Evi Dalmaijer*, Wyke Stommel, Berber Pas and Wilbert Spooren

Ethical challenges in collecting pre-existing digital data for linguistic research

https://doi.org/10.1515/ling-2023-0132
Received June 30, 2023; accepted July 7, 2024; published online December 10, 2024

Abstract: Pre-existing digital data are a valuable resource for linguistic research. Collecting these materials is often thought of as straightforward ("the data exist anyway") and ethical dilemmas are given little consideration. In this article, we discuss microethical issues we encountered while collecting electronic text messages, photos, and videos posted on a digital platform and app during paramedical treatment. Since ethics and methods are intertwined, we discuss the various ethical and methodological aspects of collecting these sensitive digital data for our linguistic research project and reflect on the benefits and limitations of the choices we made during this process. We specifically highlight the interdependence of ethics with technology and discuss how this can be even more challenging when working in a specific institutional context characterized by different conceptions of ethics and technology. Our article highlights the importance of microethics complementing prevalent ethical guidelines. We show that when pre-existing digital data are available in non-public digital spheres, it is difficult for researchers to define in advance in ethical protocols or guidelines how the data can be collected and what ethical measures should be taken. We argue that ethical reflections should be at the center of research, including research on pre-existing digital data, guiding the decisions to be made at all stages.

***Corresponding author: Evi Dalmaijer**, Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands; and Interdisciplinary Research Hub on Digitalization and Society, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands, E-mail: evi.dalmaijer@ru.nl. https://orcid.org/0009-0007-3099-0446

Wyke Stommel, Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands; and Interdisciplinary Research Hub on Digitalization and Society, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands, E-mail: wyke.stommel@ru.nl. https://orcid.org/0000-0003-2345-1691

Berber Pas, Interdisciplinary Research Hub on Digitalization and Society, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands; and Institute for Management Research, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands, E-mail: berber.pas@ru.nl. https://orcid.org/0000-0002-7563-904X

Wilbert Spooren, Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands, E-mail: wilbert.spooren@ru.nl. https://orcid.org/0000-0002-2982-3970

Keywords: microethics; pre-existing data; digital data collection; technology

1 Introduction

Pre-existing, multimodal digital materials, such as email logs, YouTube videos, Instagram pictures, or WhatsApp conversations, provide a rich resource for linguistic research (Herring 2004; Spilioti and Tagg 2017; Spooren and van Charldorp 2014). Since these materials already exist prior to the start of the research, the general belief may be that they can be accessed and collected fairly easily. As a result, there is often little attention paid to the ethical issues involved (Stommel and de Rijk 2021). While gathering data, linguists face numerous obstacles and must constantly make ethical and methodological decisions that can, to varying degrees, affect the course of the research (Rose et al. 2019). It is therefore important to reflect on these choices and to learn from each other's experiences (Page et al. 2022; Spilioti and Tagg 2017). However, as various types of linguistic research pose very different ethical questions, no discussion of research ethics in the use of pre-existing data for linguistic research can be exhaustive and no particular set of guidelines can foresee unique circumstances (De Costa 2015).

This article discusses microethical considerations that are made before and during the collection of pre-existing multimodal digital materials that are not publicly available.[1] We rely on the distinction between macro- and microethics (Kubanyiova 2008), with macroethics referring to protocols of ethical boards and principles established in ethical guidelines, and microethics to everyday ethical dilemmas arising from the roles and responsibilities of researchers and research participants in specific research contexts. In the specific case of research into digital materials, ethical guidelines have been established for Internet research, the best known of which being those of the Association of Internet Research (AoIR, www.aoir.org/ethics). However, due to the multifaceted nature of digital data, such guidelines leave room for interpretation (Sugiura et al. 2017). Ethical practice is essential to responsible research, but the criteria for such practice are neither universally accepted nor simple (Markham and Buchanan 2015). Ethical and legal regulations vary across academic disciplines, and regulations and laws are modified or replaced over time (Stommel and de Rijk 2021). Therefore, although a number of ethical

[1] We are aware that the term 'collecting' is also used to refer to the generation of linguistic material, but in this article we use the term 'collecting' to refer to the gathering of pre-existing material. Further, we are aware that there are also ethical challenges in other phases of the research (e.g., analysis or dissemination), but we focus the discussion on the data collection phase because (a) this was the phase that was challenging for us and where we thus encountered problems and (b) this was the only research phase completed at the time of writing.

guidelines prove to be useful in articulating good practices, such guidelines should also be supplemented with microethical practices, that is, actual examples of how researchers deal with ethical dilemmas in specific research contexts (De Costa 2015). As a response to repeated calls for ethical reflection and conversation (Markham and Buchanan 2015, 2017; Spilioti and Tagg 2017), this article will discuss some of the microethical issues one may encounter in using already existing digital research materials for linguistic research.

Ethics is an actively debated topic when it comes to Internet research, including when it involves applied linguistics and discourse analysis (e.g., Page et al. 2022; Paulus and Wise 2019; Spilioti and Tagg 2017; Stommel and de Rijk 2021). However, previous articles mainly focused on publicly available online data, while our article focuses on *private* digital materials.[2] We draw on a doctoral research project that uses multimodal digital materials (interaction on a website and via an app) produced in the context of paramedical treatment situated in the Netherlands. The goal of this project is to study the impact of digitalization on (para)medical interaction and professional practice. The research is qualitative in nature and the data collection discussed in this article is studied using a conversation analytical (CA) approach, to investigate how interaction in healthcare is impacted by technology and remoteness. The digital data consist of digital text messages between paramedical professionals and parents of children who are patients in a Dutch hospital, sent via two types of media: (1) a (secure) instant messaging app and (2) a non-publicly accessible forum-like website. In addition to text messages, also videos and photos showing the

Table 1: Details of research context.

Location	Time scale	Participants	Participant access	Nature of data	Data access
A hospital situated in The Netherlands.	September 2021 till February 2022.	(a) Parents of children treated in the hospital and (b) paramedical professionals involved in the children's treatment.	Obtained through gatekeepers, who in turn were also participants in the research.	Text messages, video clips, and pictures from two types of media: a secure instant messaging app and a non-publicly accessible forum-like website.	Obtained through participants. They sent materials to researchers.

[2] There are also some earlier exceptions to this, such as the article by Tagg et al. (2017) that looks at the ethics of digital ethnography.

children are exchanged in the app and on the website. See Table 1 for more details on the context of the research. Because our materials consist of written interactions combined with videos and photos, the exchanges "persist", meaning that participants[3] can be approached after the interaction has taken place (Boyd 2008; Tagg et al. 2017).

Considering that ethics and methods are intertwined, we discuss the various ethical and methodological aspects of our process of collecting these sensitive digital data for our linguistic research project and reflect on the merits and limitations of the decisions we made during this process. It should be noted that collecting private digital data is still a fairly new approach, particularly in the context of hospital treatment, and as such, the methods we used were exploratory. Therefore, ethical issues were difficult to predict, despite the positive assessments of various ethical committees related to the institutes we worked with (such as the EACH, http://tinyurl.com/mr4xxesk). Our discussion is not intended to be turned into a set of rules or guidelines, but as an illustration of microethics in practice. Our article is thus mainly descriptive in nature, aiming to provide insight into which ethical and methodological choices and dilemmas arise prior to and during collecting pre-existing digital data for linguistic research. As do other scholars on ethics and technology (e.g., Introna 2005), our contribution specifically highlights the interdependence of ethics with technology. Additionally, we address how this can be even more challenging in an interdisciplinary context characterized by different conceptions of ethics, technology, and research.

Although thus framed as an illustrative case study of microethics in an ethically complex research context, we raise four key points that go beyond the details of our study and have broader significance for our understanding of ethical research in general. These include the following points: (1) interdependence of ethics with technology, (2) distinction between active and passive participants, (3) the concept of the 'non-observer paradox' and (4) understanding ethics not only as a process but as a discussion.

We thematize our findings by broadly discussing four main obstacles in the collection of pre-existing digital data in a medical context, although many of the problems could arise in other contexts where personal, privacy-sensitive data are at stake. To illustrate, we describe several ethically sensitive situations we encountered while collecting materials. We thus provide a microethical case study of an ethically complex research context and consider the issues that may arise when collecting this still relatively new type of data. In doing so, we demonstrate the need for such microethical accounts to inform future research, alongside the macroethical frameworks that typically guide research. We point to a particular need for such

[3] Although we speak of 'participants', we are aware that those who shared their material with us are not participants in the sense of the word often used in experimental research.

accounts in qualitative projects, where research is iterative and data-driven, and not all ethical issues can be identified in advance.

2 Putting ethics into practice when collecting pre-existing digital data

In the research project under discussion, we were aware that digital communication tools had been used in a hospital for some time and that a large amount of interactional data (videos, pictures, and electronic text messages) had been archived for the purpose of treatment. Such data provide valuable and useful resources for applied linguists striving to understand language use and improve professional practices (Antaki 2011; Jol and Stommel 2016). Utilizing interactional data generated and stored for purposes other than the study (e.g., in institutional contexts) is also referred to as secondary or pre-existing data use (Jol and Stommel 2016).[4] An important methodological advantage of using pre-existing data is that the materials were generated without the intervention of the researchers, thus avoiding Labov's observer paradox (Jol 2020). Natural conduct can potentially be influenced by the presence of researchers (or their cameras) on stage, making their observations less reliable. Avoiding this influence as much as possible is eminently important in research interested in naturally occurring interactions, such as conversation analysis (Jol and Stommel 2016). The absence of researchers (or their cameras) also benefits the treatment itself. Since the treatment has already been completed by the time the data are collected, the research cannot affect, delay, or otherwise interfere with the treatment people are receiving, for example by causing additional anxiety or distress (Parry 2010).

Using pre-existing digital materials for linguistic research thus offers considerable methodological and ethical advantages. However, the question is how these already existing data, which were not recorded or stored for research purposes in the first place, can be used ethically. Ethical considerations are more obvious in some situations than in others. In drug research on human subjects, for example, as well as other research in which a participant is physically present, potential harm is anticipated, and the vulnerability of subjects is more readily recognized. In other situations, where a participant does not need to be physically involved in the research, ethical dilemmas are not always obvious in advance (Markham and Buchanan 2015). In research involving the Internet, ethical considerations start with

[4] We emphatically do not mean secondary data use in the sense of cases where researcher A uses data that have been collected by researcher B.

the question whether creators of online materials (in our case, [a] paramedical professionals and [b] parents of patients) should be seen as “research subjects” (in the senses common in human subject research in biomedical and social sciences) or as authors or creators (AoIR 2019). There are important differences between the two situations. Those who generated the digital materials have indeed at some point participated in digital interactions by typing, filming and posting videos, and so on, but they were not performing research-directed tasks for the purpose of the study, are not answering interview questions, and are not physically present during the data collection. Since the interactions were generated for purposes other than research, and thus existed prior to the research, studying these data is not an active intervention in the lives of the people. Therefore, in the case of pre-generated digital material, the concept of human research subjects is not a meaningful starting point for ethical research conduct (Jol and Stommel 2016). The concept of authors or creators in Internet research (AoIR 2002) does not apply, because the materials were never created with the intention that they be made public.

Therefore, in our research, we choose to use the term participants, because we want to account for the very human side of the research (data). We are interested in their products (i.e., the digital interactions in which they participated), and not in their individual behaviors. Although it later appeared that the participants inevitably played a more active role rather than a fully passive role, in our view they still remained participants. After all, they did not generate new data, but helped us make available data that had already been generated. We sought post-hoc informed consent from all participants before we collected their data. Informed consent is based on the principle that potential participants should be able to make an informed decision whether or not to participate in a study. While this is desirable, it is not always essential.[5] Our decision to ask for post-hoc informed consent from participants was based in part on the sensitivity of the materials. Furthermore, although within the field of Internet research, studies that focus on patterns of discourse are considered low risk for participants compared to studies that focus on individuals and their lives, qualitative research with “small” data is still considered high risk because of the potential traceability to individuals (Buchanan 2011; Page et al. 2022). To ensure that all participants were fully informed and would share their materials with us voluntarily, we approached everyone with an account through which text messages, photos, or videos had been sent to ask for post-hoc informed consent.

However, because our research was related to a medical setting, we believed the content of the interactions was most sensitive for those who were non-professionally

[5] See, for example, Jol and Stommel (2016), who describe how they ethically use existing material without asking for consent, as their material is from police witnesses and asking for consent is not possible due to legal reasons and the stress it causes participants.

involved in the interaction. Consent was therefore obtained in two stages: first, those not professionally involved in the digital interaction were asked for consent (in our case: parents of children treated in the digital setting); second, we sought consent from the professionals involved in the digital interaction (in our case: paramedical professionals).[6] Only in those cases in which everyone who had participated in the interaction had given their permission were the materials collected.

There may be practical difficulties in obtaining informed consent from all participants in an online environment. In our study, a dilemma arose when we considered how to approach parents to ask for their consent. There are several ways to get in touch with potential participants, but in our case, using a “gatekeeper” was the most ethical and practical way. Gatekeepers can act as intermediaries between researchers and participants (de Laine 2000). Because of privacy regulations, it is not possible as a researcher to obtain contact information from patients in a hospital unless there is a treatment relationship between the patient and the person approaching the patient for participation in the study. Professionals at the hospital therefore asked the parents of their patients whether they were interested in participating in our study. In our case, gatekeepers were thus initially hospital employees who had a treatment relationship with the child in question. Because our gatekeepers were insiders with practical knowledge and experience, they were more likely to be trusted and respected by potential participants. It is possible that acquaintance with the gatekeeper encouraged them to participate in the study (McDermid et al. 2014). In a pragmatic sense, using gatekeepers also allowed easy and quick access to potential participants (Greene 2014). The gatekeepers knew the potential participants and were able to approach them in an appropriate way. Moreover, participants’ identities did not have to be disclosed to the researchers until they had agreed to be contacted, which is in accordance with the General Data Protection Regulation (GDPR), i.e., the legal context for this study. Also, the use of a gatekeeper ensures that participants’ identities were verified, which allowed us to be sure that they were who their email address said they were by the time we received that information from the gatekeeper. Confirming participant identity is very important in the case of research on digital materials (Henderson et al. 2012). After all, as a researcher, you need to be sure that the person giving permission to use the materials is the same person who produced the materials.

However, there are also some risks associated with a gatekeeper strategy. The most important of these is to ensure that participants do not feel coerced into participating in the study (Eckert 2014). This risk may even be more pronounced in a

[6] This means that initially professionals acted as gatekeepers regarding which parents were asked for consent, and subsequently these parents acted as gatekeepers regarding which additional professionals were asked for consent.

healthcare setting, when participants feel that choosing whether or not to participate in the study may affect their relationship with their healthcare provider (Henderson et al. 2012). Because our study involved collecting pre-existing materials, there was minimal risk that this affected the care a person received. Still, the relationship with a caregiver can be a sensitive issue. In our case, for example, one participant emailed that they wanted to participate in the study because they felt they could "give something back" to the professional who had helped them. To hopefully reduce feelings of pressure on participants, we used an opt-in strategy: people who had indicated through the gatekeeper that they were interested in our research were approached by us, but if there was no response, we assumed that they did not want to participate (cf. Speer and Stokoe 2014).

Another disadvantage of this gatekeeper strategy is that the gatekeepers decide which participants will be approached. This means that we needed to be careful and aware of which people were selected. For example, the gatekeepers may have specifically asked individuals who were known to have experienced problems during treatment, or specifically individuals who reported no problems. This may be consequential to the types of interactions we were able to collect. In addition, the gatekeepers may themselves be participants in the digital interactions (as they are professionals working at the hospital). This carries the risk that they may have selected participants with whom, for instance, they did not experience problems during treatment. It is therefore important to instruct gatekeepers not to be selective and, for example, to give them guidance on what criteria to use. We specifically asked the gatekeepers to ask all parents of children with whom they were still in contact and with whom they (had) worked via digital means whether they wanted to participate in the study. However, it was impossible for us to assess whether they used other selection criteria in the end.

A further dilemma arose in the issue of how to ask participants for their consent. Researchers have to choose between asking for signed and verbal consent, as well as between asking for consent in a setting where both researcher and participant are physically present, or not. In our case, we asked for signed consent by sending all participants digital consent forms via email. This was mainly for practical reasons, namely COVID regulations were still being imposed in the Netherlands at that time. Furthermore, potential participants were not physically in the same place (i.e., the hospital) on a regular basis because the treatment took place in a digital sphere, or because the treatments were already finished. There are several limitations to the use of email as a recruitment and consent-seeking strategy. For example, it implies a lack of personal contact, which may make people feel less engaged and possibly more inclined not to return the forms. However, the use of email can also have benefits. It may reduce the pressure to participate, because an email makes it less cumbersome to indicate that you do not want to participate (any longer), or even to not respond at all. As a result, it makes it more likely that the participants who did return their consent forms to us actually voluntarily and confidently chose to donate their materials for our study.

Email as a method for obtaining consent also raises ethical concerns. For example, how can one ascertain that people have read the information and thus can be considered informed when they give their consent? In a non-digital context, the researcher has the opportunity to verify the participant's understanding of the information (for example, by asking questions). Via email, however, the researcher cannot even be sure whether participants have opened the document at all, let alone whether they have understood the information. While this can also be a risk in a non-digital setting, it might be easier for participants to ask for clarification in co-presence. We decided not to engage in a process of back-questioning (i.e., the process of asking participants questions to ensure that they have an adequate understanding of the study [Henderson et al. 2012]) via email, because we wanted to minimize the imposition on our participants.

In this section, we reflected on some of the ethical issues of seeking consent for the use of pre-existing digital materials for a linguistic research project. While collecting our data, it proved quite challenging to implement ethical standards derived from existing guidelines. Our discussion above of the distinction between research subjects and authors in the AoIR is an example of this. Based on the exploratory nature of our methods of collecting digital data, we consider it worth reflecting on this by discussing both advantages and disadvantages of the methods we decided to use. It is important to weigh in on each step: in the realm of ethics and in the realm of feasibility (the former leading). By considering the writers of digital texts and producers of digital videos not primarily as human subjects, but emphasizing their role as creators of the material, we believe it is ethical to approach them for post-hoc informed consent to use their materials for a study of discourse patterns in digital interactions. However, this did not mean that we did not encounter further ethical challenges. In what follows we address more specifically ethical issues that arise when pre-existing digital materials are not publicly available or accessible.

3 The "non-observer's paradox" of pre-existing digital data

In Section 2 we generally addressed the question of whether pre-existing digital materials can be used in an ethically sound manner for linguistic research and discussed ethical dilemmas in trying to apply considerations in practice. In this section we address more specific issues that arise when pre-existing digital materials are password protected. Existing literature on Internet research makes little mention of non-publicly accessible or visible materials. The AoIR, for example, focuses primarily on publicly accessible digital data. Also, in literature on the ethical use of pre-existing digital data, publicly accessible online materials are overrepresented (e.g., Page et al. 2022; Paulus and Wise 2019; Spilioti and Tagg 2017; Stommel and de Rijk 2021, 2022; Sugiura et al. 2017). Our study involved digital materials that were only accessible through an account with password. This meant that virtually any information about the materials remained unknown to us as researchers until the materials reached us. The collection of the materials therefore involved a number of ethical dilemmas specifically related to our unfamiliarity with the data upfront, hence, blinding us for possible ethical considerations (e.g., we could not "unsee" what we had accidentally seen). We refer to this as the "non-observer's paradox": these are issues that arise from not being able to observe the data prior to data collection. Our discussion in this section will illustrate ways in which technology can play both an enabling and constraining role with respect to ethics and that there is a continuous dialogue between the two, where we may in the end have to settle for the best possible rather than the perfect solution.

The first issue that arises when collecting non-publicly accessible digital materials consists of whether, as a researcher, you can be sure that the digital materials in question still really exist. In our case, we found out during data collection that all communications from one specific application were automatically deleted from a person's device after a month. This was related to privacy reasons and was activated on every device on which the app was installed. The functionality had to be switched off actively by a user of the application to preserve the written interactions, photos, and videos on the device. This meant that the pre-existing digital materials that we had planned to collect actually no longer existed when we began collecting the data. Because we could only access the materials after we had permission from all participants in the app conversations, we therefore only found out afterwards that what we had asked for no longer existed. As a consequence, we decided to wait three months for new material to be generated during the treatment, which we could then collect.[7]

Other ethical problems arising from the unfamiliarity of the materials have to do with its unknown content. In our case, the digital materials consisted of text messages, videos, and images. When it comes to the content of these digital materials, it is difficult to estimate the level of sensitivity of the materials in advance. After all, as a researcher, you do not know what is being talked about or what can be seen in visual

[7] Of course, another option would have been not to collect these cases. However, we chose not to because we had already put so much time and effort into arranging informed consent and thus it would have been a waste of a lot of research time (i.e., taxpayers money).

material. It may be the case that after collecting the material, it turns out that it contains problematic or even illegal practices visible in the data, about which legal advice should be sought. In addition, visual or written materials may also reveal sensitive issues that are not related to bad practices, but, for example, to certain sensitive or intimate (family) matters. For example, in one case from our material, the videos displayed arguments and friction between two parents.[8] Because of the unfamiliarity of the materials to the researchers, one option might be to ask participants to send only materials that they are comfortable with and not to send videos, photos, or messages that they feel are private. The question then is whether participants have the same criteria as the researchers for considering things as particularly intimate or sensitive. Moreover, participants do not always share the same ethical concerns as researchers because they know their own context and know, for example, that their friends and family will not read our academic papers anyway (cf. Tagg et al. 2017). Data selectivity can also be problematic: there is no way to control what is and is not selected by participants, and the result may not allow for proper analysis of interactions.

Another issue resulting from unfamiliarity with the data is that the researcher is unable to anticipate what personal information is present in the materials. In other words, the researcher will not know in advance who is visible in the material, who is being talked about, and who is participating in a conversation. This may raise ethical issues. In our case, a consequence was that it was not always clear whose consent should be asked for prior to collecting the data. As mentioned, we sought consent in two stages: first we asked parents for their consent, and then we asked parents which professionals were involved in the digital treatment, after which we contacted these professionals. Although a more obvious solution would seem to be to use a gatekeeper here as well, this was not possible for several reasons. First, gatekeepers must be given the time by the participating organization to go through the data. It is difficult to estimate in advance how much time, effort, and expertise this will take and moreover, it is often forgotten to discuss this in advance when planning data collection. In addition, a gatekeeper (due to the GDPR) cannot access all the materials either. For example, in the case of medical data, that

[8] Although this can also happen when applied linguists record naturally occurring data, this is a different situation than in the case of collecting pre-existing material. When the data are recorded on-site, the participant has just been asked for permission to record and a recording device is present. Thus, we can then assume that the participant is aware of what is being recorded. In the case of the pre-existing material, the argument may have taken place months ago and the participant may have long since forgotten about it. Moreover, in the case of an argument being recorded on the spot, the researcher and the participant are near each other which gives both the participant the opportunity to ask afterwards to remove the recording and the researcher the opportunity to ask the participant immediately if they are comfortable with sharing this on video.

person can only access (suitable) materials from patients with whom they have a treatment relationship.

In our study, the best solution seemed to be to simply ask the participants who did have access to the materials who co-participated in the interactions. Nonetheless, some participants said they could not remember who created the materials because the treatment had taken place some time ago, or because their child had undergone so many treatments that they lost track of who was involved in what. Another problem was that some parents were not thorough in passing on the names of all professionals involved. In some cases, we only found out afterwards, when looking at the materials, that there were other professionals involved in the treatment, who had not yet given consent. In those cases, these additional professionals were contacted post hoc, which we considered to be an acceptable compromise. A related issue concerns ambiguity about whose information is shared and what kind of potentially personal information is involved. This can result in inadvertent collection of personal information. An example is a participant congratulating another participant on their birthday. If the date on which this message was sent is visible, and especially if the participant's age is also mentioned (*congratulations on your 30th birthday*), a date of birth can be "accidentally" traced from the material.

There is also an issue resulting from the non-observer's paradox that is more technological and has to do with data storage. When collecting already existing materials, it is very difficult to estimate what the size of the data will be and therefore how to handle it in a responsible and safe way. For example, data storage often requires an estimate of the capacity of the server and thus an estimate of the volume of the data. In our project, in which videos play a crucial role, it was very difficult to estimate their size in advance. At the beginning of data collection, we did not know how large the dataset would be that we were going to collect, let alone the size of the files (it depended on how many people were willing to share their materials, but also on the timeframe of a treatment). At multiple times during the retrieval of data there appeared to be too little space on our server for the videos. Thus, the amount of material is a practical issue for researchers to consider, especially when it comes to video material. This becomes an ethical issue to consider when it comes to secure data storage. When participants shared their materials with us, we wanted to be able to store them on a secure server immediately. This was hampered because new space had to be requested and obtained first.

In addition, problems may arise in transferring the material. For example, it may be unknown what the materials look like when they arrive with the researchers and whether all relevant information is (still) present. In our case, for example, initially several text messages from different treatments ended up in one document, which meant that different "threads" were mixed up. We saw all the messages in the order they were sent (according to the date they were sent), but text messages had been sent in different threads. This made it impossible to figure out which message was responding to which earlier message (or video). For our conversation analytic approach (examining how participants respond to each other), it is highly relevant that the order of messages in a digital conversation remains intact. Also, things such as time and date stamps can be relevant for analyzing the interaction and therefore should not be lost when transferring the material. It may be relevant to an analysis whether, for example, much time has elapsed between two messages. Moreover, special features of online interaction such as the use of emojis may be lost when a text is downloaded and sent in a particular format, which is also undesirable. Through trial and error, and with the help of an IT specialist, we eventually found a format that maintained the correct order of messages without loss of relevant information and features of online interaction.

The time investment required to resolve trouble due to unfamiliarity with the data prior to collection may be significant. In our case, it only appeared possible to receive the videos separate from the written interaction, whereas they had originally (during the treatment) been sent together with a specific text message. In the file that contained the written interaction, only the name of a video was visible at the point where a video had been sent. That name referred to the video file that accompanied the message, but all video files (which we received separately) had incomprehensible names. This was eventually solved using a script to rename the files, but it took a lot of time to develop this solution. In addition, we had not sought ethical permission to also share the video files with the university's technical support staff, so the first author had to do all the actual running of the program which changed the names of the videos. This requires that a program be written in such a way that it can be used by someone without advanced technical computational knowledge, which again complicates the solution.

In sum, we have argued that there are ethical aspects to technical barriers. Although using pre-existing data suggests that the data may be easily transferred, it may still require advanced technical skills from researchers and a significant time investment. While at first this may seem like a problem that can be solved technically by, for example, designing a system in such a way that this type of problem does not occur, ethical concerns may arise (see our concluding remarks at the end of the article). Moreover, research methods cannot always be based on progressive insight. In the case of pre-existing digital materials, it is therefore key to reflect on how to deal with existing drawbacks of technological systems in an ethical way. Furthermore, problems arising from unfamiliarity with the data require more time and effort not only from researchers but also from participants. This will be discussed further in the next section.

4 Reliance on participants to share pre-existing digital data

Collecting pre-generated data may suggest that participants need to do little or nothing at all for the study. After all, the materials have already been generated and only need to be transferred. In actual practice, this does not always prove to be the case. During our study, we found that collecting the pre-existing digital materials demanded more from the participants than we had anticipated. Therefore, another ethical issue pertains to obtaining pre-existing digital material that is not publicly available. When the materials can only be accessed through the participants, what does this say about their role in the research? As our study involved private digital data, accessible only through a private, password-protected account, we depended on the participants themselves to share their materials with us. Therefore, our project demanded unanticipated time and effort from the participants. Tasks that shifted towards the participants included determining who is considered a participant and the responsibility to retrieve and collect the digital material. In this section, we reflect on the relatively active role of participants in our research.

As the process of data collection in our study unfolded, we became increasingly aware that participating in our study meant to some extent that the role of participants changed from a relatively passive to a more active role. For instance, we depended on the parents, the participants in the treatment without a professional role, to find out which professionals had been involved in the digital communication. Because the content of the materials was unknown to us as researchers, we needed those who could access the materials to find out whose accounts were used in the digital treatment, in order to obtain consent to use their digital materials. This means that we relied (partly) on participants to decide who should be considered a participant in the study and whose consent should be requested.

In addition, we also depended on participants in the retrieval of the materials. They sent us the text messages, videos, and pictures. This proved to be a time-consuming and sometimes complicated task for many, with signing and sending the digital consent forms as a first barrier for some. Several participants asked how they were supposed to sign the form. There were also a number of participants who printed the form and signed it with a pen, then scanned it and sent it back through email. While this solution was perfectly fine, it did take more time for the participants than we intended. In response to questions, we sent instructions on how to sign a consent form digitally. After receiving these instructions, some participants still thought they were not capable of signing or complained about it. In some cases, participants had to send the forms multiple times because their signature did not come through correctly or because they forgot to add other information such as a date.

Accessing the materials was also difficult for some participants. In some cases, it had been a long time since they had used their account and logged into the digital environment, or sometimes participants found out that the materials were not stored. In one case, the participant thought the materials should be there, but was not able to find them, because they had somehow disappeared or been deleted.[9] Despite our instructions on how to access and save the data on their own devices, several things went wrong. For example, to save the app conversations on their phones, participants had to turn on a slider. In some cases, participants thought they had turned on the slider, but it later turned out that they had not. The text messages also had to be converted into PDF in order to send them, but this was not clear to everyone. Sending the materials took participants far more time and effort than anticipated, especially since all the video files and photos had to be sent separately to the PDF files containing the text messages. Also, the PDF files sometimes had to be sent multiple times before they were complete. It also occurred that a version was initially sent while the participant found out later that certain messages were missing. This was because each message to be stored in the PDF file had to be selected separately. Moreover, it was necessary that the videos and photos were also selected (despite the fact that they also had to be sent separately), for the time stamps to come through. This also went wrong regularly, which implied that the whole process had to be repeated. One participant explicitly complained about the procedure, which is understandable.

Our experiences with data collection raise the ethical question as to how much one may impose on participants. As a result of our digital data collection, a responsibility was placed on participants who were not trained in doing research. This can be specifically troublesome when it comes to issues such as what private information is visible in the data and who should be asked for permission to use materials. It remains difficult to determine how much may be expected of participants, and sometimes there is no other solution that is ethically sound. The problems we encountered highlight the importance of clear instructions prior to data collection, as this can avoid unnecessary and annoying practical problems for participants (such as signing digital consent forms, accessing data, sending data). However, more and better information may not be sufficient to prevent problems entirely, as it does not make the job less time consuming (it could even add to it), nor is it always foreseeable which technical dilemmas might arise.

[9] We cannot be sure, moreover, that the material actually disappeared. It could have been a way of refusing to participate in the study without losing face, for example.

5 Working in a specific institutional context

A final category of ethical issues we discuss here is related to doing research in a specific institutional context. As applied linguists, we often work with material from contexts that do not reflect our own disciplinary or professional backgrounds, for example, when examining police interrogations, therapy talk, or statements in courtrooms. Approaching research questions prevalent in institutions from multiple disciplines is often seen as the holy grail for solving societal issues in academia and beyond. For example, the Dutch Organization for Scientific Research states that "complex societal challenges such as the corona crisis, climate change, peace and security do not [adhere] to disciplinary boundaries. Inter- and trans-disciplinary collaboration is needed to provide solutions to these from within science" (NWO n.d.). However, in practice, working in and with disciplines and institutions other than your own often proves to be challenging. We will discuss some ethically sensitive issues we encountered.

The first issue relates to the concept of informed consent in inductive and qualitative research. The principle of informed consent originated in the discipline of biomedical research but has since been adopted in virtually all other disciplines. The term implies that all relevant information is known and described before the study begins. This is often the case in hypothesis-driven, or deductive research, but in case of an inductive approach, data collection and analysis will often develop during the study, and some information may not be known prior to study initiation (Byrne 2001).

In the case of inductive analysis, which by definition uses a data-driven approach, it is impossible to inform participants in advance about the specific research question that will be answered using their data. As informed consent involves telling people what you intend to study, this could be contradictory when taking an inductive approach characterized by an emergent and iterative character which allows you to remain open to unforeseen research avenues that stem from data (analysis). Such an approach, however, is not common to all participants, nor to all scientific disciplines. Although we tried to be as clear as possible when informing participants as to what the research would focus on, we found out while collecting consent forms that it was not clear to all participants what we were interested in. For example, one participant emailed us that they were eager to participate because they were happy that something could be done about the technological difficulties they experienced in uploading videos during treatment. Solving technical difficulties was beyond the scope of our research project and was never a research goal we communicated. This suggests that there are certain pre-existing expectations among participants regarding what the focus and aims of research on digital communication are (cf. Stommel and de Rijk 2022). We believe this is also related to the institutional context of the hospital within which participants were asked to participate in the study, as this may create or reinforce the impression that there is a practical benefit to them personally in participating in the study (which is potentially the case in, for example, biopharmaceutical studies involving drug trials in patients). We therefore suggest that in future studies collecting pre-existing data (from a completed treatment), potential participants are explicitly informed that they will not personally benefit from participating in the study, to avoid disappointment.

Another dilemma has to do with the possibility (and desirability) that the materials collected may be used for very different, diverse research questions which are not clear from the start. This can be difficult when materials are collected in, for example, a healthcare setting, as different ethical standards apply here. During our project, a hospital researcher indicated that according to their standards all research questions had to be clear in advance in order to be ethically reviewed. These ethical standards are not necessarily common in other disciplines, such as linguistics. In linguistics, data can be used for many different research questions, which often emerge only when looking at the data. Sometimes, issues that at first (or, from another discipline's perspective) seem to be just an "afterthought" of the research, or just a means for obtaining data, may become the subject of analysis in their own right. Examples are how people ask for consent to collect data (Speer and Stokoe 2014) or negotiations between participants about the presence of a video camera (Hutchby et al. 2012). When obtaining consent for the use of pre-existing digital data, it may therefore be appropriate to ask participants consent for a relatively broad analysis of communications and activities, rather than for a restricted research question, despite the fact that this is not typical for a biomedical research context. Thus, when research takes place in a specific institutional context, in close contact with other disciplines, adaptation of certain concepts, such as the precise formulation of the research question, may be required.

There are more ethical issues that arise when different institutions are involved in the research. The location of the research is particularly important for qualitative studies in a health context, as they take place in settings where ethical standards for other research methods (e.g., randomized controlled trials or preregistration of hypotheses to be tested) are well established and standard practice. An additional problem is that regulations are usually tied to physical locations (such as hospitals or universities), while professional ethics usually apply to individuals moving between different locations (Stark and Hedgecoe 2010). This means that some conduct is permissible when you are physically inside a hospital, but not when you are not, even if the researcher is performing exactly the same task (e.g., looking at potential research material). How to deal with this becomes even more complicated when the research takes place in a digital health environment. Consider the case where the researcher holds a position at a university but not at the hospital where the research is conducted: logically, such a researcher is legally much more restricted in what is allowed compared to researchers who are also employed by the hospital (medical doctor/researcher). This could be resolved by working with gatekeepers or asking participants to collect data themselves, but, as discussed above, this raises other ethical dilemmas. These issues could for instance be legally resolved by giving those in a treatment relationship with the patient the time (and by extension financial and/or other resources) to collect, examine, and transfer the materials to the researchers.[10]

There are also ethical issues involved in an institutional collaboration with multiple parties that have different interests. In the case of reflecting on ethical practices, this can have complicated consequences. In a hospital, the interests of patients will always be paramount, but fear of reputational damage or financial consequences due to laws and regulations also play into doing research. This may conflict with standards for transparency of methodology, for example. The primary interest of the linguist is the quality of methodological choices in analyzing the materials. When publishing articles from an inter-institutional perspective, the primary interests of different project stakeholders may cause friction. When doing deductive research (common in a medical context) one can anticipate most things and thus avoid ethical dilemmas most of the time. The emphasis is therefore on macroethics. When doing inductive, qualitative research many ethical dilemmas cannot be anticipated. Therefore, in this type of research, the conventional conception is precisely that one has to report on methodology (emphasizing micro-ethics). Ethical issues may thus be viewed and treated differently by various project stakeholders. One solution is to obtain the agreement of all parties on the final presentation of an article, for example, but it also requires making concessions. This implies that ethics is not only a process, but also a discussion.

6 Concluding remarks

We have discussed how pre-existing digital data for applied linguistic research can be used in an ethically sound way. Our article shows, however, that this is not as easy as may be presumed and that several ethical dilemmas may arise during the collection of these materials. Our article underscores the importance of microethics as complementary to guidelines, of which many already exist. Microethics requires transparency about ethical dilemmas and problems encountered in the practice of research. We have shown that when pre-existing digital data are stored in non-public

[10] There are sometimes agreements between universities and professional organizations that regulate such tasks, but this is not always the case.

digital spheres, it is difficult for researchers to define in advance in ethical protocols or through guidelines how the data will be collected and what ethical measures should be taken upfront. This emphasizes how important it is to view research ethics not just as a set of external guidelines to be applied when assessing a research project, but also as being at the center of the research, guiding decisions to be made in all phases of the research trajectory. We therefore call on linguists to recognize the ability to think about the ethical implications of their research practice as one of their core competencies. Academic journals should allow for, even encourage, reflections on these microethical processes (cf. Stommel and de Rijk 2021).

In the future, technology within (healthcare) institutions probably will be designed in a way that bypasses many of the barriers and dilemmas discussed here. This would make research on pre-existing digital data within institutions such as a hospital easier to implement, which would benefit the kind of interdisciplinary and inter-institutionally research we discussed in this article. In some areas such developments are already under way. Because of the increase in digital care during the COVID pandemic, there has been a breakthrough in the development of digital care systems, which means that pre-existing digital data can now be stored and collected in more efficient ways. For example, through a digital care portal, parents of children who participated in a remote intervention can all be emailed at once by a physician asking if they would like to participate in a research project. Portals like this could potentially also allow data retrieval with participants' consent but not requiring their practical involvement.

However, while such developments may facilitate use of pre-existing digital data, we do not make a case for technological determinism. The fact that technology does not provide seamless, but instead "seamful", access to pre-existing digital materials forces researchers and the organizations involved to reflect on and consciously address ethical issues such as those discussed in this article. If technology were developed in such ways that all barriers that we discussed in this article were removable, the possibility of unethical events would increase (see also Hoffman et al. 2021). Precisely those barriers in collecting pre-existing digital data keep us aware of the ethical risks involved in our research. It is our ethical responsibility as researchers, despite the trouble involved, to accept that very trouble because it keeps us on our toes in detecting ethical dilemmas. We should remain critical towards apparent "quick fixes" to pertinent methodological and ethical challenges. In other words, we should not want to make it too easy on ourselves.

Acknowledgments: We would like to thank the three employees of the hospital who wish to remain anonymous for critically commenting on earlier versions of this text. We would also like to thank the editors and anonymous reviewers for their critical and helpful feedback on earlier versions of this article and state that remaining errors are ours. Finally, we would like to thank CLS and iHub at Radboud University for funding the research.

Author contributions: BP and WSt were authors of the project proposal. WSt applied for ethical assessment. BP, WSp, and WSt all supervised the project. ED collected the data for the project and drafted all versions of the paper. All authors helped revise drafts.

References

Antaki, Charles (ed.). 2011. *Applied conversation analysis: Intervention and change in institutional talk*. London: Palgrave Macmillan.

AoIR. 2002. Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee. https://www.aoir.org/reports/ethics.pdf (accessed 21 June 2023).

AoIR. 2019. Internet research: Ethical guidelines 3.0. http://aoir.org/ethics3/ (accessed 21 June 2023).

Boyd, Danah Michele. 2008. *Taken out of context: American teen sociality in networked publics*. Berkeley: University of California Press dissertation.

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: Wiley-Blackwell.

Byrne, Michelle. 2001. The concept of informed consent in qualitative research. *AORN Journal* 74(3). 401–403.

De Costa, Peter I. (ed.). 2015. *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.

Eckert, Penelope. 2014. Ethics in linguistic research. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 11–26. Cambridge: Cambridge University Press.

Greene, Melanie J. 2014. On the inside looking in: Methodological insights and challenges in conducting qualitative insider research. *The Qualitative Report* 19(29). 1–13.

Henderson, Ellen M., Emily F. Law, Tonya M. Palermo & Christopher Eccleston. 2012. Case study: Ethical guidance for pediatric e-health research using examples from pain research with adolescents. *Journal of Pediatric Psychology* 37(10). 1116–1126.

Herring, Susan C. 2004. Content analysis for new media: Rethinking the paradigm. *New Research for New Media: Innovative Research Methodologies Symposium Working Papers and Readings* 2(12). 47–66.

Hoffman, Andrew S., Bart Jacobs, Bernard van Gastel, Schraffenberger Hanna, Sharon Tamar & Pas Berber. 2021. Towards a seamful ethics of Covid-19 contact tracing apps? *Ethics and Information Technology* 23(Suppl 1). 105–115.

Hutchby, Ian, Michelle O' Reilly & Nicola Parker. 2012. Ethics in praxis: Negotiating the presence and functions of a video camera in family therapy. *Discourse Studies* 14(6). 675–690.

Introna, Lucas. 2005. Phenomenological approaches to ethics and information technology. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-it-phenomenology/ (accessed 29 June 2017).

Jol, Guusje. 2020. *Police interviews with child witnesses: A conversation analysis*. Amsterdam: LOT.

Jol, Guusje & Wyke Stommel. 2016. Ethical considerations of secondary data use: What about informed consent? *Dutch Journal of Applied Linguistics* 5(2). 180–195.

Kubanyiova, Magdalena. 2008. Rethinking research ethics in contemporary applied linguistics: The tension between macroethical and microethical perspectives in situated research. *The Modern Language Journal* 92(4). 503–518.

Laine, Marlene de. 2000. *Fieldwork, participation and practice: Ethics and dilemmas in qualitative research*. London: Sage.

Markham, Annette N. & Elizabeth A. Buchanan. 2015. Ethical concerns in Internet research. In James D. Wright (ed.), *The international encyclopedia of the social and behavioral sciences*, 2nd edn 606–613. Amsterdam: Elsevier.

Markham, Annette N. & Elizabeth A. Buchanan. 2017. Research ethics in context: Decision-making in digital research. In Karin van Es & Mirko Tobias Schäfer (eds.), *The datafied society: Studying culture through data*, 201–209. Amsterdam: Amsterdam University Press.

McDermid, Fiona, Kath Peters, Debra Jackson & John Daly. 2014. Conducting qualitative research in the context of pre-existing peer and collegial relationships. *Nurse Researcher* 21(5). 28–33.

NWO. n.d. Kennisplatform voor inter- en transdisciplinair onderzoek. https://www.nwo.nl/kennisplatform-voor-inter-en-transdisciplinair-onderzoek (accessed 21 June 2023).

Page, Ruth, David Barton, Carmen Lee, Johann Wolfgang Unger & Michelle Zappavigna. 2022. *Researching language and social media: A student guide*. Abingdon: Routledge.

Parry, Ruth. 2010. Video-based conversation analysis. In Ivy Bourgeault, Robert Dingwall & Raymond de Vries (eds.), *The Sage Handbook of qualitative methods in health research*, 373–396. London: Sage.

Paulus, Trena M. & Alyssa Friend Wise. 2019. *Looking for insight, transformation, and learning in online talk*. New York: Routledge.

Rose, Heath, Jim McKinley & Jessica Briggs Baffoe-Djan. 2019. *Data collection research methods in applied linguistics*. London: Bloomsbury Academic.

Speer, Susan A. & Elizabeth Stokoe. 2014. Ethics in action: Consent-gaining interactions and implications for research practice. *British Journal of Social Psychology* 53(1). 54–73.

Spilioti, Tereza & Caroline Tagg. 2017. The ethics of online research methods in applied linguistics: Challenges, opportunities, and directions in ethical decision-making. *Applied Linguistics* 8(2–3). 163–167.

Spooren, Wilbert & Tessa van Charldorp. 2014. Challenges and experiences in collecting a chat corpus. *Journal for Language Technology and Computational Linguistics* 29(2). 83–96.

Stark, Laura & Adam Hedgecoe. 2010. A practical guide to research ethics. In Ivy Bourgeault, Robert Dingwall & Raymond de Vries (eds.), *The Sage handbook of qualitative methods in health research*, 589–607. London: Sage.

Stommel, Wyke & Lynn de Rijk. 2021. Ethical approval: None sought. How discourse analysts report ethical issues around publicly available online data. *Research Ethics* 17(3). 275–297.

Stommel, Wyke J. & Lynn de Rijk. 2022. Ethisch verantwoord onderzoek aan de hand van posts op social media: Hoe raak je hierover in gesprek met gebruikers van een platform? [Ethical research using posts on social media: How do you get into a conversation with platform users about this?]. *KWALON* 27(2). 122–133.

Sugiura, Lisa, Rosemary Wiles & Catherine Pope. 2017. Ethical challenges in online research: Public/private perceptions. *Research Ethics* 13(3–4). 184–199.

Tagg, Caroline, Agnieszka Lyons, Rachel Hu & Frances Rock. 2017. The ethics of digital ethnography in a team project. *Applied Linguistics Review* 8(2–3). 271–292.