Getting “good” data in a pandemic, part 1: assessing the validity and quality of data collected remotely

Kostadinova, V.; Gardner, M.H.

DOI
10.1515/lingvan-2023-0170

Publication date
2023

Document Version
Final published version

Published in
Linguistics Vanguard

License
CC BY

Link to publication

Citation for published version (APA):
Kostadinova, V., & Gardner, M. H. (2023). Getting “good” data in a pandemic, part 1: assessing the validity and quality of data collected remotely. *Linguistics Vanguard*, *9*(s4), 329–334. Advance online publication. https://doi.org/10.1515/lingvan-2023-0170

Viktorija Kostadinova* and Matt Hunt Gardner

Getting “good” data in a pandemic, part 1: assessing the validity and quality of data collected remotely

https://doi.org/10.1515/lingvan-2023-0170
Received November 21, 2023; accepted November 28, 2023; published online December 22, 2023

Abstract: The articles presented in this special issue contribute to recent scholarship on remote data collection. The topics covered can be described in terms of two focal areas. The first focus is on the ways in which research can be adapted to remote data collection, and the second on the ways in which data collected remotely should be considered alongside data collected using “traditional” methods. The overarching epistemological question uniting these focal areas is whether remote data collection yields data of substantive quality. While varied in their aims and approaches, the studies presented suggest that remote data collection methods can be used on a par with traditional approaches, thus aligning with the findings of already existing scholarship on remote data collection methods. The detailed findings presented in the papers provide valuable methodological information for further development of methods in sociolinguistics and related areas. Though these studies originated in conditions when remote data collection was the only option, they reveal the suitability of remote data collection methods beyond the COVID-19 pandemic. Remote methods can get “good” data; the experience of conducting fieldwork during the pandemic, while a challenge, was a catalyst for creativity, innovation, and enrichment in the field.

Keywords: remote data collection; sociolinguistics; methodology; corpus linguistics; acoustic data quality

1 Introduction

Sociolinguistics as a broad research area is characterized by methodological plurality; specific sociolinguistic strands, however, have conventions for central methods and types of evidence. First-wave variationist sociolinguistic research (Eckert 2012) primarily relies on naturalistic conversational language data, collected through face-to-face interviews in settings in which speakers feel comfortable. Face-to-face interviews are similarly highly valued in third-wave studies, though they are typically situated within wider ethnographic and participant-observation methodologies. Studies of attitudes and perceptions, on the other hand, have a preference for experimental methods, where data collected in laboratory settings is the first choice. High-quality acoustic recordings with maximum acoustic fidelity, (historically) most easily produced using sound booths or other laboratory-like settings, are also prized by sociophoneticians. All of these regular data collection regimes were disrupted in 2020 by the COVID-19 pandemic.

The safety measures required to limit the spread of the virus (travel restrictions, masking, mandatory social distancing, institutional shutdowns, etc.), along with the (often repeated) infections among possible participants and researchers themselves, made interacting with participants *in situ* or at a university laboratory incautious at best, reckless or dangerous at worst, and ethically dubious. If research was to be conducted amid these measures, sociolinguists and researchers in related disciplines needed to pivot, either to remote data collection or to significantly modified *in situ* protocols focused on minimizing risk to participants, to the community, and to the researchers themselves. For researchers, especially PhD students and postdoctoral

***Corresponding author: Viktorija Kostadinova**, Modern Foreign Languages and Cultures, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands, E-mail: v.kostadinova@uva.nl
Matt Hunt Gardner, Linguistics, Philology and Phonetics, University of Oxford, Oxford, UK. https://orcid.org/0000-0002-1878-4232

researchers for whom simply not collecting data was impracticable, this need to adapt and improvise in real-time led to palpable unease.

In fall 2020, recognizing the growing need for unity and collective creativity, an international group of early-career researchers came together virtually for a workshop series called Getting Data: Linguistic Data Collection in the Age of Pandemic, with the aim of confronting the disruption to data collection through collaborative ideation and mutual support. This new network for knowledge exchange became a springboard for small-scale feasibility studies of new or modified methods and culminated in a series of virtual showcase meetings in March 2021. The ideas sparked and fleshed out by this process, and others that have since been developed, are presented in a two-part special issue of *Linguistics Vanguard*, including the present issue and a second part which is expected to come out in 2024. Presenting this research after a return to "normality" is an opportunity to share what worked to get "good" data during the pandemic and also to reflect critically on the qualities that make certain data "good" in the first place.

2 The state of the art

The articles in part 1 of this special issue join recent scholarship specific to remote data collection in the context of the pandemic. The earliest of these discussed the successes and challenges of collecting linguistic data with smartphone applications (Leemann et al. 2020) and evaluated the acoustic quality of sound data collected across different devices (smartphones, tablets, laptop microphones, professional equipment) and contexts (mainly in person vs. using videoconferencing software like Zoom, Microsoft Teams, and Skype; Calder and Wheeler 2022; Calder et al. 2022; Freeman and De Decker 2021; Zhang et al. 2020). Later articles provided further insights into technological options and innovation in remote data collection (Hilton and Leemann 2021), the effects of masking on production and perception (Smiljanic et al. 2021), as well as the ways in which sociolinguists adapted to remote data collection (e.g., Sneller 2022). Overall, the use of videoconferencing platforms to collect sociophonetic data appears to be viable with judicious choices of microphone and normalization method. It is appropriate for research questions involving the relative arrangements of vowels and categorical determinations of merger, though more caution is warranted for questions that focus on low vowels, that rely on small differences to determine relative distance or overlap for vowels, or that scrutinize sibilant qualities (Calder and Wheeler 2022; Freeman and De Decker 2021; Sanker et al. 2021). Directly comparing acoustic measurements (rather than general patterns) between in-person professional recordings and videoconferencing recordings may also be inadvisable. Word recognition and recall of speech produced with a mask also appears to be as accurate as without a mask in optimal listening conditions for L1 speakers (Smiljanic et al. 2021).

While the above studies were indeed spurred by the pandemic and the exigencies of doing research in lockdown, many of the methods used and evaluated were not new. Leemann et al. (2020), for instance, follows their previous work on dialect data collection using mobile apps (Leemann et al. 2018), while Hall-Lew et al.'s (2022) use of self-recording builds on a longer period of using these methods (Hall-Lew and Boyd 2017, 2020). What the pandemic did was to restrict researchers to the use of only remote data collection methods, and thus for these methods to be employed to an unprecedented extent.

Using smartphone apps or videoconferencing software to collect data largely solved the issue of safe access to participants and the fidelity of such recordings was straightforward to compare to professional recordings to determine suitability for acoustic analysis; however, a remaining lacuna was whether a change from in-person data collection to remote data collection affected how participants themselves spoke, signed, or reacted to experimental stimuli. Bleaman et al. (2022) refer to this as "medium shift", though the authors did hypothesize that medium shift is likely below the level of awareness and that it may be short-lived without lasting consequences: as people become more familiar with videoconferencing software, the instinct to medium shift will wane. This sentiment was echoed by Sali A. Tagliamonte, who has published widely on variationist sociolinguistics methods, in an interview with the second author about research methodologies during the pandemic:

“I’m really hoping that people have become so used to interacting on Zoom or Google Hangouts or whatever, that they’ll just talk like they normally talk, but that bears testing later on” (quoted in Gardner 2020).

The domain of qualitative research, which also employs face-to-face interviews, has equally needed to rethink data collection in the wake of the COVID-19 pandemic. The experience of qualitative researchers who have transitioned to videoconferencing software suggests sociolinguists have nothing to fear and much to gain. For example, Nguyen et al. (2022: 1), who call remote fieldwork “the ‘new normal’ in the COVID-19 pandemic era”, found success in hiring local Vietnamese research assistants (RAs) to conduct interviews and training/mentoring them via videoconferencing software, echoing methods proposed by Tagliamonte (in Gardner 2020). The authors concluded that the quality of their data was sufficient and that the hiring of local RAs led to additional benefits: the RAs came from the same ethnic minority groups as the research participants so they were also the target beneficiaries of the research; the process helped develop the RAs’ confidence, knowledge, and research skills; and the RAs also learned how to use their knowledge to help women and girls in their communities recognize their value. Khan and MacEachen (2022) contend that, for ethnographic interviews, videoconferencing is more affordable for research teams and, for many research participants, more accessible than face-to-face interviews. They provide “a unique opportunity for researchers and participants by compressing the time-space divide, facilitating safety, reducing travel-related expenses, accessing transnational participants, maintaining social distance, and protecting personal space and privacy” (Khan and MacEachen 2022: 1; see also Archibald et al. 2019). Nesbitt and Watts (2022: 344) came to similar conclusions after using various online tools for remote sociolinguistic data collection among Black Bostonians. For example, their two student interviewers produced 56 one-hour interviews and 15 ten-minute follow-up interviews in four months (June–October 2020), while, for a prior research project using traditional in-person methods (Nesbitt 2021), the same number of interviews took three times as long to collect. Watson and Lupton (2022: 10), who label remote fieldwork “agile research”, found that interviewing participants via Zoom permitted a rich sense of rapport and mutual empathy, “as well as an informality that may not have been as well achieved during an in-personal home visit”. While under normal conditions ethnographic research would be conducted in participants’ homes, using Zoom created a “mutual window of feeling” because it allowed participants to also see into the home of the researcher. Żadkowska et al. (2022) also note that “oversensitivity to the interaction context – practices which particularly women are socialized to follow (e.g., taking care of the space in which the partners interact, ensuring the sense of comfort, attention to principles of friendly conversation or paralinguistic communication) – becomes less of an issue in the online interview situation”. Minimizing the distracting pressures of hospitality is presumably also a desideratum for sociolinguistic research that could be achieved through online remote data collection.

3 New insights

The general theme of the present issue is the remote collection of linguistic data in a range of linguistic subfields. The topics covered by its contributors can be described in terms of two focal areas. The first focus is on the ways in which research can be adapted to remote data collection and the second on the ways in which data collected remotely should be considered alongside data collected using “traditional” methods. The overarching epistemological question uniting these focal areas is whether remote data collection yields data of substantive quality. Validity can be evaluated through technical criteria, such as the extent to which Zoom recordings result in acceptable acoustic fidelity or the relative dropout rates of smartphone app surveys versus those administered in person. Additionally, it can be substantiated by demonstrating that data obtained rigorously and remotely yields informative and consequential findings and conclusions.

The first paper in the collection, Robert Marcelo Sevilla’s “Yiyang Xiang vowel quality: comparability across recording media”, is doubly motivated. The more immediate aim is to compare acoustic vowel quality captured through Zoom to that captured with professional equipment in sound booth recordings. The secondary aim is to consider this question in light of the practical advantages of using videoconferencing software to connect with speakers of lesser studied languages or who live in remote or (geographically, socially, medically, etc.) isolated communities. Sevilla focuses on Yiyang, a dialect of Xiang (Sinitic) spoken in the northern Hunan Province of China. His acoustic analysis of recordings generated by the same informant retelling the events of the Pear Stories video resulted in consistent F1 measurements across recording condition (recorded over Zoom vs. recorded with professional equipment in a sound booth); however, F2 and F3 were somewhat distorted in the Zoom recordings. Despite this, Sevilla found that F2 and F3 distortion could be mitigated considerably if the formants were determined with manual inspection of spectrograms rather than using Praat's built-in formant tracker. His conclusion is that the sacrifice of acoustic precision with Zoom data must be weighed carefully on a case-by-case basis against the advantages that videoconferencing offers for reaching remote speakers.

The second paper, "Using social media as a source of analysable material in phonetics and phonology – lenition in Spanish" by Karolina Broś, shows the potential of using audio recordings made via social media as a way of obtaining good quality data for phonetic and phonological analysis. Broś compares recordings made in a lab specifically for research purposes to voice recordings that participants produced and sent as WhatsApp messages to their friends. She compares both the quality and quantity of lenited (i.e., voiced) voiceless consonants, a feature of vernacular Gran Canarian Spanish, across both data types. For duration, intensity, voicing, and formant structure, both recording types produced equivalent measurements in Praat; however, the actual rates of the occurrence of lenition were much higher in the WhatsApp recordings. Broś's study thus provides compelling evidence that, given the similarity in *relevant* acoustic qualities, naturally occurring voice messages may provide "better" data than data recorded in a laboratory.

The paper by Tessa Bent, Holly Lind-Combs, Rachael F. Holt, and Cynthia Clopper, titled "Perception of regional and nonnative accents: a comparison of museum laboratory and online data collection", tests the extent to which speech is intelligible in perception tasks when listeners are in different research settings: retelling what they heard in audio stimuli while in a museum laboratory versus writing down what they heard in audio stimuli presented online. The authors' careful design incorporated different accents and noise levels in the stimuli in order to assess how both play a role in intelligibility. Despite small differences across accent conditions, the intelligibility in both settings was very high in the quiet condition. With minimal background noise, there were bigger differences between the museum laboratory and the online setting, though here, too, differences were not striking. Finally, increased background noise resulted in increased discrepancy, with online participants scoring higher for intelligibility than museum laboratory participants. The authors provide a rich discussion of these findings in light of the effect of participants' own accents and differences between writing responses and providing responses vocally. The authors also suggest that their online participants' exposure to a wider variety of accents may have played a role in their superior ability to parse speech correctly in noisy conditions, a finding likely unavailable to the authors had only the museum laboratory experiments using local participants been conducted. This highlights the unexpected benefit of conducting online experiments and the potential for new and exciting linguistic insights such methods make possible.

The next paper is "Brazilian Portuguese-Russian (BraPoRus) corpus: automatic transcription and acoustic quality of elderly speech during the COVID-19 pandemic" by Irina A. Sekerina, Anna Smirnova Henriques, Aleksandra S. Skorobogatova, Natalia Tyulina, Tatiana V. Kachkovskaia, Svetlana Ruseishvili, and Sandra Madureira. Their paper documents their work in collecting a corpus of a moribund heritage Russian variety spoken in Brazil by a small number of elderly speakers. Data from their participants was collected remotely during the pandemic, through either Zoom or over the telephone, and was used to test different software options for automatic transcription. The authors found that automatic transcription with Sonix had the lowest word error rates, and that there was no difference in its performance on Zoom recordings and on phone call recordings. In terms of the acoustic quality of the data, the authors found a marginal advantage of Zoom recordings, as the sound quality was judged better overall than that of the phone recordings. Nonetheless, both types of recording were determined to be suitable for capturing intonation patterns and speech rate. This study highlights that new communication media, like Zoom, can be used successfully to conduct studies with remote elderly participants – a boon to researchers across linguistic fields.

The final paper in this collection presents another project involving corpus creation, this time addressing work with young participants. In "Outsourcing teenage language: a participatory approach for exploring speech and text messaging", Kadri Koreinik, Aive Mandel, Maarja-Liisa Pilvik, Kristiina Praakli, and Virve-Anneli Vihman document the process of collecting a corpus of Estonian teenager speech, as part of the Teen Speak project. The project's aim is the compilation of the first corpus of Estonian teenager language, including spoken and text message conversations. The authors deploy a citizen science approach that incorporates remote data collection. "Outsourcing" data collection to citizen sociolinguists resulted in a corpus of adequate size for robust sociolinguistic analysis. The authors therefore suggest that it is a good option for data collection. The authors also suggest that a similar approach can be applied to other studies, though they point out that building trust with the citizen scientists is paramount.

While varied in their aims and approaches, the articles in part 1 of this issue suggest that remote data collection methods can be used on a par with traditional approaches, thus aligning with the findings of already existing scholarship on remote data collection methods referred to above. The detailed findings presented in the papers on the comparability and validity of data collected in different settings provide valuable methodological information for further development of methods in sociolinguistics and related areas. Though these studies originated in conditions when remote data collection was the only option, they reveal the suitability of remote data collection methods beyond the pandemic. Remote methods can get "good" data; the experience of conducting fieldwork during the pandemic, while a challenge, was a catalyst for creativity, innovation, and enrichment in the field.

References

Archibald, Mandy M., Rachel C. Ambagtsheer, Mavourneen G. Casey & Michael Lawless. 2019. Using Zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods* 18(1). 1–8.

Bleaman, Isaac L., Katie Cugno & Annie Helms. 2022. Medium-shifting and intraspeaker variation in conversational interviews. *Language Variation and Change* 34(3). 305–329.

Calder, Jeremy & Rebecca Wheeler. 2022. Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for sibilant analysis. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2021-0014.

Calder, Jeremy, Rebecca Wheeler, Sarah Adams, Daniel Amarelo, Katherine Arnold-Murray, Justin Bai, Meredith Church, Josh Daniels, Sarah Gomez, Jacob Henry, Yunan Jia, Brienna Johnson-Morris, Kyo Lee, Kit Miller, Derrek Powell, Caitlin Ramsey-Smith, Sydney Rayl, Sara Rosenau & Nadine Salvador. 2022. Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2020-0148.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41(1). 87–100.

Freeman, Valerie & Paul De Decker. 2021. Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *Journal of the Acoustical Society of America* 149(2). 1211–1223.

Gardner, Matt Hunt. 2020. In conversation with Professor Sali A. Tagliamonte [blog post]. In Matt Hunt Gardner & Viktorija Kostadinova (eds.), *Getting data: Linguistic data collection in the age of pandemic*. https://gettingdata.humanities.uva.nl/?p=1331 (accessed 30 April 2022).

Hall-Lew, Lauren & Zac Boyd. 2017. Phonetic variation and self-recorded data. *University of Pennsylvania Working Papers in Linguistics* 23(2). 86–95.

Hall-Lew, Lauren & Zac Boyd. 2020. Sociophonetic perspectives on stylistic diversity in speech research. *Linguistics Vanguard* 6(s1). 20180063.

Hall-Lew, Lauren, Claire Cowie, Catherine Lai, Nina Markl, Stephen Joseph McNulty, Shan-Jan Sarah Liu, Clare Llewellyn, Alex Beatrice, Zuzana Elliott & Anita Klingler. 2022. The Lothian Diary Project: Sociolinguistic methods during the COVID-19 lockdown. *Linguistics Vanguard* 8(s3). 321–330.

Hilton, Nanna Haug & Adrian Leemann. 2021. Editorial: Using smartphones to collect linguistic data. *Linguistics Vanguard* 7(s1). 20200132.

Khan, Tauhid Hossain & Ellen MacEachen. 2022. An alternative method of interviewing: Critical reflections on videoconference interviews for qualitative data collection. *International Journal of Qualitative Methods* 21. 1–12.

Leemann, Adrian, Péter Jeszenszky, Carina Steiner, Melanie Studerus & Jan Messerli. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard* 6(s3). 20200061.

Leemann, Adrian, Marie-José Kolly & David Britain. 2018. The English Dialects app: The creation of a crowdsourced dialect corpus. *Ampersand* 5. 1–17.

Nesbitt, Monica. 2021. The rise and fall of the Northern Cities Shift: Social and linguistic reorganization of trap in twentieth-century Lansing, Michigan. *American Speech* 9(3). 332–370.

Nesbitt, Monica & Akiah Watts. 2022. Socially distanced but virtually connected: Pandemic fieldwork with Black Bostonians. *Linguistics Vanguard* 8(s3). 343–352.

Nguyen, Phuong, Regina Scheyvens, Alice Beban & Samantha Gardyne. 2022. From a distance: The "new normal" for researchers and research assistants engaged in remote fieldwork. *International Journal of Qualitative Methods* 21. 1–13.

Sanker, Chelsea, Sarah Babinski, Roslyn Burns, Marisha Evans, Jeremy Johns, Juhyae Kim, Slater Smith, Natalie Weber & Claire Bowern. 2021. (Don't) try this at home! The effects of recording devices and software on phonetic analysis. Language 97(4). e360–e382.

Smiljanic, Rajka, Sandie Keerstock, Kirsten Meemann & Sarah M. Ransom. 2021. Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *Journal of the Acoustical Society of America* 149(6). 4013–4023.

Sneller, Betsy. 2022. COVID-era sociolinguistics: Introduction to the special issue. *Linguistics Vanguard* 8(s3). 303–306.

Watson, Ash & Deborah Lupton. 2022. Remote fieldwork in homes during the COVID-19 pandemic: Video-call ethnography and map drawing methods. *International Journal of Qualitative Methods* 21. 1–12.

Żadkowska, Magdalena, Bogna Dowgiałło, Magdalena Gajewska, Magdalena Herzberg-Kurasz & Marianna Kostecka. 2022. The sociological confessional: A reflexive process in the transformation from face-to-face to online interview. *International Journal of Qualitative Methods* 21. 1–12.

Zhang, Cong, Kathleen Jepson, Georg Lohfink & Amalia Arvaniti. 2020. Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings. *Journal of the Acoustical Society of America* 148(4). https://doi.org/10.1121/1.5147535.