Tatjana Scheffler

Social media corpora for analyzing linguistic variation

Abstract: Computer mediated communication (CMC) has become a popular source of data for analyses in linguistics and social science, aided by convenient access to large-scale ad-hoc corpora. While the medium has been in focus as an influencing factor on linguistic expression in CMC for a long while, I argue that other factors have similarly significant effects on individual linguistic variation in online texts. In the paper, I address the interplay of variation by topic, register, and individual user with the medium of social media communication. I develop best practices for constructing CMC corpora that allow research into intra-author variation, by controlling for other factors that may confound results based merely on the comparison of different pre-existing corpora.

I then present one case study for the construction of a CMC corpus that demonstrates linguistic variation across two social media within the same group of authors. In particular, there is considerable inter- and intraindividual variation in linguistic features of informal, spontaneous and situated communication such as the use of emojis. Large CMC corpora with open research licenses, rich metadata and linguistic annotations thus make it possible to tease apart the particular effect of the factors medium, register, topic, and individual author on linguistic phenomena.

Keywords: CMC, corpora, individual variation, medium, sociolinguistics, emojis

1 Introduction

Individual language users show distinct individual patterns in their linguistic expressions. These individual patterns may align with the patterns shown by others sharing certain demographic properties (e.g., gender, age, community, etc.), or may be idiosyncratic. The field of sociolinguistics has its central focus on this kind of individual linguistic variation. Sociolinguistics and corpus linguistics as applied to corpora of computer mediated communication (CMC) share a common goal, the collection and study of 'the language used by ordinary people in their everyday

Tatjana Scheffler, Ruhr-University Bochum, e-mail: tatjana.scheffler@rub.de

https://doi.org/10.1515/9783111434018-014

affairs' (Labov 1972: 69). Where traditionally, variational sociolinguists have been primarily interested in elicited and everyday spoken language, studied in-depth for individual speakers, CMC corpus researchers study spontaneous (i.e., not elicited) language use in the written mode, often produced by a large number of different users. The intersection between these two interests, investigating the effects of social properties on linguistic expression and the study of naturally occurring spontaneous written data in CMC corpora, has been an active area of interest as well (c.f. Androutsopoulos 2000; Tagliamonte and Denis 2008; Androutsopoulos 2006, 2011; Herring, Stein and Virtanen 2013; Bock, Busch and Truan 2023, and many others).

The methodologies of these two subdisciplines differ: The primary method in classical sociolinguistics is the sociolinguistic interview as well as additional in-depth observations of the linguistic behavior of individuals, for example their conversations at work or in their friend group (discussed in detail in Meyerhoff 2016, where she also distinguishes sociolinguistic and corpus linguistic approaches). In contrast, social media corpus research mainly assembles language data from many individuals on one platform and analyses this data in an aggregated way. The disadvantage of such pure corpus research based on one source is that it can be difficult to draw conclusions about the linguistic system used by each speaker. Even when comparing several such corpora, two corpora may differ for many reasons, including language and speaker external ones such as the medium or topic of conversation. In addition, two separately collected corpora contain data from distinct sets of language users, and in the case of CMC corpora, these language users often belong to distinct communities, based on their age, place in society, or their interests.

A very much simplified view of the matter may therefore be that on the one hand, the sociolinguistic method permits the study of individual linguistic variation and enables us to draw conclusions about the underlying linguistic systems, but does not typically have access to the large data sets of CMC. On the other hand, CMC corpus linguistics studies spontaneous, natural linguistic expressions in social media corpora, but so far shows limited potential for investigating individual linguistic variability due to excessive aggregation, which permits only group-level comparisons. In the following, I will propose an approach for analyzing individual linguistic behavior on social media. I will develop some best practice recommendations for collecting social media corpora that support this kind of research.

2 Causes of linguistic variability in computer mediated communication

It is a well-known and well-studied fact about language that linguistic expressions depend on the communicative situations they are uttered in. For example, it is immediately obvious that (1) was originally a spoken utterance while (2) surely originated from a written source (both examples have been translated from their German originals from the parallel blog and podcast corpus PARADISE; see Seemann et al. 2023). We can understand this immediately, even though both samples are quite short and presented outside of their original context in written form.

(1) So a lot of practical policy has been made here
A lot of language policy been made
Specialist terminology developed
For example, also, soccer terminology comes from this
Take a look at French, they say 'penalty', a real anglicism
We say 'Strafstoß'
They say 'futbol', we say 'Fußball', etc. These were all fairly welldeveloped core German words at the time that were invented and developed
[FG007_Transkript]

(2) Time and again we read and hear of displeasure about the state of the German language: new spelling rules have been introduced - and then immediately withdrawn. German is losing its international significance. English words are flooding everyday language usage. In short, the selfimage of our language seems to have developed unsightly cracks. Peter Eisenberg, however, takes a relaxed view of linguistic developments and is not afraid of anglicisms.
[FG007_Blog]

But spoken vs. written presentation mode is not the only axis of variation for linguistic expressions. Koch and Oesterreicher (1985) argue that while the distinction between speech and writing is a categorical, binary one, individual media of communication differ more gradually between prototypically, 'conceptually' oral language (such as a spoken conversation between friends) and prototypically, conceptually written language (such as a legal text). According to Koch and Oesterreicher (1985), the pole of conceptual orality is characterized by communicative conditions typical for the 'language of closeness': spontaneity, dialog, expressivity, co-presence, etc. In contrast, the pole of conceptual writing is characterized by the 'language of distance': prior planning, monologue, separation of place and time, detach-

ment, objectivity, etc. These different conditions for utterances are reflected in the linguistic phenomena that we can observe in them, for example, first and second person pronouns in conceptual orality and almost exclusively third person pronouns in conceptually written material (Yates 1996; Tagliamonte and Denis 2008). It is proposed that communicative settings (telephone call with a friend, diary entry, job interview, scientific talk, academic paper) can be arranged along the conceptual orality dimension, as they correspond to varying degree to one of the two poles in both conditions of communication and means of linguistic expressions.

This multi-faceted view of communicative setting and corresponding linguistic phenomena carries over to social media as well, which typically use written mode,[1] but can vary a lot wrt. their communicative conditions and the linguistic phenomena they exhibit. To investigate both the communicative conditions of specific media platforms, as well as the linguistic phenomena which occur based on the affordances the specific media offer to their users, a whole range of CMC corpora have been collected and linguistic research using these corpora has been documented, not least in the CMC Corpora conference series (Hendrickx, Verheijen and van de Wijngaert 2021, and previous editions).

Linguistic variation can be observed between social media, but also within individual media, due to the fact that communicative situations differ starkly even within a medium (Koch and Oesterreicher 1985; Dürscheid 2003): a text message chat with a friend will exhibit linguistic features that cannot be observed in a chat with a prospective new landlord when applying for an apartment, maybe starting with the frequency of emojis. Thus, comparing two CMC corpora is likely to lead to systematic changes in some linguistic variables if their communicative situations do not match. The point that register affects linguistic variables has also been empirically demonstrated within a given social medium, which may be used to interact in different registers, such as narrative, informative, or persuasive. Scheffler, Kern and Seemann (2022) show that the use of German modal particles and intensifying particles varies along these register dimensions, even within a given CMC medium (blog posts or tweets).

Since other factors influence linguistic expression in CMC, comparing linguistic data across different corpora invites the intrusion of confounds. We can not always be sure that the differences that are necessarily found between two corpora, since no two sets of text can be identical, can be linked back to the medium distinctions.

1 Social media can be implemented in a variety of modalities, including written text (blogs, Facebook, forums), speech (podcasts, voice messages), images (Instagram, Facebook, Pinterest, chat programs), video (Youtube, TikTok, Instagram reels), or combinations of all of them. In this paper, I focus on written social media corpora for practical reasons.

Even within a single medium, significant differences are found between subcorpora, and the dimensions of variation are not limited to communicative situation and register. For example, Schler et al. (2006) investigated gender and age effects on blog texts and found clear differences between blogs written by female and male authors. However, the most noticeable differences on the level of word frequencies they found are at most indirectly related to gender, since they mainly reflect topic differences (tech related words like *linux*, *programming* lean heavily male, while words such as *shopping*, *mom* predict a female author). In authorship analysis or profiling, which aims at identifying linguistic variation based on individual preferences or linked to demographic properties of the individual author, content words are often ignored for this reason, as they may reflect the topic of a text more than properties of its author. We can thus note topic as an additional cause of variation in CMC texts.

As an interim summary, we can note that linguistic expressions may be affected by many aspects that characterize a given text, such as mode, text type, register, topic, demographic (age, gender) or group properties of the author (hobbies, subculture), or individual idiosyncrasies. In order to tease apart which part of the linguistic variability observed in a corpus is due to the underlying mechanisms of the linguistic system, and which part is affected by these to some extent language-external factors, it is important to try to control for these effects when comparing data in corpus linguistics. I will therefore aim to construct a CMC corpus which exhibits individual variability but is matched for mode, register, topic, as well as author properties.

3 Constructing CMC corpora for individual linguistic variation

To investigate naturalistic language use in CMC, corpora of CMC have to be constructed and studied. The availability of such naturally occurring language in everyday use from CMC is simultaneously affected both positively and negatively: On the one hand, linguistic production by users on various digital media is constantly increasing in quantity. On the other hand, recent tendencies in restricting the open web make it harder for researchers to access and ethically source that data. However, it is necessary for academic research not only in linguistics to continue to strive for broad access to a representative sample of CMC, since restricting ourselves to only the most accessible, abundant data sources would lead to a very biased view of language (as well as of the topics and contents represented in online media).

3.1 Principles of CMC corpus construction

Principles of open datasets (such as the FAIR principles, Wilkinson et al. 2016) apply with particular urgency to the construction of CMC corpora, since only openly available and reusable data can be *sustainable* in two senses of the word: (i) CMC corpora must be sustainable as in cost-effective, since it is often complex and expensive to construct them, much less to pre-process and annotate the data for linguistic analysis. Only by making our data available to other researchers can we get our "money's worth" and speed up scientific progress. (ii) Results of CMC corpora should be long-lasting and verifiable even after a project's end. The case of the demise of Twitter and subsequent closure of its APIs has shown that corpora that are not openly shared in the research community can be rendered worthless in a minute. Not only reviewers, but also other researchers must be able to verify and build on previous research results.

In the case of CMC corpora, though, the data is typically produced by a large number of private persons who are using language for their idiosyncratic, private communication needs. A third consideration in constructing CMC corpora is therefore the ethics of data collection and redistribution:[2] (iii) CMC corpora for studying individual linguistic behavior must necessarily contain data produced for personal purposes by many individuals, who may also reveal potentially identifying information. Ethical (as well as legal) concerns dictate that personal data collected should be as minimal as possible, and that if possible, data authors should be consulted or at least informed.

Luth, Marx and Pentzold (2022) develop best practice guidelines for ethically and legally responsible CMC data collection for research which they instantiate in several case studies, from the researchers' perspective. Fiesler et al. (2024) conducted a meta analysis of studies using Reddit data and conclude with ethical guidelines for such research. As a main takeaway, they co-opted Reddit's first user clause, "Remember the human". In order to preserve the linguistic features under investigation, the expressions often cannot be transformed enough to fully conceal authorship to a dedicated observer. Thus, fully anonymized corpora are not possible, and

2 I will not really touch on legal issues here. While important, they underlie dynamic processes and furthermore in this domain are often subsumed by ethical obligations scientists have wrt. the individuals they are studying as well as the society that benefits from their research. What I mean is that ethical constraints on data collection and use are often much more restrictive than legal constraints; in addition, the consequences of the violation of ethical constraints in my view weigh more heavily. For a detailed discussion of legal considerations see e.g. Beißwenger et al. (2017).

linguistic researchers, as well, should "remember the human" behind the data they collect and analyze.

3.2 How not to construct CMC corpora for investigating individual variation

There are several non-ideal ways in which individual variation in social media could be investigated, which I will briefly sketch here to contrast them with the approach proposed below. One may try to work exclusively with existing corpora and compare them to find systematic linguistic effects. As discussed above, this will not reliably lead to the observation of inter- or intra-individual variation, as external factors and the selection criteria of corpora may carry too much weight. For example, many social media corpora have been collected using specific keywords (e.g., "Corona, COVID") which by necessity limits the kinds of linguistic expressions found based upon such a search. In addition, many existing corpora contain only very few contributions or even single posts from each individual author, and/or do not contain sufficient metadata to link authors between posts. In such corpora, intra-individual variability cannot be observed.

In order to study linguistic variability, a larger set of CMC posts is needed for any given author – similar to the in-depth sociolinguistic interviews and observations across various situation used by sociolinguists studying the oral mode. If the dataset doesn't contain clear identifying information, computational linguists have tried to reconstruct demographic information about the authors, either by using human crowd workers, for example to manually annotate profile pictures for gender (Ciot, Sonderegger and Ruths 2013), by using the available textual data (Nguyen et al. 2016, 2021), or by using network effects such as homophily (Li, Ritter and Hovy 2014). All methods share several drawbacks. First, they exhibit significant error rates; second, inducing demographic properties from the data in order to then use it in comparing the linguistic behavior of demographic subgroups may lead to circular reasoning that serves to confirm pre-existing biases; and third, it is not clear that users consent to or are even aware of the possibility of tracking their long-term behavior on social media and their personal identity information (such as age, gender, occupational status, ethnicity, etc.).

In the past, certain platforms have tried to aggregate information on social media profiles across other platforms for a given user, which enables linking different profiles. The most usable version of this was Google+, which provided an API for reading the profiles and made it possible to scrape individualized corpora across various CMC platforms. However, the interface was closed in early 2019 and similar tools have not become available. In a more limited fashion, users some-

times self-identify alternative accounts on other platforms by linking to them in their profiles. In the next section, I will use this selflinking to create a cross-media corpus of individual CMC communication.

3.3 How to construct CMC corpora for investigating individual variation: Principles of best practice

It has been noted, for example via surveys, that authors interacting on social media, even public ones, are often not aware of the possibility that their data may be scraped and used by companies or researchers (Fiesler and Proferes 2018).[3] If they are asked about the use of their text for research, authors generally state that they would like to be asked or at least informed.

Legal as well as ethical considerations further require that the personal information of authors should be protected. Thus, private information cannot be ethically collected by researchers, unless the authors explicitly agree to this use, via donation,[4] or if users are active in a platform specifically meant for research use.[5] Private information that is not publicly shared should also not be reconstructed automatically or with the help of crowd workers, as authors may not intend to share this personal information.

Finally, the case of Twitter's demise (only the most impactful in a long string of CMC platforms that have disappeared or changed ownership) has demonstrated that relying on software interfaces provided by technology platforms puts researchers at constant risk for losing their data or losing access to published corpora and losing the ability to reproduce research results. Even large data platforms such as GitHub are just one sale or strategic decision away from making years of research disappear. Thus, it is important to develop methodologies that use free and open tools and simple techniques of the internet (web scraping) as much as possible to collect data. Further, data should be shared as widely and comprehensively as possible in order to ensure both the reproducibility of existing results as well as the reuse of precious resources.

Based on these observations I propose the following best practice principles for sustainable CMC corpus research for linguistic variation.

---

3 Breuer et al. (2024) showed that virtually none of the servers on the decentralized platform Mastodon discuss whether or not their data can or should be used for scientific research.

4 One example is the MoCoDa2 chat database https://db.mocoda2.de (last accessed 14 February 2025).

5 E.g. the platform used in (Beißwenger and Pappert 2019), if users were informed about the text collection

3.3.1 Data collection

Consent. Gather consent prior to data collection if possible (opt-in), or at minimum inform authors of the data collection and give them the option to have their data deleted (opt-out).

Prior consent is possible in the case of clearly delineated author groups such as in more typical sociolinguistic studies, or when linguistic research is carried out within and benefits a specific community. At least a minimum effort should be made to post-hoc inform authors of data collection if prior consent is not feasible. For example, it is often possible to inform a community via their platform moderators and to provide contact options for opting out of datasets.

Use public data. Gather only public data; private data, e.g. from chat systems, should be collected only via donations or with explicit consent prior to the production of the data. In general, private data that requires a login is not freely viewable and underlies specific restrictions.

Web scraping. Collect textual data via the web, making use of legal permissions for text and data mining for research (e.g., the German §60d UrhG).

Web scraping operates by simulating a web browser interface and the clicks a human user would make, while collecting the data presented to a human user viewing the platform's content. As long as it considers public data, this interface should always be open and available for data collection, for example by using JavaScript tools. However, additional effort may be required, particularly for creating scraping tools and data representations that can capture this content. Also, such tools must often be adapted when an interface changes.

Limit metadata. Collect only self-identified user metadata. Do not attempt to reconstruct personal information such as gender or sexual orientation (and many more) beyond the information explicitly shared by the authors themselves.

Metadata is often very valuable for (socio-)linguistic research. Self-provided metadata is easy to collect when it is presented to the public on a platform, and much more reliable than automatically inferred metadata which may be prone to enhance biases.

Anonymization. Anonymize or pseudonymize the data thoroughly in order to avoid harm, if necessary manually. Remove personal information from own records.

This requires a significant effort on the researchers' behalf but can be mitigated by sharing resources and corpora. Some semi-automatic tools can help, but

typically a corpus must be anonymized manually if there is a danger of exposing personal information, for example when a platform uses real-world user names.

3.3.2 Research and distribution

After data collection, ensuring the sustainability of the data is in the researchers' hands:

- Archive everything.
- Annotate full datasets ensuring high data quality and save annotations in a reusable format, independently of tools via which the data can be used (e.g., XML, tabular text formats such as CSV).
- Share all (anonymized) data, including annotations, with reviewers and other researchers on request via private links, or if possible openly on the web.
- Extract the informative or relevant linguistic data from CMC posts to share freely and separately from any personal information (e.g., see the data set of extracted English it-clefts in Bevacqua and Scheffler 2020).
- Develop derived text formats that can be freely shared with anyone via repositories and the web (Schöch et al. 2020).

4 A corpus for linguistic variation in CMC

In the following I will describe an approach for collecting a cross-media corpus aimed at investigating individual linguistic variation. The best practices described in the previous section have been developed in part by drawing upon the experiences in constructing this corpus, as well as others. They are therefore not yet all followed to the fullest in this effort.

Given the observations in Section 2, the central goal for the corpus consisted of the following criteria: It should contain naturalistic CMC data from a selection of people. It should cover at least two different media in order to enable cross-media comparison. To investigate intra-speaker adaptation to the medium, the identical users should be represented in the subcorpus for each medium, and we should be able to identify authors across the two (or more) media. The topics and registers should match as much as possible across the social media, in order to minimize the influence of topic and register on the linguistic expressions. This would mean that any remaining variation could be traced either to the medium or to the individual behavior of users. The texts itself should be spontaneous and not too restricted (e.g., public speeches or newspaper articles adhere to many externally imposed norms

and rules and only partially reflect "everyday language" used for free communicative goals).

The corpus was constructed by collecting tweets and blog posts from users belonging to the German parenting blogger community. It draws centrally on a human curated list of "parenting bloggers" on Twitter from 2017, called the "Elternbloggerkarte".[6] The list aggregated Twitter accounts around a single topic or community, parenting, and already imposed the filter that the users included are active both on the platform Twitter (their accounts were listed) as well as operated a blog (inclusion criterion for the list). The focus on this community has the advantage that it is not too specialized: Tweets and blog posts by users from this community generally relate to family life and daily events.

Constructing the corpus consisted of several steps:

1. The Twitter list was read out using Python and the Twitter API, and included 195 members. Three additional prolific parenting bloggers and tweeters were manually added to the list.
2. All available tweets (in cases of very active users, the most recent 3200 tweets) were collected for each profile using tweepy.
3. Each Twitter profile was crawled using the tweepy package and the Twitter API, to collect the URL linked in there. In the parenting blogger community, this URL in most cases links to a personal blog.
4. The URL was manually cleaned and links to Facebook and other non-blog websites were removed. For the remaining URLs, the Python package feedreader was used to automatically scrape the RSS feed, if available, and retrieve the most recent available blog posts (usually, 5–10 posts).
5. Blog posts were cleaned from boilerplate via BeautifulSoup.

The collection process used the APIs available at the time. More recently, those programmatic interfaces to the Twitter platform's content have been closed and the new instantiation, X, does not provide this kind of access in the same way. A possible alternative to the use of APIs is web scraping.[7] Scraping has the advantage that it exploits the public interface of a platform, which is always available via a web browser for a public social media platform. A possible disadvantage is the limited availability of metadata (such as user networks and other internal information that

6 *Elternbloggerkarte* ('parenting blogger map') is a project to log the physical locations of bloggers active in the parenting community, starting from Germany. The map is still available here: https://familiert.de/elternbloggerkarte/ (last accessed 14 February 2025).

7 Luca Hammer regularly makes scraping tools available: https://github.com/lucahammer (last accessed 14 February 2025).

is not openly displayed to viewers). However, a lot of the data interesting for linguistic research is still available on the web, albeit with more effort than before.

In addition, it must be noted that it is not always possible to link authors' accounts on different platforms to each other in order to capture cross-platform linguistic adaptation. There are two ways in which this can be achieved: First, via self-tagging by users as in this case. On many CMC platforms, user profiles include links to other providers. For example, this is the case for many Fediverse platforms such as Mastodon, on platforms such as Wikipedia, GitHub, or on personal blogs. Users are often happy to identify their alternative activities elsewhere on the web. For some communities, dedicated personal identifiers have even been created to uniquely identify each individual (such as with OrcID for scientists). Second, in studies based on data donation, users can be asked to provide textual data from not only one platform but several, such that the linguistic output can be linked.

The data collection for the parenting corpus was carried out in Spring 2017; tweets were acquired between February 14–16, blog posts on February 20, 2017. Both tweets and blog posts were obtained for 62 users, after some quality checks (for example, users who primarily posted in a language other than German were removed). While the data collection for research purposes follows §60d of the German *Urheberrechtsgesetz* (copyright law), we nevertheless retroactively informed the users and obtained their consent prior to analysis and distribution of the corpus. All blogs were consulted in 2020 to retrieve contact information for their operators.[8] All users were contacted and asked to explicitly respond if they do not want their data included in the corpus (opt-out). Out of 50 users who could be contacted (some blogs had become inactive), three asked for their data to be removed, six explicitly agreed after some additional questions to be included in the corpus, and the remainder quietly acquiesced to inclusion. Indeed several people even responded positively by actively agreeing to being part of the corpus and showing an interest in the results. All data from users that could not be contacted or asked to be removed was deleted, yielding a final corpus of 44 users for whom both tweets and blog posts in sufficient quantity could be collected. The corpus data is summarized in Table 1, the corpus is available as the "TWItter and BLOgs COrpus: Parenting" (TwiBloCoP),[9] in raw text format or TEI-XML.

8 Contact information is a legal requirement for public or commercial web sites in Germany.

9 https://staff.germanistik.rub.de/digitale-forensische-linguistik/forschung/textkorpus-sprachliche-variation-in-sozialen-medien/ (last accessed 14 February 2025).

Table 1: Size of the TwiBloCoP corpus.

	blog posts	tweets
users	44	44
posts	468	81,440
tokens	~360,000	~1,200,000

In the following, several data preprocessing steps were carried out, including anonymization of all data, sentence splitting and tokenization, as well as part of speech tagging. Anonymization was applied manually by replacing all personal names, blog names, emails, places, @usernames, urls, and phone numbers with placeholders in brackets such as [NAME], [PLACE] etc. The users were assigned random 4-digit ID numbers to link tweets to their corresponding blog posts that share the same author. The sentences and tokens were then automatically split using the Python package SoMaJo,[10] and part of speech tagged with SoMeWeTa.[11]

Topic-wise, the corpus is quite homogeneous: Both blog posts and tweets are concerned with family life and parenting, see examples (3)–(4).[12] Any remaining linguistic variability can then be traced to individual variability (by observing authors across their different texts) or cross-medium variability (by showing tendencies across different authors within a medium).

(3) Children are our mirrors. If you want to change your child, change YOUR behavior, not the child's. My son has these tantrums all the time. Regularly. Then it is very difficult to get him out of it. And that is exactly what I would like to do. […] Hm. At some point I asked myself why these fits upset me so much.
[blog-4421-10]

(4) Alarm rang every 5 minutes since 6 am. Got up right before 8. Great. Worked like a charm 🙈
[tweets-7291]

Looking specifically at intensifying and modal particles, Scheffler et al. (2022) have shown that in addition to the medium and the individual author, register still remains as an additional source of variability in the corpus. They show that within

10 https://github.com/tsproisl/SoMaJo (last accessed 14 February 2025).
11 https://github.com/tsproisl/someweta (last accessed 14 February 2025).
12 All examples have been translated from German.

the general parenting topic, individual texts and text parts differ wrt. whether they are intended to convey information in a relatively neutral fashion, whether they are meant to convince or argue, or whether they are mainly meant to tell a narrative story about the author's personal life. The corpus has been completely manually annotated for register, so that this dimension can also be taken into account in studies of variability.

5 Individual variation in social media

A corpus with texts from the same authors in two social media makes it possible to study both how individuals differ from each other when communicating about the same topics in the same media, as well as how authors adapt to the medium while communicating similar content. In the aggregate, simple complexity measures such as type-token ratio (computed over the first 1000 tokens of each text, tweets are aggregated by user) and average word length show that the blog posts and tweets are of medium linguistic complexity, right in between spoken conversations and newspaper texts (see Figure 1).

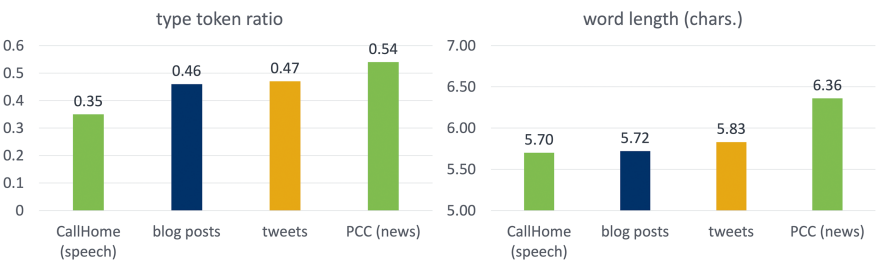

Figure 1: Type-token ratio (left) and average word length (right) for German tweets and blog posts from the TwiBloCoP corpus, compared to telephone conversations (CallHome) and newspaper commentaries (PCC).

While the blog posts and tweets are quite similar to each other in complexity (interestingly and maybe unexpectedly, the tweets are slightly more complex than the blog posts according to these measures), they show stark differences wrt. the frequency of non-standard spelling such as word lengthening by letter reduplication (*niiiiiice*), across-the-board capitalization (*AWESOME*), or socalled inflectives marked with asterisks (**yawn**), and particularly the presence of emojis (Figure 2).

These phenomena are virtually nonexistent in standard monological written media such as newspaper texts.

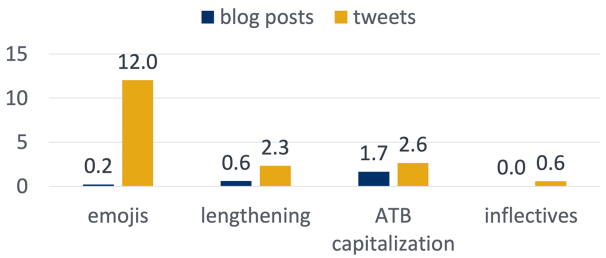

Figure 2: Frequency of non-standard items in blog posts and tweets, per 1000 tokens.

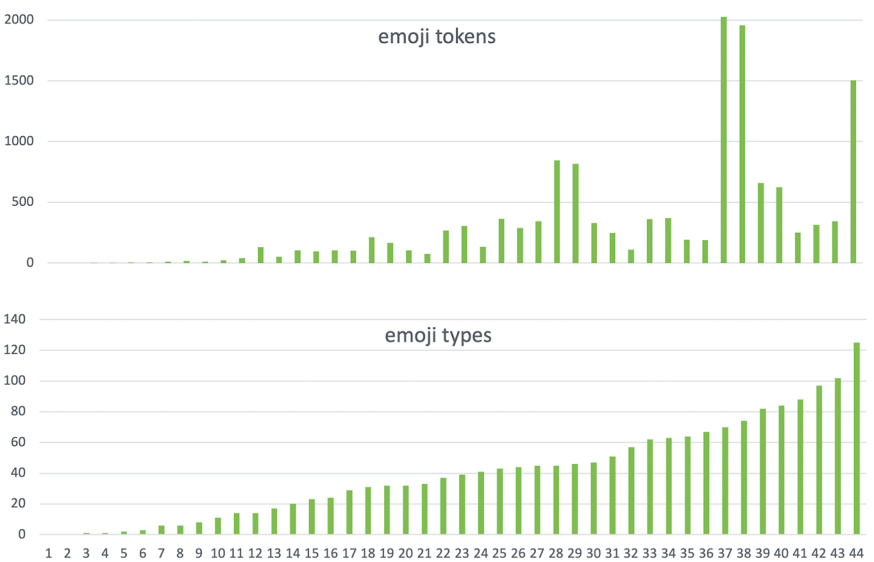

Figure 3: Individual variation in emoji usage frequency among authors in TwiBloCoP; each bar in the top graph represents the same author as in the corresponding bar in the bottom graph.

It is interesting to have a closer look at these differences between the two CMC subcorpora, because both media contain the same users writing about very similar topics. However, while they use almost exclusively standard graphematic tools in their blog posts, the same users are much more likely to use non-standard writ-

ing phenomena in their tweets. Looking specifically at the emojis (as the most frequent phenomenon), we can observe significant individual variation in their use. For example, the 77 emojis that occur in blog posts are used by only 13 of the 44 authors, and only 3 of them use emojis more than 5 times in their blog (the large majority of these emojis are red hearts). In Twitter, the distribution of emojis is also not uniform among authors. 14 authors use fewer than 100 emojis in total in their tweets, while the overall high frequency of emojis is mostly due to three power users, see Figure 3. In addition, authors also show diverse amounts of internal emoji variation: most use fewer than 50 different emojis, while some pick from a much larger variety. The individual style of emoji use can be compared between these three authors by observing the "emoji clouds" generated from their subcorpora (Figure 4).

6 Conclusion

While the fact that CMC as a domain of linguistic research shows great variety is perhaps well-known, this paper made the point that even individual media or text types of computer mediated communication are not monolithic. In each medium, many individual users congregate to express their own personalities and idiosyncratic linguistic strategies. We can only characterize the linguistic system employed in a corpus if we can make reference to this individual linguistic variability. And adversely, we should be able to know to what extent linguistic variability observed in large corpora is due to the medium, register, topic, or individual properties of the author.

To enable such research, I have presented an approach for collecting CMC corpora that expose individual linguistic variability. One case study is the Twitter and Blog Corpus – Parenting, in which the same authors are represented across two social media. Constructing it has helped develop a list of best practices for the collection and distribution of social media corpora for research into the everyday language use in the digital domain.

Acknowledgements

I would like to thank the reviewers of this volume for their helpful comments and the editors for their patience. I am grateful to the audience at the CMC Corpora conference in Mannheim for their questions.

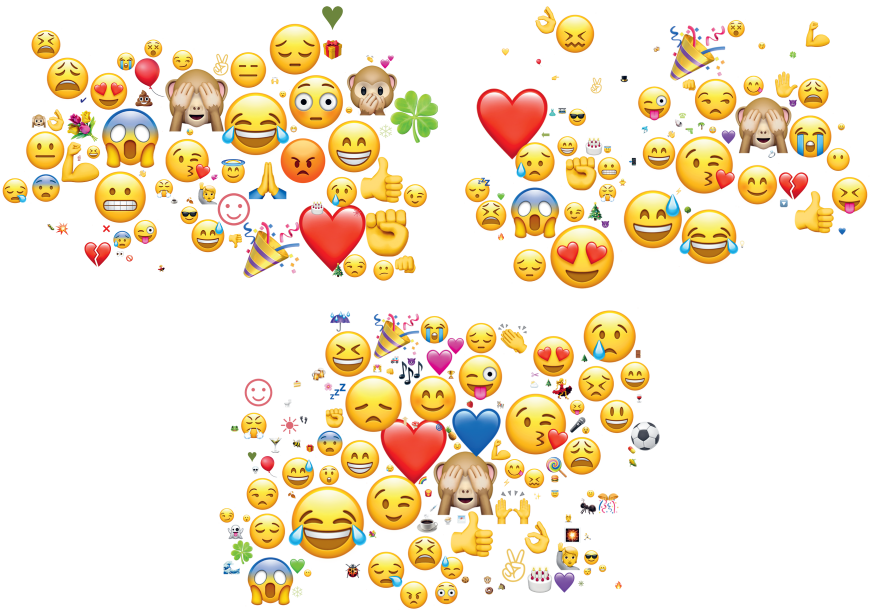

Figure 4: Emojis used by the three authors with the most emoji tokens.

Hannah Seemann has contributed significantly to the creation and development of the TwiBloCoP corpus, many thanks to her and to Lesley-Ann Kern for their work on the corpus and to our student annotators for their contributions. All remaining errors are my own.

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

References

Androutsopoulos, Jannis. 2006. Introduction: Sociolinguistics and computermediated communication. *Journal of Sociolinguistics* 10 (4). 419–438.

Androutsopoulos, Jannis. 2011. From variation to heteroglossia in the study of computer-mediated discourse. In Crispin Thurlow & Christine Mroczek (eds.), *Digital discourse: Language in the new media*, 277–298. Oxford University Press.

Androutsopoulos, Jannis K. 2000. Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics* 4 (4). 514–533. doi: 10.1111/1467-9481.00128. http://doi.wiley.com/10.1111/1467-9481.00128.

Beißwenger, Michael, Harald Lüngen, Jan Schallaböck, John H. Weitzmann, Axel Herold, Pawel Kamocki, Angelika Storrer & Julia Wildgans. 2017. Rechtliche Bedingungen für die Bereitstellung eines Chat-

Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In Michael Beißwenger (ed.), *Empirische Erforschung internetbasierter Kommunikation*, 7–46. Berlin & Boston: De Gruyter.

Beißwenger, Michael & Steffen Pappert. 2019. How to be polite with emojis: a pragmatic analysis of face work strategies in an online learning environment. *European Journal of Applied Linguistics* 7 (2). 225–254. https://doi.org/10.1515/eujal-2019-0003. https://www.degruyter.com/view/journals/eujal/7/ 2/article-p225.xml (last accessed 14 February 2025).

Bevacqua, Luca & Tatjana Scheffler. 2020. Form variation of pronominal itclefts in written English: A corpus study in Twitter and iWeb. *Linguistics Vanguard* 6 (1). 20190066. https://doi.org/10.1515/lingvan-2019-0066. Bock, Cornelia F, Florian Busch & Naomi Truan. 2023. Introduction: The sociolinguistics of exclusion–indexing (non) belonging in mobile communities. *Language & Communication* 93. 192–195.

Breuer, Johannes, Marco Wähner, Annika Deubel & Katrin Weller. 2024. Collecting and archiving Mastodon data: Ethical enquiries on decentralized networks. Talk presented at the DNB conference "After Twitter". https://wiki.dnb.de/pages/viewpage.action?pageId=337119232 (last accessed 14 February 2025).

Ciot, Morgane, Morgan Sonderegger & Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu & Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. Seattle, Washington, USA: Association for Computational Linguistics. https://aclanthology.org/D13-1114 (last accessed 14 February 2025).

Dürscheid, Christa. 2003. Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit: Theoretische und empirische Probleme. *Zeitschrift für Angewandte Linguistik* 38. 37–56.

Fiesler, Casey & Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media + Society* 4 (1). https://doi.org/10.1177/2056305118763366.

Fiesler, Casey, Michael Zimmer, Nicholas Proferes, Sarah Gilbert & Naiyan Jones. 2024. Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. *Proceedings of the ACM on Human Computer Interaction* 8 (GROUP). 5:1–5:33. https://doi.org/10.1145/3633070. https: // dl.acm.org/doi/10.1145/3633070 (last accessed 14 February 2025).

Hendrickx, Iris, Lieke Verheijen & Lidwien van de Wijngaert (eds.), 2021. *Proceedings of the 8th Conference on Computer-mediated Communication CMC and Social Media Corpora (CMC-Corpora 2021)*. Nijmegen, NL: Radboud University. https://surfdrive.surf.nl/files/index.php/s/Lcgx6d3EwGMjugR (last accessed 14 February 2025).

Herring, Susan C., Dieter Stein & Tuija Virtanen. 2013. Introduction to the pragmatics of computer-mediated communication. In Susan C. Herring, Dieter Stein & Tuija Virtanen (eds.), *Pragmatics of computer-mediated communication*, 3–31. Berlin & Boston: De Gruyter.

Koch, Peter & Wulf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36. 15–43.

Labov, William. 1972. *Language in the inner city: Studies in the Black English vernacular* 3. Philadelphia: University of Pennsylvania Press.

Li, Jiwei, Alan Ritter & Eduard Hovy. 2014. Weakly supervised user profile extraction from Twitter. In Kristina Toutanova & Hua Wu (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 165–174. Baltimore, MD: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1016. https: //aclanthology.org/P14-1016 (last accessed 14 February 2025).

Luth, Janine, Konstanze Marx & Christian Pentzold. 2022. Ethische und rechtliche Aspekte der Analyse von digitalen Diskursen. In Eva Gredel (ed.), *Diskurse digital: Theorien, Methoden, Anwendungen*, 99–134. Berlin & Boston: De Gruyter. https://doi.org/10.1515/9783110721447-006.

Meyerhoff, Miriam. 2016. Methods, innovations and extensions: Reflections on half a century of methodology in social dialectology. *Journal of Sociolinguistics* 20 (4). 431–452. https://doi.org/10.1111/josl.12195.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé & Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics* 42 (3). 537–593. https://doi.org/10.1162/COLI_a_00258. https://aclanthology. org/J16-3007 (last accessed 14 February 2025).

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg & Theo Meder. 2021. "How old do you think I am?" A study of language and age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 7 (1). 439–448. https://doi.org/10.1609/icwsm.v7i1.14381. https://ojs.aaai.org/index.php/ICWSM/ article/view/14381 (last accessed 14 February 2025).

Scheffler, Tatjana, Lesley-Ann Kern & Hannah Seemann. 2022. The medium is not the message: Individual level register variation in blogs vs. tweets. *Register Studies* 4 (2). 171–201. https://doi.org/10.1075/rs.22009.sch.

Schler, Jonathan, Moshe Koppel, Shlomo Argamon & James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Aaai spring symposium: Computational approaches to analyzing weblogs, vol. 6*, 199–205.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann & Jörg Röpke. 2020. Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/ 2020_006. https://zfdg.de/2020_006 (last accessed 14 February 2025).

Seemann, Hannah, Sara Shahmohammadi, Manfred Stede & Tatjana Scheffler. 2023. PARADISE: A German PARAllel DIScoursE annotated multi-media corpus. https://doi.org/10.17605/OSF.IO/59ACQ. https://doi.org/10.17605/OSF. IO/59ACQ.

Tagliamonte, Sali A & Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83 (1). 3–34.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (1). 160018. https://doi.org/10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618 (last accessed 14 February 2025). Publisher: Nature Publishing Group.

Yates, Simeon J. 1996. Oral and written aspects of computer conferencing: A corpus based study. In Susan C. Herring (ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, 29–46. Amsterdam: Benjamins.