

[WLG]

WIENER LINGUISTISCHE GAZETTE

Resolving ethical issues in an online corpus of mixed public-private messages

A reflexive account

Jenia Yudytska

Sonderdruck aus: *Wiener Linguistische Gazette* 95 (2024): 111–135

Eigentümer und Verleger:

Universität Wien, Institut für Sprachwissenschaft
Sensengasse 3a
1090 Wien
Österreich

Herausgeberschaft:

Jonas Hassemer, Florian Grosser & Carina Lozo (Angewandte Sprachwissenschaft)

Erweiterte Redaktion:

Markus Pöchtrager (Allgemeine Sprachwissenschaft)
Stefan Schumacher (Allgemeine und Historische Sprachwissenschaft)

Kontakt: wlg@univie.ac.at

Homepage: <http://www.wlg.univie.ac.at>

ISSN: 2224-1876

NBN: BL,078,1063

Die *Wiener Linguistische Gazette* erscheint in loser Folge im Open-Access-Format.
Alle Ausgaben ab Nr. 72 (2005) sind online verfügbar.



Dieses Werk unterliegt der Creative-Commons-Lizenz CC BY-NC-ND 4.0
(Namensnennung – Nicht kommerziell – Keine Bearbeitungen)

Resolving ethical issues in an online corpus of mixed public-private messages: A reflexive account

Jenia Yudytska*

Wiener Linguistische Gazette (WLG)

Institut für Sprachwissenschaft

Universität Wien

Ausgabe 95 (2024): 111–135

Abstract

Dieser Beitrag befasst sich mit den ethischen Fragen, die bei der Arbeit mit privaten und öffentlichen Daten derselben Personen auftreten können. Für eine Untersuchung des Einflusses von Kommunikationsgeräten (Computer, Smartphone) auf die Sprache in der computervermittelten Kommunikation erstellte ich ein neues Korpus, welches aus Nachrichten von denselben Personen auf zwei Plattformen besteht: Twitter (eine öffentliche Plattform) und Discord (ein privater Server). Die Pseudonymisierung von Namen reicht nicht aus, um die Nutzer:innen zu schützen, da ihre Identität durch die öffentlichen Tweets angreifbar ist. In dem Beitrag werden diese und ähnliche ethische Fragen sowie mögliche Lösungen vorgestellt, bei denen die Achtung der Privatsphäre der Teilnehmer:innen mit der Wahrung der akademischen Integrität in Einklang gebracht werden muss.

Schlagwörter: Digitally-mediated communication, ethics, affordances

* Jenia Yudytska, Department of Languages, Literature and Media (SLM I),
Universität Hamburg, yevgeniya.yudytska@uni-hamburg.de

1 Introduction

The last four decades have seen both huge growth and rapid developments of digitally-mediated communication (DMC): from the beginning of personal computing and access of the internet to the general public in the 1980s and 90s, to the rise of mobile computing in the late 2000s, and to our current daily life full of various digital devices and increasing digital convergence (Bröhl et al. 2018; Jenkins 2008; Kjeldskov 2013; MacKenzie 2013). DMC research too has changed significantly over time, both in terms of research interests and methodological approaches. While the first wave was primarily interested in simply cataloguing “characteristic” features of DMC like emoticons and acronyms (<LOL>), the state of the art is far more complex, with increasing focus on how such resources are used for social and interactional purposes (Androutsopoulos 2006). This has been accompanied on the one hand by the deployment of sophisticated online ethnographic methods (ibid.; Bolander & Locher 2014), and on the other, by the compilation and analysis of increasingly large corpora, as technological improvements have made it easier for linguists to download messages from a variety of platforms (cf. Nguyen et al. 2015). The speed of this progress, particularly in the ability to collect more data for study from individuals, makes it prudent to continue to re-examine the ethical dimensions involved in DMC research (cf. Tagg & Spilioti 2022).

The current paper presents a reflexive account of some ethical considerations regarding the analysis of a corpus of DMC messages, compiled as part of my dissertation project on the influence of the communication device on microlinguistic features. What distinguishes this particular corpus is that it comprises both public and private messages from the same participants. Specifically, the messages are collected from one of two platforms: Twitter/X,¹ a micro-blogging

1 Note that among other changes to this platform, it has recently been renamed from “Twitter” to “X.” I continue to use “Twitter” throughout this paper, both as that was its name during data collection and to enable easier searching for readers interested in this platform specifically.

platform with a maximally large audience possible (cf. boyd & Crawford 2012; Pavalanathan & Eistenstein 2015), and Discord, a platform on which private chat servers can be created by communities such as the one the participants are part of (cf. Kiene et al. 2019). That is, while any person with an internet connection can read the Twitter messages used in the corpus, only fellow server members have access to the messages taken from Discord.

The data provides a rare opportunity to examine messages from the same users in such different dimensions of context. Most studies of language use in DMC analyse a corpus comprising of messages from a single platform or mode, whether it be a *public* platform such as Twitter (cf. Ilbury 2020; Pavalanathan & Eisenstein 2015; Shoemark et al. 2017) or online forums (cf. Androutsopoulos 2023; Bieswanger 2016) or *private* chat apps like WhatsApp (cf. Busch 2021; Siebenhaar 2020) or SMS (Thurlow & Brown 2003). While some studies do use messages from multiple modes, the corpora structure differs here as well. For example, Verheijen (2018) compares linguistic variation across four modes, Twitter, WhatsApp, SMS, and instant messaging; all but Twitter are private communication. However, each subcorpus corresponding to a platform has been collected from a different group of donors. In contrast, Tagliamonte (2016) examines messages from same users across three different modes, email, SMS, and instant messaging, but all three are private.

In short, when compared with other DMC corpora, the current data is relatively unique in respect to the private-public factor. However, it also provides some complex methodological challenges when keeping with the Association of Internet Researchers' original main ethical guideline: do no harm (Ess & the Association of Internet Researchers 2002). To explore the ethical dimension thoroughly, the paper begins with a brief overview of the project that the corpus was compiled for, as well as introducing the corpus and participants (Section 2). The topic of private-public data is also explored in more depth, following the approach of Landert and Jucker (2011) of differentiating between access and content (Section 3). With this, it is possible to explore the key issue of how to respect the participants' privacy (Section 4.1), and the possible

solutions to do so (Section 4.2). The paper concludes with some brief thoughts on the importance of reflecting on and openly discussing ethical issues related to our research (Section 5).

2 Overview of the project and corpus

The aim of my doctoral project is to explore the influence of the communication device on linguistic variation in DMC; the “communication device” here refers to the physical technology used to produce and send messages, most commonly a phone or computer (cf. Jucker & Dürscheid 2012). This topic was explored to some extent in the earliest wave of linguistic research on DMC. Overall, the characteristic features of DMC (Section 1) were analysed as developing as a result of the communication being mediated by technology. As typing is slower than speaking, the principle of parsimony and linguistic economy are especially important in DMC, which leads people to use abbreviations, omit punctuation and capitalisation, etc. (cf. Androutsopoulos 2011; Crystal 2004; Thurlow 2001; Werry 1996). Furthermore, as paralinguistic cues used in face-to-face communication, e.g., laughter, body movements, tone, are unavailable in DMC, new text-specific contextualisation cues were developed, such as the repetition of letters and punctuation (<good morninggg!!!!>), non-standard capitalisation (<GOOD MORNING!>), and emoticons (Carter 2003; Ferrara et al. 1991; Herring 2001). Comparing linguistic variation across the device types, the consensus was that the phone’s smaller keyboard and screen led to even greater linguistic economy on the phone (Cougnon & Farin 2012; Frehner 2008; Herring 2004; Herring & Zelenkauskaite 2008; Ling & Baron 2007).

These explanations were eventually criticised as overly technologically deterministic: they described technology as having an inevitable, autonomous effect on language use, while ignoring or minimising the role of social factors and users’ agency (Squires 2010). Current research has thus adopted the concept of affordances to describe the influence of technology on human behaviour more generally, and language variation

specifically. Affordances are action possibilities; they are based in the material properties of a technology and shape what is easier or harder to accomplish, without ultimately constraining it (Bucher & Helmond 2018; Hutchby 2001). Furthermore, as Section 1 notes, the focus has shifted from technology to exploring the use of linguistic features across different contexts, by different groups of people, and for different interactional purposes within DMC (Androutsopoulos 2006; Bolander & Locher 2014; Squires 2010). However, this shift has meant that there is little current systematic research on linguistic variation across the computer and phone. My dissertation seeks to fill this research gap without returning to technological determinism. Rather, I examine the affordances of a device type as one influence among many on language use.

For the empirical study, I decided on a mixed-methods approach: both qualitative but especially quantitative methods are used to investigate linguistic variation across device types and other dimensions of context. In particular, the project re-examines earlier claims about the effect of device type on microlinguistic features more robustly. For example, I compare the statistical frequencies of non-standard capitalisation across device types, but then also qualitatively compare the *motivation* for non-standard capitalisation in individual messages on the computer and phone.

Part of the project thus involves constructing a novel corpus that can be used for such analyses – one which avoids the Observer’s Paradox (cf. Bolander & Locher 2014). This means the social media platform(s) from which messages are collected must somehow display the device type used to write the message with within its metadata, which narrows down the choice of platform to only several possibilities. For example, the popular messaging service WhatsApp can be accessed both on the phone via an app and on the computer via the “WhatsApp Web” site. However, WhatsApp does not display what device type the interlocutor is using as part of its user interface. At the time of data collection, two sites that did were Twitter and Discord; note that Twitter stopped doing

so soon after it was acquired by its new CEO.² Both platforms were chosen rather than only one in order to explore the interaction between the influence of device affordances and other contextual factors thoroughly.

A small group of users who post on both platforms were approached regarding the project. These individuals are members of the book community: specifically, they engage in online fandom of sci-fi and fantasy books, either as book bloggers or as authors themselves. Book bloggers review books online and thus promote them via electronic word-of-mouth (cf. Kelly-Holmes 2016; Murray 2016). Promotion via such (micro-)influencers has become an increasingly important part of the marketing branch of the book community; they are typically not paid, although they may receive free ARCs (“advance review/reader copies”) of the book from the publisher or author (Jaakkola 2022; Moody 2019; Steiner 2010). Instead, book blogging is both a hobby, part of their online fandom engagement (Kutzner et al. 2019), but also a way to earn symbolic capital within their community, building an online identity as a trusted expert and micro-celebrity (ibid.; Albrecht 2017; Moody 2019; Reddan 2022; cf. Khamis et al. 2016).

The eleven users whose messages comprise the corpus are members of a Discord chat server of a few dozen book bloggers and authors. The server thus provides a private space for the users to chat privately about books, book blogging, and (events within) the broader fandom community; it is also used by the members to chat about other topics such as their private life, other forms of media, politics, etc. In contrast, the public platform Twitter is used primarily to promote books, and their own blogs, to a greater audience of fans. These differences are illustrated by the examples below. In Example (1), a short Discord conversation, the users are discussing their opinion on a book they both moderately enjoyed; the extract is clearly an informal conversation between friends





2 Musk (2022): “And we will finally stop adding what device a tweet was written on (waste of screen space & compute) below every tweet. Literally no one even knows why we did that ...” Retrieved from:
<https://twitter.com/elonmusk/status/1592178009410531330> [Accessed 26.06.2024]

who are both aware of each other's past reading. Example (2) is a fairly typical tweet within the corpus: an update promoting the user's new blogpost by listing several books. In short, the platforms are used for very different purposes by the participants, and the users have a different audience in mind when writing the message. The corpus consists of roughly 25,000 messages per platform, gathered sporadically over the course of a year.

Ex. 1:

[Nora 44051 Computer Discord]	oh, Leila, i finished reading witchmark
[Nora 44052 Computer Discord]	i see what you mean. it was good, but not great
[Leila 44053 Phone Discord]	A bit rushed at the end right?
[Leila 44054 Phone Discord]	Yeah

Ex. 2:

[Tereza 9211 Computer Twitter]	Final batch of mini-reviews and I am caught up!
	 The Hod King
	 The Lady's Guide to Petticoats and Piracy
	 Cursed Bunny
	 The Emperor's Babe
	[URL LINK TO BLOG]

3 Privacy and publicness in DMC

As Examples (1) and (2) show, there is a difference between the platforms Discord and Twitter on several levels in terms of the degree of publicness and privacy. At the basic level, the platforms differ as to who has access to the content of the messages. This distinction has been long-standing in DMC research: in her classification scheme for DMC, Herring (2007) differentiates between public, semi-private, and private communication. Public messages are searchable (cf. boyd 2010); that is, the message can be found again if the reader searches for the text within

the platform. Such data is often used by researchers studying DMC without asking for the user's consent, e.g., in large-scale Twitter studies (cf. Nguyen et al. 2015; Pavalanathan & Eisenstein 2015; Shoemark et al. 2017). However, boyd & Crawford (2012: 672) note, "[j]ust because content is publicly accessible does not mean that it was meant to be consumed by just anyone," and warn that accessibility should not be used to justify the ethics of collecting data without consent.

For this small-scale corpus, I asked the users for permission to collect both their Twitter and Discord data. At the time of data collection, Twitter was a maximally public platform: typically, anyone with internet access could read any message posted, although users did have the option to lock their account to be read by followers only, and one-to-one private messaging also existed. Since then, Twitter has changed its privacy rules, and now an account is necessary to access tweets; while accounts are free and simple to create, this does now technically make the platform semi-private. The platform Discord has the option of creating public, semi-private, or private servers; public servers are searchable via the platform's server discovery page, while private servers require an invite link. Some invite links may be posted publicly: for example, communities on the public, asynchronous platform Reddit may add a Discord community server for faster-paced chatting (cf. Kiene et al. 2019). The particular server examined here does not have an open invite link posted anywhere, and is thus fully private.

However, Landert & Jucker (2011) argue that accessibility is not the only dimension along which the public-private distinction must be analysed. Another important axis is the topic or content of the messages, which they describe as follows: "Private topics are those that affect single individuals or very small groups of people while public topics are those that lack this concentration on a private individual or a very small group" (Landert & Jucker 2011: 1427). Private topics are more likely to involve sensitive and personal information, while public topics include, for example, scientific facts or international sporting events. Private topics are more likely to be discussed within privately accessible communities, and vice versa. However, there can be a certain amount of blurring of boundaries, and the differentiation along the axes should be

considered a continuum rather than absolute categories (ibid.; Bolander & Locher 2014; Tagg & Spilioti 2022).

Most of the community members also explicitly describe a difference between what they are willing to discuss on Twitter and Discord. As part of the project, I conducted a questionnaire with them after data collection was complete. While it focused on their device habits and ideologies, I also asked about how they perceived the two platforms. Many of their answers centre around the public-private distinction in topic:

- *“I’m awkward as hell on twitter because I’m very conscious that anyone can see what I write there. It also feels more formal, which I’m less comfortable with.”* (Michael)
- *“I find twitter is more shouting into the void and discord is for conversations with friends. I am always aware on twitter that people I don’t know will be reading what I put out there, and while I’m fairly unfussed about what I share, there is a line between public and private information.”* (Eliza)
- *“I’m way more down to earth on discord. I’ll usually proofread my tweets a bunch, vs discord which is just... type and go!”* (Nora)
- *“Hmm I’d say I’m less guarded on Discord. If only because I know I’m among a set group of people, and nobody I don’t know is going to jump on something I say, or take it out of context based on a misreading.”* (Roy)
- *“I think each platform has its differences. Public vs private is a big one, I always take more care with what I’m saying on twitter. [...] I’m also pretty shy so much more likely to just like and retweet rather than answer, but discords are safer spaces when I can be my true awkward self. I avoid commenting on controversial subjects on twitter because I don’t have the energy for that.”* (Tereza)

The users make clear that they are more careful about what they write on Twitter, especially in regard to potentially controversial topics and sensitive information. In contrast, they treat Discord like a safe place to interact with friends, and thus are less careful about their interactions. This division can be seen in the topics discussed within the corpus, with

a much larger proportion of Discord messages concerned with the everyday. Furthermore, even when discussing media, they differ in how they express their opinions between the platforms. As illustrated in the examples below, they are more likely to express a strong negative opinion in the private Discord (<holy shit do I hate> in Example 3), while hedging negative evaluations on the public Twitter (<Unfortunately, it wasn't my cup of tea!> in Example 4).

Ex. 3:

[Leila 44032 Phone Discord]	I...
[Leila 44033 Phone Discord]	I mean I did find something but holy shit do I hate these soap opera romances
[Leila 44034 Phone Discord]	It's soooo over the top angsty and dramatic

Ex. 4:

[Leila 22803 Phone Twitter]	Unfortunately, it wasn't my cup of tea! Hopefully you'll get to watch it soon!!
-----------------------------------	---

One final aspect to note here is the association, although also not absolute, with the difference in standard language use across each of the platforms: language in messages directed at a larger, public audience has been found to be more likely to adhere to orthographic norms (Pavalanathan & Eisenstein 2015; Shoemark 2017; cf. Landert & Jucker 2011). In Examples (2) and (4), written on Twitter, the participants use standard capitalisation; Leila even uses a comma in Msg. 22803. In contrast, Examples (1) and (3) on Discord contain all-lowercase messages (<i see what you mean> in Msg. 44052) and tokens with letter repetition (<soooo> in Msg. 44034). The analysis within the dissertation finds that these examples reflect a broader statistical trend regarding linguistic variation across the two platforms, as does Nora's comment on proofreading her tweets in a way she does not with Discord messages.

Altogether then, it can be concluded that for the participants there is a very clear contrast in public-private between the two platforms. First and foremost, there is a technological difference regarding accessibility. However, this difference is also reflected in the topics the users choose to discuss on each platform, in their explicit metalinguistic understanding of the platforms, and in the style of language they use on each platform. Furthermore, maintaining the privacy of the Discord messages means not only making sure the users are not somehow identified in real life, but also, and to some extent more importantly, that their privately shared opinions do not become public among their broader online community.

4 The ethics of a private-public corpus

With the importance of the division between publicness and privacy in DMC thus established, this section turns to discussing the ethics pertaining to an empirical analysis with public-private data. The first half (Section 4.1) introduces the ethical issues which may arise: searchability of public messages, possibility of participant identification via researcher, and danger of participant tracking due to the large quantity of messages in the corpus. The second half (Section 4.2) discusses some potential solutions: avoiding certain types of analysis, reproducing only public or private messages, substituting participants' public messages for unrelated others', altering reproduced messages so they become unsearchable, avoiding reproducing certain private messages, but also potentially *heightening* risk to meet participants' desires for culture sensitivity.

4.1 Ethical issues

Discussing how to do DMC research ethically, Tagg and Spilioti (2022: 96) describe a general guideline: "the more public the site and the more open the access to it, the less urgent is the need to protect participants' privacy." The strict division between the public and private platforms

for these participants thus indicates that the privacy of their Discord messages must be handled with utmost care. However, a crucial issue arises here due to the searchability of the public tweets (cf. boyd 2010). Pseudonymisation, that is, changing the names/nicknames, is the most important and basic way to protect the anonymity of participants posting privately: using a pseudonym when discussing a user and reproducing their messages within a paper or dissertation prevents their identity from being discovered (cf. Bolander & Locher 2014; Buchanan 2011; Tagg & Spilioti 2022). However, in this case it is not enough. Anyone searching for the text of the Twitter message itself would be able to find it easily; they would thus immediately know who the pseudonyms “Leila” and “Tereza” actually belong to.

A further issue is my own involvement with this community: the reason that I have access to the private server and these participants is that I also do book blogging. While I have been “on hiatus” since starting my dissertation (and never did any message collection around the time periods I was active in the community spaces so as to minimise any accidental influence from my own device use), I am still on close terms with some of the participants and other members within the book community. While I am not active in those fandom spaces under my full name, that is, I do not blog as “Jenia Yudytska,” neither am I very careful about hiding my identity; my Twitter profile, for example, identifies me as both a book blogger and a linguistics doctoral student. The problem here is that as I myself am findable, so too is my broader network, and thus potentially the participants.

Robson (2017) describes just such a problem with regard to his own role as digital ethnography researcher when he conducted a long-term study of a public forum used by Religious Education teachers. While the access to the forum posts is public, the topics discussed by its members can be relatively sensitive, and thus are relatively private (cf. Landert & Jucker 2011). Therefore, when reporting his findings, he never published direct quotes from the participants, using paraphrases instead. However, he had interacted with the participants on the forum publicly in his role as researcher, under his real name. Thus, googling for him and the forum topic meant that the participants could be found, and

subsequently identified based on the paraphrases. His solution was to simply delete his messages on the forum and thus sever the connection. However, this is impossible in my case, as it would be extremely difficult to remove all traces of my involvement in the book community across multiple platforms: even if I were to remove my own messages, I am also on occasion mentioned in others' messages and blogposts.

Finally, there is a more general problem with the quantity of metadata within the corpus. As mentioned in Section 2, the corpus is quite large at ca. 50,000 messages; that is, the more prolific of the participants have contributed two to five thousand messages on each platform. The participants were informed that I would be collecting messages over the course of a year, and were also informed about the topic of my dissertation, that is, that I would specifically be tracking their use of computer and phone. However, as researchers on internet ethics have pointed out, users are not always aware of how much information is being collected in aggregate, which complicates the notion of "informed consent" (boyd & Crawford 2012; Buchanan 2011; Tagg & Spilioti 2022).

In particular, one of my original research interests was to explore the motivation for and potential impact on linguistic variation of device switching: participants switching from the computer to the phone or vice versa during a conversation. For example, several participants indicated within the questionnaire that they may switch to the computer when typing longer messages. As part of my preliminary attempts at investigating such device switching, I used the `ggplot2` R package (R Core Team 2022; Wickham 2016) to create Figure 1, based on the timestamps of the participants' messages. Each row in Figure 1 is one participant, and each dot is a message that they wrote; the graph shows a timespan over the course of several days.

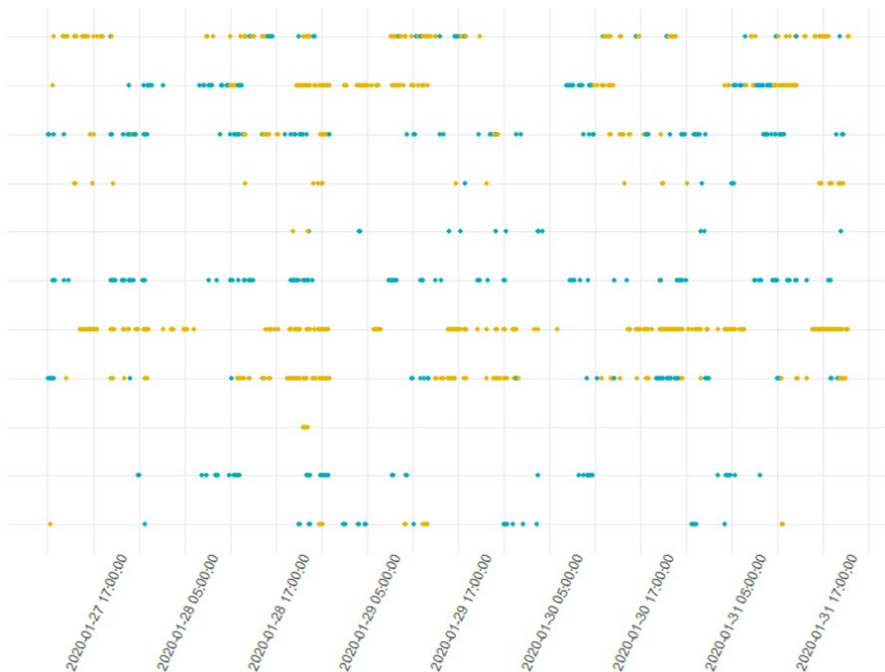


Figure 1: Messages posted by each participant over the course of several days; blue messages have been posted via the phone and yellow via the computer

The graph in Figure 1 was meant to provide a rough overview of device switching, so that timespans with more switching could be analysed qualitatively in more detail. During this period, the community was very active and the participants were posting constantly throughout the day. Thus, what I had inadvertently created was a graph to track the users', who are all in different time zones, sleeping patterns. In his blog article, "How you can use Facebook to track your friends' sleeping habits," the software engineer Louv-Jansen describes producing a similar graph (Louv-Jansen 2016). He had realised that his friends checked Facebook upon waking up and right before bed; thus, with a simple script to record their online status, he could track their sleeping patterns with a high degree of accuracy. With the data in my corpus, it is possible to go

one step further: some of the participants describe using their phone primarily in a mobile context of use, that is, when they are away from the computer. Examining device switching in this way thus means tracking their movements, albeit with a very basic categorisation of stationary vs. mobile.

4.2 Potential solutions

I believe that the most important and basic part to solving an ethical dilemma within the research is to weigh whether the analysis at hand is worth doing at all. In my project, there are two major issues, the issue of a potential discovery of private data via public data, and the issue of unintentional tracking. These issues were resolved in different ways. For the unintentional tracking, I simply decided to abandon the investigation into device switching within the dissertation. While I am still interested in this topic, an additional round of consent gathering would be required to ensure the participants are truly informed (cf. Tagg & Spilioti 2022), and device switching may be better analysed via screen recording instead in any case. In regard to the principal guideline of “do no harm” (Ess & the Association of Internet Researchers 2002), the main problem would be if participants were somehow recognised and their constant online communication throughout the day would cause difficulties at their workplace. In short, the potential risks and obstacles severely outweigh the potentially limited research benefits; in fact, I decided to avoid using timestamps altogether when showing message examples, as another way to limit tracking.

Similarly, it would be possible to resolve the issue of the public-private data by taking a purely quantitative approach to the analysis. Graphs and statistical models can be used to investigate trends of linguistic variation across device types and platforms, without any messages being shown. A related potential solution would be to reproduce only private or public messages. While I do primarily take a quantitative approach in the dissertation, I rejected the idea of not showing any messages. Importantly, a reproduction of the messages

helps to clarify the analytic argument being made. For example, the quantitative analysis finds a significant difference between messages produced via the computer and phone in their use of capitalisation, with more omission on the computer than the phone, due to the latter's auto-capitalisation. Nora's Discord message in Example (1) (< i see what you mean. it was good, but not great>) helps explain this finding more clearly than only a graph would.

Furthermore, while not the primary focus, I *did* compare phone-based and computer-based messages qualitatively as an important part of the analysis. Messages should thus be shown so as to support the reproducibility of the findings (cf. Weller & Kinder-Kurlanda 2016; Winter 2020). Winter (2020) describes *replicability* of research as the ability to reproduce the findings of a study on *novel* data, and *reproducibility* of research as the (more basic) ability of another researcher to reproduce the findings of a study given the *same* data. The minimal requirement for reproducible research is thus that the data is made available in some way, so that other researchers can come to their own conclusions, thus disagreeing with or reproducing my analysis. Due to the mixed public-private nature of the corpus as described in this paper, the full corpus cannot be shared openly, but the bare minimum is to show certain pertinent messages for others to examine.

One potential solution I considered but also ultimately discarded was to show Discord messages from the participants and the corpus, but to show Twitter messages from unrelated users within the broader book community instead. As tweets are public, informed consent is arguably less important (cf. Spilioti & Tagg 2022). Many others in the community use similar linguistic strategies to promote books. For example, the tweet below (Figure 2) is from a publishing company, that is, from a public company posting on a public platform. Like the tweet in Example (2), it uses emoji as bullet points, and thus could technically be used in its place to illustrate this stylistic choice within the book community. However, this approach was rejected for two reasons. On the one hand, boyd and Crawford's (2012) warning that accessibility should not be taken as justification is valid, especially when considering that I would be using data no one gave me consent for to protect the corpus data that

I *did* have consent for; on the other, finding illustrative, unrelated examples is extremely time-consuming and difficult.

What fanfiction teaches writers:

- ✍️ to be humble: you're bending characters to your will, but within someone else's context
- ✍️ structure
- ✍️ to be ENTERTAINING and keep people coming back for the next chapter
- ✍️ characterization, diving into the interiority of characters' heads





Figure 2: Tweet from a publishing company, not part of the corpus

Ultimately, I took two main precautions in order to protect the users' identity and privacy. The first was to anonymise the public tweets even further. Names and any locations were all pseudonymised. Moreover, every tweet was altered slightly when reproduced, so that it would become impossible, or at least far more difficult, to search for. That is, every tweet within this paper has been changed slightly. Example (5) below illustrates this procedure: the tweet from Example (1) has been changed one more time. This involved changing all book titles, and sometimes changing emoji and adjectives or nouns to their synonyms. As the focus is on microlinguistic features, the exact book or adjective used is deemed less important than the overall structure, and graphic features, of the tweet. For example, the original book title Tereza mentions is neither <The Emperor's Babe> nor <Assassin's Apprentice>, but it does use both standard capitalisation and an apostrophe. To check that this step of the anonymisation worked, I tried searching for sections of each altered message on Twitter's built-in search engine; if the tweet still appeared in the search results, I altered the message further and rechecked it until this was no longer the case.

While this means the data does not *completely* fulfil the criteria for reproducibility described above, analysing the private-public divide for the participants in depth led me to conclude that it is more academically sound to protect my participants' privacy than to reproduce the examples one-to-one. Each message is shown with an ID, however (e.g., Msg. 9211 in Example 5). This allows the original to be found easily within the corpus, so if absolutely necessary to answer any questions, it could potentially be briefly shown to specific individuals.





Ex. 5:

[Tereza | 9211 | Computer | Twitter] Final batch of mini-reviews and I am caught up!

-  The Hod King
-  The Lady's Guide to Petticoats and Piracy
-  Cursed Bunny
-  The Emperor's Babe

[URL LINK TO BLOG]

[Tereza | 9211 | Computer | Twitter] Last batch of mini-reviews and I have caught up!

-  The Constant Rabbit
-  The Gentleman's Guide to Vice and Virtue
-  Prometheus Bound
-  Assassin's Apprentice

[URL LINK TO BLOG]

The Discord messages I left unaltered, other than changing book titles if the participants were discussing a negative review, or if the Discord message was somehow linked to a reproduced Twitter message. This is part of the other precaution taken, which is to simply avoid or minimise reproducing messages from private topics within the private data (cf. Buchanan, 2011; Landert & Jucker 2011). As described in Section 4.1, even if the Twitter data is altered enough to become untraceable, I as the researcher and the community member am still a vulnerable point of

access to the participants' identity. Moreover, those most likely to recognise the participants through me are fellow members of the wider book community – and it is among them that the reputation of the participants could be damaged if private opinions became known. Again, the focus of the dissertation is various microlinguistic features, and not larger discourses; it serves no scientific purpose to use the most controversial, sensitive, or otherwise private material from the Discord messages. Therefore, it should be and is avoided. With this, even if the participants are found, the risk of harm to them should be minimised even further.

One final point, however, concerns heightening risk rather than lowering it. Two of my participants are from a first/second-generation immigrant background, now living in Western Europe. I had originally planned to pseudonymise them using a name traditional to the country they currently live in, as there is an increased risk of identification with using a name from their home culture. Firstly, there are overall comparably fewer sci-fi and fantasy book bloggers of their cultures in the (English-speaking) online book community: the pool of potential “suspects” that these pseudonymised users could be thus becomes far smaller than if they are given stereotypically white (Anglo) names. Secondly and more crucially, the participants can be found through their connection to *me*, and as my own network of book bloggers is overall not exceedingly large, the pool of “suspects” from their cultures now becomes limited to a few persons.

Nevertheless, past guidelines point out the need for cultural sensitivity when conducting research; paradoxically, marginalised users may at times desire greater visibility (franzke et al. 2020; Tagg & Spilioti 2022). When taking part in a friend's research project, I had had my own experience of an Anglo pseudonym being chosen for me, and feeling oddly uncomfortable at seeing a quote from me published under an Anglo name. Consequently, I asked each user directly what they would prefer, after explaining my thoughts about potential risk of identification. Despite the warning, both wanted a name from their home culture, and even supplied me with a suitable pseudonym themselves. Because the participants expressed their preferences so clearly, and also

because I have done my best to mitigate risk in other ways, such as avoiding reproducing sensitive messages, I decided that the increased risk of identification was outweighed by the need to respect the participants' cultural identity. This example underlines the importance of, where possible, working with the participants to ensure that they are protected ethically in a way that matches their preferences.

5 Conclusion

The aim of this paper was to provide some ideas about potential issues and potential solutions for researchers interested in working with DMC data. While all (empirical) linguists face ethical dilemmas throughout our research, it is rare for us to have the opportunity to discuss the deliberations behind our choices in-depth. In particular, there is little public space for us to admit to *not* undertaking analysis specifically out of ethical considerations. Thus, for me, the decision to reject investigating device switching via timestamps was accompanied by the strong worry that I was over-thinking the issue, abandoning a promising novel direction of research over nitpicky moral qualms. In addition to offering some concrete potential ideas on how to tackle ethical issues in DMC research, more broadly, I hope that this paper is useful to other young researchers as a transparent illustration of the thought-process behind the choices taken and *not* taken in such studies.

Nevertheless, it is important to stress that the decisions taken here are not necessarily right for all studies. Most importantly, my study is primarily quantitative in nature, and the reproduced text messages are just one part of the analysis. If my dissertation were to focus on close reading or other qualitative methodologies, it would arguably be far more important scientifically to reproduce the message accurately; hence, another approach to dealing with the data ethically would have to be chosen. As discussed already in the Association of Internet Researchers' original recommendations from 2002, a "recipe" for ethical research of DMC is impossible, but that does not mean there are no guidelines or responsibilities for researchers either (Ess & the

Association of Internet Researchers 2002). Rather, a sometimes-complicated series of choices is involved in ensuring that the best possible measures are taken to ensure the participants' privacy.

Acknowledgements

I would like to express my warm appreciation to the participants of the reading workshop, for the productive and insightful discussions.

References

- Albrecht, Katharina. 2017. Positioning BookTube in the publishing world: An examination of online book reviewing through the field theory. Leiden University: MA thesis.
- Androutopoulos, Jannis. 2006. Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics* 10(4). 419–438.
- Androutopoulos, Jannis. 2011. Language change and digital media: A review of concepts and evidence. In Nikolas Coupland & Tore Kristiansen (eds.), *Language standardisation in Europe*, 145–159. Oslo: Novus Press.
- Androutopoulos, Jannis. 2023. Punctuating the other: Graphic cues, voice, and positioning in digital discourse. *Language and Communication* 88. 141–152.
- Bieswanger, Markus. 2016. Electronically-mediated Englishes: Synchronicity revisited. In Lauren Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*, 281–300. Berlin & Boston: de Gruyter.
- Bolander, Brook & Miriam Locher. 2014. Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media* 3. 14–26.
- boyd, danah. 2010. Social Network Sites as networked publics: Affordances, dynamics, and implications. In Zizi Papacharissi (ed.), *Networked self: Identity, community, and culture on Social Network Sites*, 39–58. New York: Taylor & Francis.
- boyd, danah & Kate Crawford. 2012. Critical questions for Big Data. *Information, Communication & Society* 15(5). 662–679.

- Bröhl, Christina, Peter Rasche, Janina Jablonski, Sabine Theis, Matthias Wile & Alexander Mertens. 2018. Dekstop PC, tablet PC, or smartphone? An analysis of use preferences in daily activities for different technology generations of a worldwide sample. In Jia Zhou & Gavriel Salvendy (eds.), *Human aspects of IT for the aged population. Acceptance, communication and participation. ITAP 2018. Lecture notes in computer science* 10926. 3–20.
- Buchanan, Elizabeth A. 2011. Internet research ethics: Past, present, and future. In Mia Consalvo & Charles Ess (eds.), *The handbook of internet studies*, 83–108. Chichester, UK: Blackwell Publishing.
- Bucher, Taina & Anne Helmond. 2018. The affordances of social media platforms. In Jean Burgess, Alice Marwick & Thomas Poell (eds.), *The Sage handbook of social media*, 233–253. Sage.
- Busch, Florian. 2021. *Digitale Schreibregister: Kontexte, Formen und metapragmatische Reflexionen*. Berlin & Boston: de Gruyter.
- Carter, Kimberley A. 2003. Type me how you feel: Quasi-nonverbal cues in computer-mediated communication. *ETC: A Review of General Semantics* 60(1). 29–39.
- Cougnon, Louise-Amélie & Cédric Fairon. 2012. Introduction. In Louise-Amélie Cougnon & Cédric Fairon (eds.), *SMS communication: A linguistic approach*, 3–10. Amsterdam: John Benjamins.
- Crystal, David. 2004. *Language and the internet*, 2nd edn. Cambridge: Cambridge University Press.
- Ess, Charles & the AoIR ethics working committee. 2002. *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee*. www.aoir.org/reports/ethics.pdf
- Faraj, Samer & Bijan Azad. 2012. The materiality of technology: An affordance perspective. In Paul M. Leonardi, Bonnie A. Nardi & Jannis Kallinikos (eds.), *Materiality and organizing: Social interaction in a technological world*, 237–258. Oxford: Oxford University Press.
- Ferrara, Kathleen, Hans Brunner & Greg Whittemore. 1991. Interactive written discourse as an emergent register. *Written Communication* 8(1). 8–34.
- franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess & the Association of Internet Researchers. 2020. *Internet research: Ethical guidelines 3.0*. Available at: <https://aoir.org/reports/ethics3.pdf>
- Frehner, Carmen. 2008. *Email – SMS – MMS. The linguistic creativity of asynchronous discourse in the new media age*. Bern: Peter Lang.

- Herring, Susan C. 2001. Computer-mediated discourse. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The handbook of discourse analysis*, 612–634. Oxford: Blackwell.
- Herring, Susan C. 2004. Slouching towards the ordinary: Current trends in computer-mediated communication. *New Media & Society* 6(1). 26–36.
- Herring, Susan C. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet* 4. Available at: <https://www.languageatinternet.org/articles/2007/761>
- Herring, Susan C. & Asta Zelenkauskaitė. 2008. Gendered typography: Abbreviation and insertion in Italian iTV SMS. In Jason F. Siegel, Traci C. Nagle, Amandine Lorente-Lapole & Julie Auger (eds.), *IUWPL7: Gender in language: Classic questions, new contexts*, 73–92. Bloomington: IULC Publications.
- Hutchby, Ian. 2001. Technologies, texts and affordances. *Sociology* 35(2). 441–456.
- Ilbury, Christian. 2020. “Sassy queens”: Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics* 24(2). 245–264.
- Jaakkola, Maarit. 2022. *Reviewing culture online: Post-institutional cultural critique across platforms*. Gothenburg: Springer Nature Switzerland.
- Jenkins, Henry. 2008. *Convergence culture: Where old and new media collide*. New York & London: New York University Press.
- Jucker, Andreas & Christa Dürscheid. 2012. The linguistics of keyboard-to-screen communication. A new terminological framework. *Linguistik Online* 56(6). 39–64.
- Kelly-Holmes, Helen. 2016. Digital advertising. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge handbook of language and digital communication*, 212–225. New York: Routledge.
- Khamis, Susie, Lawrence Ang & Raymond Welling. 2016. Self-branding, ‘micro-celebrity’ and the rise of Social Media Influencers. *Celebrity Studies* 8(2). 191–208.
- Kiene, Charles, Jialun A. Jiang & Benjamin M. Hill. 2019. Technological frames and user innovation: Exploring technological change in community moderation teams. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 44 (1–23).
- Kjeldskov, Jesper. 2013. Mobile computing. In Mads Soegaard & Rikke F. Dam (eds.), *The encyclopedia of human-computer interaction*, 2nd edn. The Interaction Design Foundation. Available at:

- http://www.interactiondesign.org/encyclopedia/mobile_computing.html
- Kutzner, Kristin, Kristina Petzold & Ralf Knackstedt. 2019. Characterising social reading platforms. A taxonomy-based approach to structure the field. In *Proceedings of the 14th international conference on Wirtschaftsinformatik*. AIS eLibrary.
- Landert, Daniela & Andreas H. Jucker. 2011. Private and public in mass media communication: From letters to the editor to online commentaries. *Journal of Pragmatics* 43. 1422–1434.
- Ling, Rich & Naomi S. Baron. 2007. Text messaging and IM: Linguistic comparison of American college data. *Journal of Language and Social Psychology* 26(3). 291–298.
- Louv-Jansen, Søren. 2016. How you can use Facebook to track your friends' sleeping habits. *Medium*. Available at: <https://medium.com/@sorenlouv/how-you-can-use-facebook-to-track-your-friends-sleeping-habits-505ace7fff6> [Accessed 26.06.2024]
- MacKenzie, I. Scott. 2013. *Human-computer interaction: An empirical research perspective*. Amsterdam: Elsevier.
- Moody, Stephanie. 2019. Bullies and blackouts: examining the participatory culture of online book reviewing. *Convergence* 25(5–6). 1063–76.
- Murray, Simone. 2016. 'Selling' literature: The cultivation of book buzz in the digital literary sphere. *Logos* 27(1). 11–21.
- Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé & Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42(3). 537–593.
- Pavalanathan, Umashanthi & Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech* 90(2). 187–213.
- R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Reddan, Bronwyn. 2022. Social reading cultures on BookTube, Bookstagram, and BookTok. *Synergy* 20(1). Available at: <http://slav.vic.edu.au/index.php/Synergy/article/view/597>
- Robson, James. 2017. Participant anonymity and participant observations: Situating the researcher within digital ethnography. In Michael Zimmer & Katharina Kinder-Kurlanda (eds.) *Internet research ethics for the social age: New challenges, cases, and contexts*, 195–202. New York: Peter Lang.

- Siebenhaar, Beat. 2020. Informalitätsmarkierung in der WhatsApp-Kommunikation. In Jannis Androutsopoulos & Florian Busch (eds.), *Register des Graphischen*, 133–158. Berlin: de Gruyter.
- Shoemark, Philippa, Debnil Sur, Luke Shrimpton, Iain Murray & Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. *Proc. of the 15th conference of the European chapter of the association for computational linguistics* 1. 1239–1248.
- Squires, Lauren. 2010. Enregistering internet language. *Language in Society* 39, 457–492.
- Steiner, Ann. 2010. Personal readings and public texts: Book blogs and online writing about literature. *Culture Unbound* 2. 471–494.
- Tagg, Caroline & Tereza Spilioti. 2022. Research ethics. In Camilla Vásquez (ed.), *Research methods for digital discourse analysis*, 91–114. London: Bloomsbury Academic.
- Tagliamonte, Sali A. 2016. So sick or so cool? The language of youth on the internet. *Language in Society* 45. 1–32.
- Thurlow, Crispin. 2001. Language and the internet. In Rajend Mesthrie & Ronald E. Asher (eds), *The concise encyclopedia of sociolinguistics*, 287–289. London: Pergamon.
- Thurlow, Crispin & Alex Brown. 2003. Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*. Available at: <https://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html>
- Verheijen, Lieke. 2018. Is textese a threat to traditional literacy? Dutch youths' language use in written computer-mediated communication and relations with their school writing. Nijmegen: Radboud University doctoral dissertation.
- Weller, Katrin & Katharina E. Kinder-Kurlanda. 2016. A manifesto for data sharing in social media research. *WebSci '16*. 166–172.
- Werry, Christopher. 1996. Linguistic and interactional features of Internet Relay Chat. In Susan C., Herring (ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, 47–61, Philadelphia: John Benjamins.
- Wickham, Harold. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
- Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. New York & Abingdon: Routledge.