all conditions remained textually uniform and to avoid interface-related bias. Empathy was operationalized from a psychological cognitive framework through four linguistic markers: Emotion Validation (“It’s perfectly natural to feel that way”), Engagement Questions (“How did that make you feel?”), Echoing (“You said you were sad about the outcome of this conversation”), and Encouragement (“That was very brave of you!”). These markers were coded independently through a double-rater evaluation process and retained only if they were identified by both raters before being quantified through a frequency analysis, resulting in an Empathy Score (E-Score). The results show that ChatGPT produced the highest overall frequency of empathy markers and performed on par with—or, for some markers, exceeded—the human conversation partners in the other two conditions. Data from the post-interaction questionnaire also showed that 50% of participants in the chatbot condition identified their conversation partner as human, indicating a substantial degree of linguistically performed humanness. These findings offer empirical insight into the language-based performance of empathy by AI and contribute to current debates on the ethical and practical implications of deploying conversational agents in mental health care. The study also addresses how language shapes perceptions of humanness and explores the performative construction of identity in human–machine interaction. By situating its results in the longer history of conversational agents—from ELIZA to contemporary systems such as ChatGPT and Gemini—it offers a differentiated perspective on the societal opportunities and challenges posed by AI systems capable of convincingly performing human identity through language.

**Keywords:** Human–machine interaction, Artificial Intelligence, Large language models, AI-based therapist, Artificial Empathy

# The Positive Pulse: The Hidden Language of Scientific Social Media

Cansu Akan, Sasha Genevieve Coelho

Chemnitz University of Technology

E-mail: cansu.akan@phil.tu-chemnitz.de, sasha.coelho@phil.tu-chemnitz.de

## Abstract

Social media platforms have become crucial in shaping public understanding of scientific issues. As Brossard & Scheufele (2013) note, “A world in which more than 340 million tweets are being posted everyday is not the future of science communication anymore. It is today’s reality.” Acknowledging this reality, this study examines the intersection of academic discourse and public communication, focusing on how scientific information is disseminated and received on the Science Media Centre (SMC), a trusted source disseminating scientific information to the public and social media.

Employing a corpus linguistics approach and sentiment analysis, the current study examines a dataset of 6,736 SMC posts, comprising 3.237.484 tokens, posted on the platform between 2020 and 2023 to uncover the emotional tone and linguistic characteristics of science communication. The study builds upon the Pollyanna principle, which suggests a tendency towards positive language in scientific publishing (Matlin, 2016) and utilizes sentiment analysis techniques to understand the emotional tone in the SMC posts. To investigate the shift in emotional tone, we created two sub-corpora pre-COVID and post-COVID, to assess whether there is an increase in positive language after the onset of the pandemic. VADER sentiment analyzer was used to quantify the emotional tone of the texts. Through corpus analysis, the study identifies key collocations and recurrent language patterns, which sheds light onto how Pollyanna principle manifests in academic engagement with public. Additionally, the study investigates the linguistic competencies that future scientists must develop for effective academic engagement on social media outlets.

Our findings provide crucial insights into how academic discourse is adapted for public engagement, examining the balance between scientific accuracy and accessibility, which linguistic competencies are required for effective science communication on social media platforms and what role social media plays in bridging the gap between academics and the general public.

**Keywords:** media corpora, discourse analysis, sentiment analysis, linguistic characteristics

## References

- Brossard, D., & Scheufele, D. A. (2013). Science, New Media, and the Public. *Science*, 339(6115), 40-41.
- Matlin, M.W. (2016). “Pollyanna principle” in R.F. Pohl (ed.): *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking and Memory*. Psychology Press, pp.315-333.
- Science Media Centre. Available at: <https://www.sciencemediacentre.org/> (Accessed: 20.05.2025).

# Science Communication in Science Slams

Johanna Vogel

Leibniz Institute for the German Language

E-mail: vogel@ids-mannheim.de

## Abstract

How do scientists speak when addressing a non-specialist audience? Science communication has become a widely discussed topic in the public sphere and a prominent area of academic research (e.g. Janich/Kalwa 2018). While several handbooks provide practical guidance for communicating with lay audiences (e.g., Falkenberg 2021; Wagner/McKee 2023), systematic analyses of spoken external science communication at the lexical-syntactic level remain scarce. In contrast, internal science communication, especially in written form, has been extensively explored, notably by Weinrich (1989) and Kretzenbacher (1992). For spoken academic discourse, the GeWiss corpus (Meißner/Slavcheva 2014) offers a valuable comparative framework. This presentation is part of a larger academic project investigating the linguistic structures of spoken external science communication, with a focus on social media formats such as Science Slams and TikTok videos.

The poster addresses methodological and technical challenges involved in constructing a corpus of spoken social media content and shares preliminary findings on the linguistic construction of orality in Science Slams. The reference corpus is being built from audio material, mainly from YouTube and TikTok, pre-segmented and automatically transcribed using aTrain. Transcriptions are manually corrected and annotated according to the cGAT conventions. Further processing is conducted using the corpus annotation software EXMARaLDA, and morphosyntactic tagging is carried out with TreeTagger based on STTS 2.0 guidelines, enabling a systematic, lexical-syntactic analysis.

The corpus will continue to grow over the coming years. The current dataset comprises approximately two hours of audio from the German Science Slam Championships (2021–2024). The initial analysis concentrates on markers of orality, with a focus on modal particles (e.g., also, halt), which, following Stein (2003: 439), can serve as indicators of spoken discourse. These findings are compared with recent analyses of internal scientific speech (cf. Schwendemann & Wallner 2023) to contribute toward defining spoken external science communication as a distinct linguistic register.

**Keywords:** science communication, corpus, work in progress

## References

- Falkenberg, Viola (2021): *Wissenschaftskommunikation: vom Hörsaal ins Rampenlicht: mit Übungen und Checklisten* (utb; Hochschullehre 5670). Tübingen: Narr Francke Attempto Verlag.
- Janich, Nina & Nina Kalwa (2018): *Wissenschaftskommunikation*. In Frank Liedtke & Astrid Tuchen (Hrsg.), *Handbuch Pragmatik*, 413–422. Stuttgart: J.B. Metzler. doi:10.1007/978-3-476-04624-6\_40.
- Kretzenbacher, Heinz Leonhard (1992): *Wissenschaftssprache* (Studienbibliographien Sprachwissenschaft 5). Heidelberg: Groos.
- Meißner, Cordula & Adriana Slavcheva (2014): *Das GeWiss-Korpus - ein Vergleichskorpus der gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen*. In Christian Fandrych, Cordula Meißner & Adriana Slavcheva (Hrsg.), *Gesprochene Wissenschaftssprache: korpusmethodische Fragen und empirische Analysen* (Wissenschaftskommunikation), 15–38. Heidelberg: Synchron.
- Schwendemann, Matthias & Franziska Wallner (2023): *Mündlichkeitsphänomene in der gesprochenen Wissenschaftssprache: Korpuslinguistische Befunde und didaktische Perspektiven*. *Informationen Deutsch als Fremdsprache*. De Gruyter. 50(5). 505–524. doi:10.1515/infodaf-2023-0083.
- Stein, Stephan (2003): *Textgliederung: Einheitenbildung im geschriebenen und gesprochenen Deutsch: Theorie und Empirie*. De Gruyter. <https://doi.org/10.1515/9783110906073>.
- Wagner, Laura & Cecile McKee (2023): *How to Talk Language Science with Everybody*. 1. edn. Cambridge University Press. doi:10.1017/9781108894227.
- Weinrich, Harald (1989): *Formen der Wissenschaftssprache*. In Horst Albach, Juliane Besters-Dilger, Martin Carrier, Hans-Peter Daniel, Aleksandr G. Granberg, Wolfgang Haber, Jaakko Honko, et al. (Hrsg.), *Jahrbuch der Akademie der Wissenschaften zu Berlin (Jahrbuch der Akademie der Wissenschaften zu Berlin / Yearbook of The Academie of Sciences and Technology in Berlin)*, 119–158. Reprint 2021. Berlin; Boston: De Gruyter. doi:10.1515/9783112417843-012.



# Appraising Chinese and Italian Operas on English-Language Social Media: A Corpus-Based Multimodal Discourse Analysis

Lei Liang

University of Modena and Reggio Emilia

E-mail: lei.liang@unimore.it

## Abstract

This doctoral research investigates how Chinese and Italian operas—two emblematic forms of intangible cultural heritage—are evaluated and interpreted by English-speaking audiences across different social media platforms. The study addresses the question: How do platform-specific affordances shape the appraisal of Chinese and Italian operas in English-language user-generated discourse?

As social media plays an increasingly influential role in mediating global cultural narratives, platforms such as YouTube, TikTok, Instagram, X (formerly Twitter), and Facebook offer diverse communicative environments that affect both how operas are presented and how they are received. This research adopts a corpus-based, cross-platform comparative approach, drawing on 250 original English-language posts (50 from each platform), collected using hashtags such as #ChineseOpera and #ItalianOpera. Each post, together with its top-level comments and accompanying multimodal elements (e.g., images, video stills, filters, emojis), constitutes a unit of analysis.

Analytically, the study applies Appraisal Theory (Martin & White, 2005) to examine how language is used to convey attitudes, manage engagement, and adjust intensity (Graduation). In parallel, multimodal discourse analysis (Kress & van Leeuwen, 2006) is employed to analyze how visual and dynamic resources interact with linguistic choices to construct evaluative meanings.

Preliminary findings suggest that Italian opera is frequently appraised in terms of emotional richness, visual grandeur, and historical prestige, often communicated through high-quality imagery and minimal explanatory text. By contrast, Chinese opera tends to be framed through cultural specificity and symbolic depth, often requiring more context or explanation. Audience responses reflect this gap: while Italian opera posts generally receive uniformly positive appraisals, reactions to Chinese opera vary more widely—ranging from admiration and curiosity to confusion or cultural stereotyping—underscoring the impact of cultural familiarity on evaluation.

Platform-specific patterns also emerge. TikTok and Instagram foreground affective intensity and aesthetic appeal; YouTube facilitates reflective, long-form appraisals; X promotes succinct, often polarized reactions; and Facebook shows lower engagement but deeper commentary. These distinctions highlight how platform affordances mediate not only the multimodal presentation of operas but also the evaluative stance of audiences.

This research contributes to studies of digital cultural heritage, social media discourse, and cross-cultural evaluation by revealing how traditional art forms are linguistically and multimodally appraised in transnational contexts. Future research may extend the analysis to Chinese-language content, enabling a more comprehensive understanding of how domestic and international audiences differ in their perceptions and evaluations of opera on social media.

**Keywords:** appraisal theory, Chinese and Italian opera, social media discourse, multimodal analysis, corpus-based analysis

# Decoding Business German: A Corpus-Based Lexical and Morphological Analysis of Contemporary Job Advertisements

Kristina Krcmarevic Bogdanovic, Kristina Ilic

University of Belgrade

E-mail: kristina.krcmarevic@gmail.com, kristina.ilic997@gmail.com

## Abstract

This study aims to investigate lexical and morphological features of job advertisements written in German and published over a seven-month period on poslovi.infostud, a leading online employment platform in Serbia. These ads, directed at German-speaking applicants, provide insight into how professional language is shaped within the digital sphere of business communication.

The theoretical foundation draws on the linguistic characteristics of German as a language for specific purposes (LSP) and the move-step model within the framework of genre analysis of job advertisements. The corpus itself was compiled out of 50 job postings from three professional branches (Customer Support, Human Resources, and Sales), all published on poslovi.infostud over a seven-month period and written in German. The samples were anonymized and automatically POS-tagged and lemmatized using TagAnt. The tagging and lemmatization were then manually checked and corrected, before importing the samples into AntConc for analysis. In AntConc we further used query tools including KWIC concordancing, word list generation, collocation mapping. This approach allows us to systematically examine how specific lexical and grammatical choices operate within the different segments of job advertisements.

Special emphasis is placed on identifying and interpreting industry-based terminology, frequently occurring lexical items, prevalent derivational structures, and complex compounds that are characteristic of institutional and professional German in digital communication. The analysis also considers the frequency and distribution of anglicisms, evaluating their role in signaling international orientation, corporate identity, and alignment with contemporary professional values. In addition to the linguistic features of the texts themselves, the study also briefly considers platform-specific characteristics of poslovi.infostud and, in this context, what distinguishes online postings from traditional job advertisements.

Beyond its descriptive and analytical aims, the study also demonstrates the applicability of accessible corpus methods to authentic CMC materials, offering theoretical insights as well as pedagogical implications for language for specific purposes (LSP), business German instruction, and intercultural communicative competence development.

**Keywords:** Business German, Corpus linguistics, online job advertisements, CMC, lexical analysis

## References

- Barjaktarović, S., Cicvarić-Kostić, S., & Kostić-Stanković, M. (2021). *Komunikacija brenda poslodavca u oglasima za posao* [Employer brand communication in job advertisements]. *Marketing*, 52(4), 225–234. <https://doi.org/10.5937/mkng2104225B>
- Bhatia, V. K. (2014). *Analysing genre: Language use in professional settings*. Routledge.
- Busse, U. (2019). Typen von Anglizismen: Von der heilago geist bis Extremsparing – aufgezeigt anhand ausgewählter lexikographischer Kategorisierungen. In G. Stickel (Hrsg.), *Neues und Fremdes im deutschen Wortschatz: Aktueller lexikalischer Wandel* (S. 131–155). De Gruyter. <https://doi.org/10.1515/9783110622669-008>
- Dobrić, N. (2008). Influence of English on names of professions. In L. Mišić & V. Lopičić (Eds.), *Language, literature and globalization* (pp. 305–316). Faculty of Philosophy, University of Niš.
- Eisenberg, P. (2018). *Das Fremdwort im Deutschen*. De Gruyter. <https://doi.org/10.1515/9783110474619>
- Kochetova, L. A., Ilyinova, E. Yu., Sorokoletova, N. Yu., & Volkova, O. S. (2017). Corpus-assisted comparative study of British job advertisements: Sociocultural perspective. In *Proceedings of the 7th International Scientific and Practical Conference on Current Issues of Linguistics and Didactics: The Interdisciplinary Approach in Humanities (CILDIAH 2017)* (pp. 139–145). Atlantis Press. <https://doi.org/10.2991/cildiah-17.2017.24>
- Krome, S. (2018). Skypen, faken, toppen und liken: Anglizismen im Deutschen als Indikatoren gesellschaftlichen und orthografischen Wandels. *Muttersprache*, 128(2), 105–122.
- Löffler, H. (2010). *Germanistische Soziolinguistik* (4., neu bearb. Aufl.). Erich Schmidt Verlag.
- Łacka-Badura, J. (2015). *Recruitment advertising as an instrument of employer branding: A linguistic perspective*. Cambridge Scholars Publishing.
- Roelcke, T. (2020). *Fachsprachen* (4., neu bearb. Aufl.; Grundlagen der Germanistik, Band 37). Erich Schmidt Verlag.





## **VI. Training Session with Stephanie Evert**



# Reading concordances with algorithms

Nathan Dykes, Stephanie Evert, Michaela Mahlberg, Alexander Piperski

Affiliation Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

E-mail: stephanie.evert@fau.de.

## Abstract

Concordance analysis has long been central to corpus linguistics and other text-based disciplines, including digital humanities, computational social sciences, and computer-assisted language learning. It gives researchers a systematic lens for observing and interpreting patterns of language use, integrating both quantitative and qualitative perspectives. By focusing on a single search word or phrase in a context-limited display—commonly known as a KWIC (Key Word In Context)—scholars can investigate various aspects of its usage and meaning.

In spite of its wide applications, concordance reading has seen little innovation to date. Popular functions of concordance tools are still the traditional approaches, such as sorting lines alphabetically by the left or right context of the node or filtering for specific words. Another challenge for concordance reading is the documentation of the research process and methods applied, in order to ensure reproducibility. The tutorial addresses these challenges by introducing both a taxonomy of concordance-reading strategies and a set of computational algorithms that build on these strategies to organize large amounts of textual data efficiently and transparently. Through hands-on exercises using the new Python library FlexiConc (<https://pypi.org/project/FlexiConc/>) integrated into CLiC (<https://clic-fiction.com/>), the tutorial will demonstrate how to apply robust concordance reading approaches to a variety of research contexts.

The tutorial starts with an introduction to concordance analysis, including its place in the continuum of quantitative and qualitative research. We cover the most common general strategies for concordance analysis: selecting, sorting, and grouping lines, and show how each of them can aid interpretation. Participants will also learn about basic formal definitions and mathematical properties of the computational algorithms that underlie these strategies. We will discuss how algorithms extend beyond simple alphabetical ordering, opening up new possibilities for advanced text analysis.

The tutorial will include practical exercises, in which participants explore the functionalities of the FlexiConc library and its web interface. This library is designed to support a wide range of concordance reading strategies and to document user decisions in a systematic way. The CLiC web interface is designed to be intuitively accessible and to enable convenient interactive exploration. We introduce the concept of an ‘analysis tree’ to ensure the reproducibility and accountability of concordance research. By using a tree structure to trace the decisions taken when selecting lines from concordances, ordering, and grouping them, we can document not only the final results but also the process that led there. This approach fosters transparency, which is crucial for collaborative and interdisciplinary projects, as well as for replicating or extending research.

## 1. Tutorial outline

### *Introduction to concordance analysis: fundamentals and strategies*

- Participants are introduced to basic concepts of concordance analysis. After a brief definition of fundamental terms and concepts, we give an overview of concordance software and its functionalities and allow participants to explore selected example concordances. Participants will be encouraged to share their observations on linguistic patterns as they work with existing concordancing tools.
- We introduce strategies for organizing concordances (different types of selecting, ordering, and grouping). Each strategy is discussed with regards to its purpose, and how it may be combined with other strategies. In a hands-on exercise, participants apply different strategies themselves to example data and compare their observations to those from the step before to see how the application of dedicated strategies helps with concordance organisation and enhances systematicity.

### *Computational algorithms*

- Participants are introduced to our algorithmic approach to concordance reading, which extends the basic strategies and enhances their flexibility.
- In a hands-on exercise, participants try out different concordance algorithms, including complex applications such as clustering, which are not widely

available in current concordance tools. They can work with the web interface of our library on a public server, so no software installation is required (but advanced participants are welcome to work directly with the Python library via Jupyter notebooks, which enables them to process their own corpus data).

### *Analysis trees for research documentation (~ 30 minutes)*

- We discuss reproducibility as a central challenge for concordance analysis and how this problem can be solved with the help of the ‘analysis tree’. The tree-like display, accessible through the web interface of our library, enables users to trace and illustrate decision-making during concordance analysis.

### *Summary and outlook (~ 20 minutes)*

## 2. Target audience

This tutorial targets an interdisciplinary audience, including students and researchers in corpus linguistics, general linguistics, computational linguistics, digital humanities, and computer-assisted language learning. We will keep the technical discussion to a manageable level to accommodate participants from both technical and non-technical backgrounds. Those interested in advanced techniques, such as more low-level concordance processing using Python, will be directed towards additional online

resources and are invited to attend a follow-up tutorial at the KONVENS conference.

### **3. Technical requirements**

Participants are encouraged to bring their own laptops, and technically more advanced participants may want to install Jupyter and FlexiConc on their own computer. However, a modern Web browser is sufficient to follow all hands-on exercises.