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.[1]

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.[2] 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).[3] 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.

[1] 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

[2] https://pushshift.io/signup

[3] 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 queryies. 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 severly 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).[4] 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 “discoursemes”, defined as minimal units of lexical meaning in the context of a specific discourse. Constellations of such discoursemes then indicate framings and discursive positions (consider e.g. a combination of the discoursemes MIGRANT, FLOOD, and MENACE). Since discoursemes 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 discoursemes 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[5] aims to integrate algorithms more flexibly and more tightly into the concordance reading process. Based on a mathematical taxonomy

[4] https://github.com/cqlf-ontology/

[5] 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.[6]

3.5. An Integrated CADS Platform

The ultimate goal, though, is the creation of an in 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).[7] As a first step, the MORCDA project aims for an experimental integration of MMDA and innovative concordance reading algorithms into Swiss-AL.[8]

4. References

Arnold, T. and Tilton, L. (2023). *Distant Viewing: Computational Exploration of Digital Images*. The MIT Press.

Baker, P.; Gabrielatos, C.; Khosravinik, 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 an 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). Narrlangen 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

[6] https://pypi.org/project/FlexiConc/

[7] https://swiss-al.linguistik.zhaw.ch/

[8] 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 discoursemes 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 (OASIcs)*, 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.