International Journal of Social Research Methodology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsrm20

# Remote data collection in sociolinguistics: lessons from the COVID-19 pandemic

Annina Heini & Krzysztof Kredens

**To cite this article:** Annina Heini & Krzysztof Kredens (05 Oct 2023): Remote data collection in sociolinguistics: lessons from the COVID-19 pandemic, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2023.2265257

**To link to this article:** https://doi.org/10.1080/13645579.2023.2265257



Published online: 05 Oct 2023.

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2023.2265257



# Remote data collection in sociolinguistics: lessons from the COVID-19 pandemic

Annina Heini and Krzysztof Kredens

Aston Institute for Forensic Linguistics Aston University Birmingham B4 7ET UK

**ABSTRACT**

This article reports on our experience of collecting language data from informants in video-conferencing settings under a research design originally developed with face-to-face interactions in mind. We had set out to investigate whether individual stylistic features persist in different modes of textual production and designed a complex set of data-collection procedures, which we then adopted for use in a fully virtual environment with 112 informants, who were asked to provide language samples in eight discourse types. We conclude that the unintended shift to virtual settings had only a minimal impact on the volume and quality of the data. While the process was occasionally afflicted by IT-related technical issues and made extra demands on the data collector, it also created opportunities, notably around the management of interactional power asymmetries. Additional benefits included instant troubleshooting during data handover sessions and the ability to recruit participants who would not have been able to travel to face-to-face sessions.

**ARTICLE HISTORY**
Received
Accepted

**KEYWORDS**
Remote data collection; sociolinguistics; forensic linguistics; COVID-19

## Introduction

This article provides experiential reflections on the data collection process for a large-scale sociolinguistic study under suboptimal conditions, i.e. ones precluding face-to-face interaction with participants in a shared physical space. The aim of the study was to investigate the stability of idiolects (individual linguistic styles) across a number of spoken and written discourse types. The data collection process was designed in late 2020 and the data actively collected between January and June 2021, at a time when the COVID-19 pandemic forced significant changes in patterns of social organisation worldwide, including the UK, where the research team and participants were based. We discuss the realities of sociolinguistic data collection in an entirely virtual, pandemic-generated environment and consider issues around shifted power asymmetries, rapport between researcher and participants, researcher well-being, but also opportunities that a virtual data collection can bring.

## Study background

The study we report on was designed to be an empirically-based contribution to forensic linguistics, in its broadest sense a branch of applied linguistics interested in phenomena at the intersection of language and law, for example legal-linguistic practices in police interviews with suspects or in

**CONTACT** Annina Heini a.heini14@aston.ac.uk Aston Institute for Forensic Linguistics, Aston University, Birmingham B4 7ET, UK

courtroom interactions. Narrowly defined, forensic linguistics uses linguistic knowledge to provide investigative and evidential support in law enforcement contexts. This support often involves forensic authorship analysis, i.e. comparing language styles to comment on who the likely author of a particular text is, where much of the research base is informed by sociolinguistics, a discipline investigating links between aspects of language use and social factors (e.g. age, social class, or level of formal education).

The idea of identifying the authors of forensically relevant texts is based on two assumptions: that every language user has a unique linguistic style, or 'idiolect', and that features characteristic of that style will recur with a relatively stable frequency (Coulthard et al., 2011, p. 536). Hundreds of style markers and a variety of author-matching techniques have been proposed over the years, with some recent studies reporting correct attribution rates for the less complex closed-set tasks in the region of 90 per cent (e.g. Grieve, 2007; Koppel et al., 2013; Wright, 2017). However, one problem with those studies has been their tendency to use data from just one discourse type, whereas a forensically useful author identification system often needs to be able to capture stylistic similarities between texts created in different contexts and for different purposes and audiences. To help establish a protocol for how such a system would operate, an understanding of individual stylistic variation across, rather than within, discourse types is necessary. Thus, our study uses linguistic data produced in different media and contexts, and for different purposes and audiences. The data is sociolinguistically varied and it is reasonable to assume that if patterns of idiolectal stability are identified for any given author, the types of style markers responsible for that stability could be operationally useful in various kinds of other contexts.

Conceptually, this project is thus a radical departure from the kind of research design characteristic of recent authorship attribution studies. There are only a handful of studies using cross-genre and/or cross-domain data. For example, Kestemont et al. (2012) obtained some promising results using data from two genres (literary prose and theatre plays) but the findings were based on a sample of only five authors. Similarly, Statmatatos (2013) used texts written in two genres (opinion articles and book reviews) for 13 authors. Finally, Goldstein-Stewart et al. (2009) worked with language data from the categories of email, essay, interview, blog, chat and discussion, but they only used 21 participants and did not employ a linguistic framework to interpret their findings. Our study is the first of its kind in that it uses sociolinguistically dynamic samples to investigate the notion of idiolectal stability from an applied linguistic perspective.

We have collected language data from 112 participants, all of whom were, at the time of the data collection, undergraduate students at a British university. To identify idiolectal features we had to ensure a degree of homogeneity in the demographic characteristics of the group (Kredens, 2002) and so used judgment sampling (e.g. Schilling-Estes, 2007). The three conditions for participation were that the participants should be over 18 years old, native speakers of English, and students enrolled at our institution. We were also mindful of the potential correlation between interviewers' social characteristics and participants' style shift as reported in sociolinguistic studies (see Rickford and McNair-Knox (1994) for an early discussion) and decided all of the interviews would be conducted by the same single researcher.

In order to obtain a comprehensive picture of an individual's language, the data ought to be as varied as possible and include both spoken and written samples. As regards spoken data, each participant took part in a 20-minute interview on the topic of food, whereby the interviewer elicited particular discursive forms such as expressions of attitudes, time sequences, instructions etc. Furthermore, the participants completed an oral image description task while simultaneously monitoring the output of a speech-to-text program displaying their words on a screen in real time; the idea was to increase their cognitive load so that they would pay less attention to their stylistic output. For written data, each participant handed over a generic essay or other written assignment of between 2,000 and 3,000 words that they had written as part of their university studies. Furthermore, the participants shared 50 emails which they had sent from their university email account, and 100 text messages from their

mobile phone, 50 sent to a family member and 50 to a non-familial acquaintance or friend. An internet search task provided two written discourse types, namely the search strings inputted into online search engines, as well as a typed business memo based on the items researched online. The eighth and final discourse type was a set of two handwritten essays that the participants could complete in their own time. This eighth set of samples was added about four months into the data collection process and the production of the essays was offered to participants as an 'extra task' after the seven original types were successfully collected.

The eight discourse types together form a highly varied sub-corpus for each participant, covering different communication channels (emails, text messages), genres (university coursework, emails to lecturers), contexts (information exchange with family, friends and socially more distant audiences, phatic communication instances, professional contexts), and a variety of input types (full-sized keyboard, mobile phone keyboard, handwriting).

## Social science research during COVID-19

The COVID-19 pandemic has impacted humanity worldwide like few events in recent history. With such a global event also comes an abundance of related research from a variety of scientific fields. Recent bibliometric studies have found that, perhaps unsurprisingly, the vast majority of publications since the beginning of the COVID-19 pandemic have come from the health sciences; much less literature has been published in the social sciences (Aristovnik et al., 2020; Ruiz-Real et al., 2020, p. 862). But even within the health sciences, there exists research not only *on* COVID-19, but also about ethical or methodological considerations for research *during* the COVID-19 pandemic. Meagher et al. (2020), for example, explore ethical and social issues surrounding the conducting of clinical trials specifically during a global pandemic, and Solís-Cordero et al. (2021) describe switching to video-conferencing for a randomised controlled trial to assess the effectiveness of an intervention programme based on caregiver – child interaction.

Online data collection in *social science* was quite common before the pandemic and included e.g. online surveys (Van Selm & Jankowski, 2006), online interviews (Janghorban et al., 2014), and virtual ethnography (cf. Hine, 2008). For researchers using face-to-face data collection, the pandemic's effects meant that methodological adaptations had to be made in response to social-distancing measures and a move to online spaces was a natural solution (Nind et al., 2021). review social science publications from 2020 and collate a list of adaptations that researchers have made to their research design. The most prominent one, which aligns with the current research project, is the transition to online (virtual) data collection in order to comply with social ('actually physical') distancing guidelines (Nind et al., 2021, p. 10).

Interviews in their various forms are among the most commonly used data collection methods, and the widespread access to remote video-link platforms meant that even during the pandemic data collection with human subjects could take place remotely. Researchers have reflected on the move to virtual interviewing for data collection, noting benefits as well as limitations: Dodds and Hess (2020), reflecting on group interviews with families, note the benefit of minimising travel time for both parties when neither researchers nor participants have to be present in a designated physical space (see also Hensen et al., 2020; Richardson et al., 2021). Dodds and Hess (2020) also note that video-link platforms – thanks in part to the rapid growth in use since the beginning of the pandemic – are easy and intuitive to use for most participants.

A crucial point observed in online interviews, as opposed to a face-to-face setting, is the difficulty in reading body language and facial cues (Dodds & Hess, 2020, p. 211). In terms of the setting overall, Richardson et al. (2021) discuss the environments that both the researchers and their participants were working from in their particular study, where they interviewed members of police forces. In their experience, the plan of moving the interviews online during a time where most people were working from home caused initial worries about establishing and maintaining a level of

professionalism; however, they conclude upon reflection that they 'did not experience different levels of intimacy/distancing/professionalism within any of the environments people chose to be interviewed' (Richardson et al., 2021, p. 5).

It has also been shown that social isolation, as experienced by many during the different national lockdowns throughout the pandemic, negatively affects people's well-being (Rettie & Daniels, 2020; Rossi et al., 2020), and from a methodological perspective when conducting any kind of participant research it is crucial to appreciate that challenging circumstances can affect both participants and researchers, regardless of the type of research. Fell et al. (2020) emphasise that participants and researchers experience 'psychological stress and anxiety' and that a situation such as the COVID-19 pandemic is 'far removed from the conditions under which knowledge is ordinarily produced and applied, and questions around the validity of findings generated during this circumstance are inevitable' (p. 2). Giménez et al. (2020) focused on the psychological well-being of a very particular type of researcher, namely the forensic linguistic caseworker. Their research revealed that changes in the physical working environment and in the working schedule were factors affecting the caseworkers' professional activity most substantially.

An interesting aspect discussed by Nind et al. (2021) is survey researchers' concerns about data distorted by the events surrounding the COVID-19 pandemic, and many of these considerations are true for social science research that employs other data collection methods, such as the current project. Fell et al. (2020) voice a concern about the 'validity of findings over time', and that 'while there is always uncertainty about how closely the future will resemble the present, we argue that this uncertainty is now especially high' (p. 2). Burton et al. (2020) advocate the inclusion of guidance for participants on how to answer certain questions, e.g. 'we know that life has changed a lot for everyone in the country. When you are answering the survey, we would like you to answer according to your circumstances now, even if these are not normal' (p. 237; see also Will et al., 2020). In the sociolinguistic interview part of the current project, some questions were similarly adapted, e.g. 'in regular, non-COVID times, what would you say is your favourite restaurant, and why?' Sometimes clarification would come from the participants themselves; for example, when asked about what they ate in a typical day, some responded with 'do you mean at the moment [during lockdown] or when I'm at uni?' In addition to explicit 'pre-COVID' questions, the current project also added a specific question that took the unusual situation into account, 'how has the COVID-19 pandemic impacted your eating and drinking habits?'.

## Data collection methods

### *Recruitment*

With the COVID-19 pandemic showing no signs of abating towards the latter half of 2020, we decided to amend the data collection design to align with the fully virtual environment in which higher education institutions found themselves working. Under optimal circumstances, recruitment would have happened via email and by means of posters on notice boards; however, given the new context, the call for participation was sent out via email only, using a dedicated account with the institution's domain name, targeting different cohorts of students in the College of Business and Social Sciences.[1] Students who fulfilled the three conditions for participation (age 18+, native English speakers, and currently enrolled) and who expressed an interest in contributing to the study were emailed three documents: a Participant Information Sheet, containing a detailed description of the project and methods of data collection; a Privacy Notice, explaining how the researchers were going to store, manage and use the data; and a Consent Form, which the students were asked to sign to confirm they agreed to take part in the study, understood the voluntary nature of participation and were aware they could withdraw at any time. The documents were informed by careful ethical considerations which were sanctioned by the relevant university ethics committee.

Another issue the documents stipulated was the inconvenience payment that participants would be entitled to once they completed the data-sharing process. The payment would be made via email in the form of voucher codes for an e-commerce company. As will be discussed further below, the participants were initially offered payment for seven discourse types, and invited to provide the eighth type in return for additional remuneration.

### *Ethical issues*

The main concerns here were anonymisation processes of sensitive personal data, that is, emails and text messages that involved participants', addressees' and third parties' names, numbers, contact details, and other identifying information. We instated a rigorous data anonymisation protocol aimed at redacting the original data without delay using replacement tags so as to retain contextual information. Unredacted data would be deleted once the data collection process was complete and only redacted data was passed on to other researchers in the team for further analysis.

### *Demographic information*

The recruited cohort of participants is both homogenous and heterogenous: on the one hand, it comprises only undergraduate students of a relatively narrow age bracket (19–23) and enrolled in social-science and business degrees, with the majority of participants (81%) being female. On the other hand, there is diversity within the group in terms of their linguistic, ethnic, and geographical backgrounds: 38% of the participants consider themselves monolingual English speakers, and 57% declare speaking one or two languages in addition to English, with Punjabi and Urdu the most commonly listed additional languages. In terms of ethnicity and using a free-text response form, 13% label themselves as 'White British', 10% as 'Black African', with 'British Indian', 'British Pakistani' and more broadly 'British Asian' being represented as well. One third of the participants name the city where the university is located as the place where they grew up, others list a range of cities and towns predominantly in the south of the UK. A total of 120 students were recruited for the study, with eight withdrawing during the course of it, denoting a relatively low drop-out rate of 9.6%.

## Experiential reflections

Under optimal conditions, our data collection would have been split into a 'handover' part and an 'elicitation' part; we would have been collecting 'extant' and 'elicited' data (Salmons, 2016, pp. 9–10)[2] separately. The participants would have handed over emails, text messages, the university essay, and the handwritten essays, whereas the researcher would have obtained the participants' elicited data during the interview, and image description and Google search tasks in a physical space, i.e. a designated room on the university campus.

Each participant was then invited to schedule two data collection sessions using the university's scheduling software. This gave the students agency in picking slots whenever it suited them best. Slots were available most days of the week between 9 am and 6.15pm; for certain participants whose extra-curricular commitments or time zones meant they would be unable to attend the remote sessions within that time frame, alternative times could be scheduled easily. The data collection sessions would be held on Microsoft Teams, as all students and staff had institutional access to this software.

In the following sections, our data collection experiences related to all eight discourse types are described and reflected upon. The order represents that in which the data were collected; the first session covered discourse types 1 to 4 and the second covered 5 and 6. Discourse type 7 was collected without the data collector's direct involvement.

### Sociolinguistic interviews

The main challenges in conducting the interviews in a virtual environment concerned connectivity issues and background noise. Regarding the former, the majority of the participants, as well as the researcher, were in multi-occupancy households with other people using the internet at the same time, thereby limiting the bandwidth available. Some participants were in professional spaces, most notably the university's campus environment, where the internet connection was frequently unstable, too. A 2021 research report evaluating data from over 2,000 workers from a variety of sectors in the UK found that 43% of respondents did not have a reliable internet connection (O2 Business, 2021; see also Hantrais et al., 2021, p. 262). This rate aligns broadly with the ratio of data collection sessions that were in some ways impacted by connectivity issues. The issues in some cases meant that the participants had to switch off their cameras, which may have affected rapport-building, particularly when considering that the interview was the first encounter between data collector and participant (see also O'Connor et al., 2008; Richardson et al., 2021).

As for background noise, and this relates to the multi-occupancy household settings outlined above, the interview sessions were occasionally interrupted by doorbells, traffic noise, fire alarms, pets, and individuals other than the participant, including children, speaking in the background. It must be mentioned that connectivity issues and background noise were only a problem during the actual interview part of the first data collection session, as well as during the image description task in the second session. This is because these two discourse types relied solely on audio recordings of the participants' spoken output, which would later be transcribed.

In order to keep the audio recordings as 'clean' as possible, the data collector routinely informed the participants before the start of each interview that she would keep verbal feedback to a minimum while the participant was answering questions. With the inevitable time lag that comes with virtual interaction, minimal responses ('channelling') often result in overlapped speech, which in turn unnecessarily complicates transcribers' work. We therefore resolved not to step in to fill even the more prolonged silences (cf. Richardson et al., 2021). To counteract this unusual and, in linguistic terms, unnatural manner of interacting – after all, people are not used to holding monologues with no verbal feedback by the interlocutor – the data collector signalled active engagement with non-verbal cues such as nodding and smiling for those participants whose webcams were kept on.

### University essays

The second discourse type collected was a recent, single-authored university essay with a word count of 2,000 to 3,000. The data collector asked the participants to email as an attachment a piece of assessment they had produced and that fit those requirements. This handover took place during the video-call, so that the participants could ask questions and the data collector could verify the shared document(s) in real time. A positive outcome here was that the participants were as a rule using their own laptop, which is also where most of them stored their university assignments. In a physical setting, they would either have had to bring their laptops to a physical session or email the essays to the data collector in their own time. This method would have undoubtedly complicated matters as the project researchers would not have been available in to answer queries and verify the documents immediately.

### Emails

The participants were asked to forward 50 emails they had sent in connection with their studies at the university. In simple terms, this meant asking them to log in to their email account, go to the 'Sent Items' folder, and forward the 50 most recent emails to an email address that was set up specifically for the study. In cases where the participants had not sent enough emails from their

university account, we asked them to share emails also from a personal account. The email forwarding was also done in real-time during the first session; only in exceptional circumstances, e.g. where email accounts were momentarily not accessible or where the internet connection was not good enough to handle big attachments, could the participants forward the emails in their own time after the session. Similar to the collection of essays described earlier, the completion of this task during video-link enabled the data collector to verify the materials received and address issues contemporaneously.

### *Text messages*

For the final task in the first data collection session, the participants were asked to forward two batches of 50 text messages to a phone number that had been set up especially for the project. Due to its general popularity, as well as the convenient in-app message-forwarding feature, WhatsApp was chosen as the preferred platform for sharing the messages; however, some participants forwarded messages from iMessage or Facebook Messenger. First, they were asked to choose two people, one family member and one friend/acquaintance with whom they communicated via text message regularly on a one-to-one basis. They were then asked to select and forward multiple messages. Similar to the collection of essays and emails, this set-up enabled them to put clarification questions to the data collector in real time. Most of these concerned the inclusion or exclusion of certain types of messages, such as voice memos, images, etc.

The data controller was then easily able to export the text messages to the project database using the 'Export Chat' function in WhatsApp. The app's chatlogs are automatically saved as text files and can thus easily be anonymised and processed into xml format for further analysis.

Overall, the assisted data handover in the case of essays, emails and text messages worked well in two respects. Firstly, the move from asynchronous to synchronous communication between researcher and participant proved to be time-efficient; we would have very likely spent much more time trying to secure the data using the former modality. Secondly, the fact that the researcher's expectations could be voiced and responded to in real time, and any problems arising were solved collaboratively, quite likely enhanced the quality of the data.

### *Image description with speech-to-text software*

The second data collection session started with the image description task. For this, the participants were sent three numbered images from the university's stock image library that showed diverse groups of students in a variety of settings. They were asked to describe the images in as much detail as possible while also closely monitoring the graphic interface of a speech-to-text (STT) software that was transcribing the participants' speech into written language in real time. They had been sent a link to the open access STT software Speechnotes (*www.speechnotes.co*) and asked to share their screen so that the input could be visually captured from the data collector's laptop.

Screen-sharing was perhaps the most problematic aspect of the second data collection session, with many participants having never used that function. In addition to this, the Microsoft Teams interface differs slightly depending on the operating system being used. Giving detailed, step-by-step instructions to the participants on how to activate screen-sharing – without being able to see their screen at that point – proved difficult at times. In some cases, first-time screen-sharing proved time-consuming as it meant the participants had to update their privacy settings and often also reboot their laptops. Some experienced connectivity and audio issues (broken microphones and/or the aforementioned background noise), which also affected the quality of the data collected. Background noise, such as for example family members' or housemates' voices, was never picked up by the STT software; however, it happened occasionally that the participants commented on specific instances of background noise (e.g. apologising to the data collector for a noisy neighbour) and that the software transcribed those comments. These rare occasions are not deemed

problematic as the linguistic input was still produced by the actual participants, and the data is thus not considered to have been contaminated.

In an in-person setting, the participants would have completed this task using their own phone to look at the images, and a computer provided by the institution for the STT software. This set-up would have levelled the playing field in that it would have created the same conditions for all participants in terms of a stable internet connection and a good microphone. Screen-sharing would not have been necessary in a face-to-face setting as the screen and audio could have been captured on the computer the participant was using. In sum, the collection of this particular dataset was negatively impacted by the pandemic-induced move to online data collection.

### *Google search and business memo*

The final part of the second data collection session was a Google search task whereby the participants were asked to find specific information on the Internet while sharing their screen with the data collector. They responded to a scenario according to which they were the personal assistant for a CEO who was travelling to Helsinki, Finland for a conference, and it was their task to look for a number of itinerary items for her trip (hotel, a breakfast café, a restaurant, a museum, a souvenir shop and some evening entertainment).

Similar to the image description task described above, the data collector was recording the participants' shared screen during the task. No audio recordings were made, so there was no issue with background noise. Unsurprisingly, at times this task proved problematic with regards to internet connectivity. Participants that were experiencing weak internet signal would struggle with a slow browser while the active Microsoft Teams call and the screen-sharing were using up substantial amounts of bandwidth. Furthermore, and this was especially prevalent during the university's assessment season, many participants had a great number of other internet-based applications and browser tabs open. These connectivity issues resulted in some participants taking a very long time to access only a small number of websites, thereby limiting the amount of usable data obtained. In some rare cases, it was decided that the session would be cancelled and rescheduled for a day when the participants were in an environment with a more stable internet connection.

In a face-to-face setting, the participants would have used a computer provided by the institution and the screen would have been captured directly on that machine. Once again, this setting would have created a more even playing field with regard to internet speed.

Once the participants completed their online research, they were sent a link to a Microsoft Form, where they were instructed to '[w]rite a memo for the CEO with all the information you gathered for her trip to Helsinki'. There were no formatting restrictions, no word limit and no time limit for this task. On average, it took the participants 10 minutes to write their memos.

At the very end of the second data collection session the participants were asked to complete a brief demographic survey where they were asked about their age, gender, as well as their geographical, ethnic, and linguistic backgrounds.

### *Handwritten texts*

As outlined above, the eighth discourse type, handwritten texts, was added four months into the data collection process, at a time when many participants had already finished both data collection sessions and had been reimbursed for their time and effort.

In a non-pandemic, physical data collection environment, the cost of an additional discourse type would have easily outweighed the benefits, considering factors such as time, logistics, and the data collector's availability. However, thanks to our experience of managing the virtual data collection process until then, the additional data could be added without issue. The existing participants were contacted via email and invited to complete an eighth task for an extra

inconvenience payment. In order to secure as many participants as possible, it was crucial to keep the instructions and the effort required to complete the task as minimal as possible. Willing participants were sent a simple set of instructions, namely (1) get two pieces of A4 paper, (2) write two short essays of 300 words each on a generic topic using set headings, (3) take pictures or scans of the texts and (4) email them to the data collector. The task could be completed in their own time with no direct virtual engagement with the data collector ('self-recording', see Goldstein et al., 2020; Nind et al., 2021, p. 13). The participants used their mobile phones to capture and send the photos, which meant no issues to do with connectivity as experienced in the two synchronous data collection sessions.

The only issue observed in this task was that pictures of the two essay pages taken on smart phones were generally of a very high quality and thus had large file sizes. As a result, some participants encountered difficulties in sending their images via email. This was a problem that would not have occurred in a face-to-face setting where the participants would have been asked to simply submit the physical documents; however, it must be emphasised again that the overall convenience related to data hand-over and management was much higher in the virtual setting.

## Discussion

In this section we take stock of aspects of the data-collection experience that we found to be different or missing from what we had been accustomed to – and from what the literature on sociolinguistic data collection reports – when it comes to face-to-face research interviews. We discuss technical issues, opportunities afforded by the newly-emerged social and physical context, the question of rapport between researcher and participants, and of shifted power asymmetries.

### *Technical issues*

As described above, we encountered technical problems with internet connectivity and background noise. While both did interfere with the interactions on occasion, their impact on the data quality was ultimately not significant. As regards background noise in particular, the participants' devices used a variety of filtering-out algorithms (and microphones) and as a consequence the recordings we obtained varied in audio quality. For language data collection this was also found to be an issue in Leemann et al. (2020), who report on addressing Covid-related restrictions by getting their participants to use smartphone applications while being supervised via videoconferencing. For Leemann et al. (2020) one of the main technical issues was variation in the recording quality of the smartphones. However, whereas the focus of their study was on the phonetic variants of linguistic variables, with the audio quality a primary concern, our interest was always in the lexical, grammatical and pragmatic levels and the quality of our recordings proved to be more than adequate for a professional transcriber to produce good-quality transcripts. An important methodological lesson to do with background noise, however, is that its emergence and nature can be difficult to predict; participation protocols should ideally sensitise participants to the importance of making the relevant arrangements and include contingency plans should the noise interfere with the online session significantly.

Not unrelated to the technological pitfalls was the issue of the participants' ability to navigate situations where multiple demands were being made on their digital competencies. In hindsight, following the decision to move to a virtual project set-up, we probably spent too much time on devising the relevant data collection protocols but were less concerned about their uptake on the participants' part. We were making the unwarranted assumption that young students' experience and use of modern technology is uniform.[3] Similarly, we were taken aback by the scale of variation in the operating systems, digital platforms and devices the participants used.

### Social and physical context

An important aspect of the data collection process was its geographically dynamic and diverse character. The data collector was working from a variety of locations throughout, including private accommodation and offices. In order to minimise what might have felt like an invasive experience, she obscured her background in Microsoft. The participants were also working from a number of different locations, including offices, libraries and family homes. Those attending the sessions from a home environment were at times visually located on a bed or a sofa, the very casual nature of which was in clear juxtaposition with how a session would be conducted in an on-campus, face-to-face setting. This level of informality was reflected also in the way in which some participants presented themselves; one for example was wearing a cosmetic face mask. A related observation is that some participants would attend the meetings on their smartphone rather than on a tablet or laptop, which affected the quality of the interaction in that the input on the data-collector's screen was visually much less steady when the participants were holding their phones in their hands. This also fed into the informal, on-the-go atmosphere, which a face-to-face data collection setting would have most likely precluded.

Free from geographically-dictated constraints, the data collection process generally offered more flexibility, both for the researcher and the participants. In the case of short-notice cancellations, alternative participants could be contacted swiftly to try and fill the slot in a matter of minutes. This was a logistical benefit that would not have been available in a physical data collection setting, where parties are required to travel to designated places. We were able to interview students living a considerable distance from Birmingham, including a few based abroad at the time.

### Rapport

One of the challenges of sociolinguistic data collection lies in creating an environment where the informants feel at ease with talking about themselves while being audio- and/or video-recorded and with handing over personal data. To achieve this, the data collector needs to establish rapport, a quality 'that emerges between the researcher and the researched in interaction' (Manns, 2021, p. 139). What transpired in our study was that the move to virtual sessions provided additional opportunities for the interactional building of rapport. For example, to overcome the various technical obstacles (e.g. connectivity issues) both parties worked together, with the data collector giving instructions and the participants following them and reporting on progress. A successful completion of such joint tasks often seemed to work as an ice-breaker, with the participants 'primed' to the data collector's presence and demeanour before the various tasks actually started. With the two interactants having established a degree of mutual familiarity, this unplanned preparatory stage was perhaps one of the reasons why the task-completion stages followed a largely contiguous dialogic structure and only contained minimal instances of overlapped speech (see also the 'Sociolinguistic interview' section above).

### Power asymmetries

Rapport building is associated with the often marked power asymmetries arising when (perceived or actual) status differences get played out in interactions in an institutional context. As Kvale (2006, p. 485) finds, 'a research interview is not an open and dominance-free dialogue between egalitarian partners, but a specific hierarchical and instrumental form of conversation, where the interviewer sets the stage and scripts in accord with his or her research interests'. In the physical setting of a university campus, the professional environment and the interviewer's perceived status work in tandem in the stage-setting process and often put them in a position of dominance, more so if the interviewees are financially incentivised undergraduate students. If we had been able to go along with our original plan, we would have hosted the participants in an institutional environment,

but the virtual data collection meant that data collector and participant were essentially hosting each other in their respective private spaces. Thus, with the institutional semiotics of a university campus absent, the participants seemed to be actively taking part in the 'stage-setting' themselves and new frames of reference were being negotiated between participant and data-collector in dynamic interactional terms. A virtual guest in often intimate spaces (e.g. bedrooms or living rooms), the data-collector did not occupy a more powerful position any longer, and the participants were no longer mere 'subjects'. Importantly, the participants now had a choice regarding their preferred location, whose nature often turned out to be quite informal. It is not clear what all this meant for the quality of the data elicited but we can hypothesise that, even in this more relaxed virtual set-up, most of the interviewees produced language samples that reflect their idiolectal styles well. Leemann et al. (2020) did compare the reactions to an interview task of two groups of informants and found that 'the online-only cohort trended toward perceiving the interview as less stressful, less difficult, and more relaxed than the in-person interview cohort did' (p. 11).

## Conclusion

In this article we have shared our experience of, and reflections on, collecting elicited and existing language data from participants attending virtual project sessions. Overall, we believe the forced move to virtual data collection had only a minimal impact on the volume and quality of the data we collected. However, the process had its challenges, chief amongst which were those to do with disruptions engendered by technical issues and intrusion on participants' private spaces. The former, coupled with only partial access to body-language cues and shared physical points of reference meant an increase in cognitive load and thus fatigue for the data collector, who additionally had to be responding to a variety of technical issues in the sessions. This translated in the researcher working longer hours than anticipated (cf. Fell et al., 2020, p. 3).

We have also identified a number of benefits to the virtual sessions: we had access to a greater pool of potential participants, the instant troubleshooting during data handover resulted in time saving, and management of power asymmetries including rapport-building was improved.

It might well be the case that, in the post-pandemic research landscape, virtual data collection will be considered a fully viable and time- and cost-saving alternative to traditional language-data-gathering methods. In practical terms, a fast and reliable internet connection, a clear set of instructions for the participants right from the outset, and use of simple-to-implement platform-agnostic software solutions will all help ensure a positive experience for both researcher and participant. A sensitivity to the participant's needs, who may be initially confused by their role in the shared virtual space, will also be key but, as we have found, the opportunities to construct that space interactionally are there as well.

## Notes

1. We have found no information on the take-up rate in studies using similar-sized participant groups in face-to-face settings. We can only hypothesise that with no expectation of in-person attendance, so no travel or time commitment, interest in the study was higher than would have been the case in an on-campus scenario.
2. Elicited data 'is collected from consenting participants in response to researchers' questions or other prompts. The researcher can influence the direction or level of specificity and can probe in ways not possible with extant data' (Salmons, 2016, p. 10). Extant data 'includes documents, visuals, records, Big Data, and other materials the researcher examines to find evidence and background information related to the research problem (Salmons, 2016, p. 9).
3. Cf. Bennett et al. (2008) on the 'digital native' myth.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by Research England's Expanding Excellence in England scheme (E3) as part of funding for the Aston Institute for Forensic Linguistics 2019-2023.

## Notes on contributors

***Annina Heini*** is a Postdoctoral Research Associate at the Aston Institute for Forensic Linguistics at Aston University. As a member of the Centre for Forensic Text Analysis she contributes to research on individual language change across genres to inform practice in forensic authorship analysis. She is also interested in spoken interaction in legal contexts, especially police interview discourse, with a focus on age-based vulnerability and children's access to justice.

***Krzysztof Kredens*** is Reader in Forensic Linguistics in the College of Business and Social Sciences at Aston University. His research is based at the Aston Institute for Forensic Linguistics, where he heads the Centre for Forensic Text Analysis. He has a variety of research outputs in forensic linguistics and ample casework experience, both as an expert witness and in policing contexts. He is registered on the UK National Crime Agency's Expert Advisors Database.

## Ethics

Ethical approval was obtained from the relevant institutional ethics body.

## References

Aristovnik, A., Keržič, D., Ravšelj, D., Tomaževič, N., & Umek, L. (2020). Impacts of the COVID-19 pandemic on life of higher education students: A global perspective. *Sustainability*, *12*(20), 8438. https://doi.org/10.3390/su12208438

Bennett, S., Maton, K., & Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, *39*(5), 775–786. https://doi.org/10.1111/j.1467-8535.2007.00793.x

Burton, J., Lynn, P., & Benzeval, M. (2020). How understanding society: The UK household longitudinal study adapted to the COVID-19 pandemic. *Survey Research Methods*, *14*(2), 235–239. https://doi.org/10.18148/srm/2020.v14i2.7746

Coulthard, M., Grant, T., & Kredens. (2011). Forensic Linguistics. In R. Wodak, B. Johnstone, & P. E. Kerswill (Eds.), *The SAGE handbook of Sociolinguistics* (pp. 529–544). SAGE. https://doi.org/10.4135/9781446200957.n36

Dodds, S., & Hess, A. C. (2020). Adapting research methodology during COVID-19: Lessons for transformative service research. *Journal of Service Management*, *32*(2), 203–217. https://doi.org/10.1108/JOSM-05-2020-0153

Fell, M. J., Pagel, L., Chen, C.-F., Goldberg, M. H., Herberz, M., Huebner, G. M., Sareen, S., & Hahnel, U. J. J. (2020). Validity of energy social research during and after COVID-19: Challenges, considerations, and responses. *Energy Research & Social Science*, *68*, 1–7. https://doi.org/10.1016/j.erss.2020.101646

Giménez, R., Elstein, S., & Queralt, S. (2020). The pandemic and the forensic linguistics caseworker's wellbeing: Effects and recommendations. *International Journal of Speech, Language & the Law*, *27*(2), 233–254. https://doi.org/10.1558/ijsll.19548

Goldstein-Stewart, J., Winder, R., & Sabin, R. (2009). Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL2009)*, Athens, Greece, 30 March–3 April 2009. 336–344.

Goldstein, R. Z., Vasques, R. A., & Loschiavo dos Santos, M. C. (2020). Doing design research with youth at/from the margins in pandemic times: Challenges, inequalities and possibilities. In Kara, H., S.-M. Khoo, Eds. *Researching in the age of COVID-19 :Volume Creativity and Ethics* Vol. 3. Policy Press. https://doi.org/10.2307/j.ctv18dvt3x.16

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, *22*(3), 251–270. https://doi.org/10.1093/llc/fqm020

Hantrais, L., Allin, P., Kritikos, M., Sogomonjan, M., Anand, P. B., Livingstone, S., Williams, M., & Innes, M. (2021). Covid-19 and the digital revolution. *Contemporary Social Science*, *16*(2), 256–270. https://doi.org/10.1080/21582041.2020.1833234

Hensen, B., Mackworth -Young, C. R. S., Simwinga, M., Abdelmagid, N., Banda, J., Mavodza, C., Doyle, A. M., Bonell, C., & Weiss, H. A. (2020). Remote data collection for public health research in a COVID-19 era: Ethical implications, challenges and opportunities. *Health Policy and Planning*, *36*(3), 360–368. https://doi.org/10.1093/heapol/czaa158

Hine, C. (2008). Virtual ethnography: Modes, varieties, affordances. *The SAGE Handbook of Online Research Methods*, 24, 257–270. https://doi.org/10.4135/9780857020055.n14

Janghorban, R., Roudsari, R. L., & Taghipour, A. (2014). Skype interviewing: The new generation of online synchronous interview in qualitative research. *International Journal of Qualitative Studies on Health and Well-Being*, *9*(1), 241–252. https://doi.org/10.11.11/j.1467-8535.2007.00793.x

Kestemont, M., Luycjx, K., Daelemans, W., & Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, *93*(3), 340–354. https://doi.org/10.1080/0013838X.2012.668793

Koppel, M., Schler, J., & Argamon, S. (2013). Authorship attribution: What's easy and what's hard? *Journal of Law and Policy*, *21*(2), 317–331. https://doi.org/10.1111/j.1467-8535.20.07.00793.x

Kredens, K. (2002). Towards a corpus-based methodology of forensic authorship attribution: A comparative study of two idiolects. In B. Lewandowska-Tomaszczyk (Ed.), *PALC'01: Practical applications in language Corpora* (pp. 405–437). Peter Lang.

Kvale, S. (2006). Dominance through interviews and dialogues. *Qualitative Inquiry*, *12*(3), 480–500. https://doi.org/10.1177/1077800406286235

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., & Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, *6*(3), 1–16. https://doi.org/10.1515/lingvan-2020-0061

Manns, H. (2021). Alignment and belonging in the sociolinguistic interview. In Z. Goebel (Ed.), *Reimagining rapport* (pp. 139–158). Oxford University Press. https://doi.org/10.1093/oso/9780190917074.003.0008

Meagher, K. M., Cummins, N. W., Bharucha, A. E., Badley, A. D., Chlan, L. L. & Wright, R. S. (2020). COVID-19 Ethics and Research. *Mayo Clinic Proceedings*, *95*(6), 1119–1123. https://doi.org/10.1016/j.mayocp.2020.04.019

Nind, M., Coverdale, A., & Meckin, R. (2021). *Changing social research Practices in the context of Covid-19: Rapid evidence review*. National Centre for Research Methods.

O2 Business. (2021) Creating a dynamic workforce: Empowering employees for productivity and growths. Research Report [Online]. [Accessed June 21] Available from: https://doi.org/10.1111/j.1467-8535.20/07.00793.x

O'Connor, H., Madge, C., & Shaw, R., Wellens J. 2008. Internet-based interviewing. In N. In: Fielding, R. M. Lee, & G. Blank, Eds. *The SAGE handbook of online research methods* pp. 271–289. SAGE Publications. https://doi.org/10.4135/9780857020055.n15

Rettie, H., & Daniels, J. (2020, August 3). Coping and tolerance of uncertainty: Predictors and mediators of mental health during the COVID-19 pandemic. *American Psychologist*, *76*(3), 427–437. Advance online publication. https://doi.org/10.1037/amp0000710

Richardson, J., Godfrey, B., & Walklate, S. (2021, January-April). Rapid, remote and responsive research during COVID-19. *Methodological Innovations*, *14*(1), 1–9. https://doi.org/10.1177/20597991211008581

Rickford, J. R., & McNair-Knox, F. (1994). Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. *sociolinguistic perspectives on register*. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 235–276). Oxford University Press.

Rossi, R., Socci, V., Talevi, D., Mensi, S., Niolu, C., Pacitti, F., DiMarco, A., Rossi, A., Siracusano, A., & DiLorenzo, G. (2020). COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. *Frontiers in Psychiatry*, *11*(790). https://doi.org/10.3389/fpsyt.2020.00790

Ruiz-Real, J. L., Nievas-Soriano, B. J., & Uribe-Toril, J. (2020). Has COVID-19 gone viral? An overview of research by subject area. *Health Education and Behavior*, *47*(6), 861–869. https://doi.org/10.1111/j.1467-8535.2007.007.93.x

Salmons, J. (2016). *Doing Qualitative research online*. SAGE. https://doi.org/10.4135/9781473921955

Schilling-Estes, N. (2007). Sociolinguistic fieldwork. In R. Bayley & C. Lucas (Eds.), *Sociolinguistic variation: Theories, methods, and applications* (pp. 165–189). Cambridge University Press. https://doi.org/10.1017/CBO9780511619496.010

Solís-Cordero, K., Lerner, R., Marinho, P., Camargo, P., Takey, S., & Fujimori, E. (2021). Overcoming methodological challenges due to COVID-19 pandemic in a non-pharmacological caregiver-child randomly controlled trial. *International Journal of Social Research Methodology*, *25*(5), 687–696. https://doi.org/10.1080/13645579.2021.1933067

Statmatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, *21*(2), 421–439.

Van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. *Quality and Quantity*, *40*(3), 435–456. https://doi.org/10.1111/j.1467-85.35.2007.00793.x

Will, G., Becker, R., & Weigand, D. (2020). COVID-19 lockdown during field work. *Survey Research Methods*, *14*(2), 247–252. https://doi.org/10.18148/srm/2020.v14i2.7753

Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, *22*(2), 212–241. https://doi.org/10.1075/ijcl.22.2.03wri