Nanna Haug Hilton* and Adrian Leemann

Editorial: using smartphones to collect linguistic data

https://doi.org/10.1515/lingvan-2020-0132
Received November 24, 2020; accepted December 3, 2020

Abstract: In the last decade, we have seen a number of studies come to life in which the collection of data for linguistic research has not followed a traditional path of holding in-person interviews or experiments, or using surveys for eliciting judgements, but instead have made use of smartphone technology and applications for collecting data. The current collection is the first to include papers with reflections from the linguistics community about the use of smartphone technology for linguistic research. The scope of the projects presented in this collection is a broad one. They have the mode of data collection, i.e. through a phone, in common, but all present different opportunities and challenges. The studies discussed in this introduction use smartphones to investigate language variation and change, clinical linguistics, psycholinguistics, and the sociology of language. Our hope is that this issue will provide ideas and inspiration, as well as access to readily usable tools, to keep researchers working, in a remote fashion, towards increasing our understanding the human competence of language.

Keywords: covid linguistics; remote data collection; smartphones; web-based research

1 Introduction

As of writing this editorial introduction, in the summer of 2020, the science of linguistics has had to adapt to a new reality where close physical contact with human subjects for gathering data has become almost impossible. Along with the rest of society, our discipline has had to consider different ways of tackling the challenges brought about by a pandemic. While entirely unplanned (this collection has been in the making since 2018), we find ourselves in a situation where a special issue on how to gather data for linguistic research through the means of smartphones, as opposed to in-person interactions, is more timely than ever. Our hope is that this issue will provide ideas and inspiration, as well as access to readily usable tools, to keep researchers working, in a remote fashion, towards increasing our understanding the human competence of language.

The special issue was born out of a wish to gather insights from the researchers in different linguistic sub fields working with smartphones to collect or process data for their research. Crowdsourcing language data with smartphones has become increasingly popular in the last decade. News outlets (Newzoo 2020) indicate that close 80–85% of the population in Western Europe own a smartphone, and millions report to make regular use of social media through handheld devices. This scenario presents linguists with unprecedented opportunities for interacting with speech communities remotely. In the last decade we have seen a number of studies come to life in which the collection of data for linguistic research has not followed a traditional path of holding in-person interviews or experiments, or using surveys for eliciting judgements, but instead have made use of smartphone technology and applications for collecting data. Applications for iPhones or Android devices have, for example, been used by linguists to collect speech for development of language technology (De Vries et al. 2014), to make high-quality recordings for acoustic phonetic research (De Decker and Nycz 2011), or to record and document endangered languages (Bird et al. 2014).

***Corresponding author: Nanna Haug Hilton**, Centre for Language and Cognition Groningen, University of Groningen, Oude Kijk i/t Jatstraat 26, Groningen, 9712 EK, Netherlands, E-mail: n.h.hilton@rug.nl
Adrian Leemann, Center for the Study of Language and Society, Bern, Switzerland, E-mail: adrian.leemann@csls.unibe.ch

The current collection is the first to include papers with reflections from the linguistics community about the use of smartphone technology for linguistic research. The scope of the projects that make use of smartphones for collecting language data is a broad one. They have the mode of data collection, i.e. through a phone, in common, but all present different opportunities and challenges. A majority of the studies use smartphone applications for investigating language variation and change, while the rest showcase smartphone use for clinical linguistics, for psycholinguistic experiments, and for corpus building in sociology of language and linguistic landscapes. The languages addressed and studied using these tools are also plentiful. The case studies addressed in each paper concern Mandarin, Southern Min, Hebrew, Arabic, Urdu, English, Frisian, (Swiss) German, Dutch and Luxembourgish, in detail, but many of the tools can be easily adapted to any other language context.

The studies in this special issue are the following: Gilles, Entringer, Martin and Purschke start with the presentation of the ***Schnëssen*** *App* ('schnëssen' means 'to chat, to gossip'), developed for crowdsourcing speech data of Luxembourgish and sharing insights about variation with the public. The app includes an extensive sociolinguistic questionnaire (appropriate given Luxembourg's highly multilingual backdrop) and allows different elicitation rounds to be organized by the researcher. Hasse, Bachmann, and Glaser follow with the ***Gschmöis*** *App* for Swiss German. The app is the first attempt to crowdsource morphosyntactic variation of Swiss German and consists of written and audio-related tasks, using different types of elicitation methods (multiple-choice, open questions etc.). Next, Hilton showcases the ***Stimmen*** App ('Voices' in Frisian), which aims at developing speech corpora for a lesser used languages and collect dialect knowledge from users of minority languages in the Netherlands. The Stimmen app includes a picture naming task that can be used by speakers of any language, as well as a free speech module, a dialect quiz, and an interactive map that enables users to listen to recordings. Also concerned with the Dutch speech community are Hinskens, Grondelaers and van Leeuwen who discuss the ***Sprekend Nederland*** ('Speaking Netherlands') project. This project has an app that was initiated by a broadcasting company, with the aim of countering misconceptions and prejudices towards regional accents. Next, Leemann gives an overview of four apps in his paper, all developed for German-speaking Europe (***Dialäkt Äpp, Voice Äpp, Grüezi Moin Servus***, and ***Deutschklang***). The apps were devised as a means of engaging with the public with a topic that a lay audience find interesting: dialect variation. Britain and Blaxter's contribution, next, is perhaps the most methodologically oriented. They explore the degree to which an unsupervised dialect survey of the BBC shows similar regional patterns in its data as regional patterns that have been retrieved with the more supervised ***English Dialects App*** for smartphones. Gaiser and Matras follow with a demonstration of how linguistic landscapes can be captured with their ***LinguaSnapp App***. The app enables users to take pictures and add analytical descriptors, and the paper highlight the opportunities for smartphone applications as a teaching tool in higher education. Purschke's contribution also presents an app for linguistic landscaping: ***Lingscape***. Positioned as a project of citizen science, participants of the app are able to upload pictures of signage, view those pictures on maps, and create their own annotation categories. Following on from that, Miley, Schaeffler, Beck, Eichner and Jannetts present a smartphone app, ***Fitvoice***, for the collection of phonetic data in clinical settings. The app enables longitudinal monitoring of voices, e.g. voice monitoring in clinical depression. Finally, Chen and Myers contribution presents an online – albeit responsive on mobile devices – web-crowdsourcing platform called ***Worldlikeness***. The app enables psycholinguists to collect data on typologically diverse languages, store the data, and make that data available for other linguists as a means of building up cross-linguistic databases.

As showcased by the papers collected here smartphones have great potential for monitoring linguistic behavior in real-time: they allow for large scale and rapid collection of written text, speech, or imagery alike, and the devices are powerful, computationally-speaking, and overall unobtrusive to the user. Furthermore, smartphones allow researchers to get in instant touch with respondents and can also give respondents easier access to the researcher, even at a distance. However, a number of distinctive challenges also exist for the use of smartphones in research: ethical considerations, data quality concerns, participant recruitment and retention, as well as issues surrounding data ownership and storage. We give an overview here of the concerns and challenges identified and discussed in this special issue.

2 Data quality concerns in smartphone research

A frequently mentioned concern about crowd sourced data is that it is problematic for research purposes. There are concerns that crowd-sourced contributions can be done merely in jest; that participants can contribute several times, or that participants do not read instructions, to mention just a few. Studies of the quality of crowd sourced data indicate that these worries are not entirely justified. Lind et al. (2017) find that crowd sourced analyses from paid volunteers is comparable to analyses produced by five research assistants, but that there is variability across different types of tasks, and within groups of volunteers, Especially the complexity of the task that the data comes from is central when determining validity of crowd-sourced data (Horn 2018; Shing et al. 2018). Quality of crowd sourced language data concerns primarily its reliability and ecological validity: whether the method provides results that are the same as those found with a different method, and whether the properties of the data can be said to be equal, i.e. to a comparable standard of that of sound recordings, image and video capturing techniques used in different methodologies.

The only previous consideration of comparability between outcomes of linguistic smartphone research and that using other methods is Leemann et al. (2016) who find that dialect judgements crowd sourced with the smartphone-based Dialäkt Äpp correlate to a high degree with speech samples collected with traditional dialectological methods. When it comes to properties of recordings made with smartphones De Decker and Nycz (2011) concluded early on that iPhone recordings were usable to investigate vowel spaces, and comparable in quality for this purpose to recordings made with traditional recording equipment.

In this special issue quality of the data collected is considered in some respect in all of the contributions. When it comes to the standard of the sound recordings made with smartphones, Miley et al. conclude that voice recordings made in monitored experiments, i.e. not crowd sourced, with smartphones include less environmental noise than those made with standard microphones. They further poses that smartphone recordings might even be preferable to microphone recordings for clinical purposes. Importantly, she notes how new makes of smartphones, and smartphone microphones, are manufactured all the time, and so she calls for standardized acoustic tests that can be used to evaluate sound recordings, and their comparability, further. In the studies that consider crowd sourced speech recordings, Hinskens et al., Hilton, and Leemann all note that while a small proportion of the data is unusable (5–10% estimated by Leemann) most recordings are usable for instrumental phonetic analyses, and, additionally, that crowd sourced data can serve the purpose of training material for speech recognition developments.

When it comes to the reliability of the smartphone-based method, Hasse et al. conclude that the regional distributions of morphosyntactic variants collected with their app *Gschmöis* correspond very neatly with distributions previously identified in the Syntactic Atlas of Swiss German Dialects SADS (Bucheli Berger and Glaser 2002). Chen & Meyers find that the experimental data collected with smartphones through the web experiment *Wordlikeness* is comparable to that collected using traditional experimental software such as E-prime. Britain & Blaxter consider automatically collected user-data in their paper and ask whether meta-data and reported dialect knowledge collected through a smartphone application (The English Dialects App) correspond to data collected in a large web-based survey with meta data gathered through Google Analytics. They find that the data looks very similar in the two collection modes, and argue that online surveys with automatically gathered meta data is the modern version of the Labovian rapid anonymous surveying technique (Labov 1966).

3 Participant recruitment and retention in smartphone research

Another consideration to make when using smartphones in linguistic research, be it for data collection, processing or analysis, is the sample of the population that is reached with the tool. Worldwide smartphone ownership in 2020 is primarily found in the young and well-educated demographic of the population, and

there are large differences between nation states. While 99% of adults below the age of 35 own a smartphone in South Korea, for example, the proportion that does so in Mexico is 66%. In the demographic older than 50 years of age, less than half of the population own a smartphone in countries such as Canada, Italy, Japan and Poland (Silver 2019). What is more, the figures above do not consider signal strength, and the affordability of data, which are further restrictive factors that determine use of smartphones, beyond that of mere ownership.

The representation bias is something that is discussed at length in the contributions to this special issue. Disproportionate ownership of smartphones across age groups is reflected in samples discussed in this issue: many have a relatively high proportion of young respondents, as opposed to middle aged and older respondents. A number of the studies also report a majority of female respondents (e.g. Gilles et al., Hilton and Leemann). Furthermore, Gilles et al., Purschke and Leemann point out in their contributions that respondents' intrinsic motivation to take part in scientific research is unequally distributed across populations. This is likely to further contribute to a higher proportion of respondents with higher education levels participating in research with smartphones applications.

Gamification is an approach that can be taken to create additional motivation for users to participate in, as well as continue to use, web-based experiments and scientific studies. Games with a purpose, so-called GWAPS, offer a relatively efficient way of gathering information online, if the user is sufficiently clear about their role and their mission (Lafourcade et al. 2015). Leemann finds in the review of his smartphone-based studies from the last decade that gamified tasks are by far the most successful at attracting users. In the contribution about *Lingscape* Purschke emphasizes that gamification can also have an adverse effect on scientific outcomes. He notes how a reward mechanism such as a high score or points system can interfere with the goal of the researcher to create data sets that reflect the intrinsic motivations of a language community.

Another challenge when sampling language data with smartphone applications is to reach all the members of the speech community. Hasse et al. discusses how, compared to an interview setting, the simple questionnaire in *Gschmöis* is restrictive and respondents' dialect and language learning histories are often lost or incomplete. Gilles et al. discuss representation in their sample in the *Schnëssen* application and conclude that through the relatively complicated tasks in the app, data from second language speakers and speakers of contact varieties are missing. Hilton uses a picture naming task in the *Stimmen* app, as opposed to relying on written language, to include more users with oral languages and respondents who do not want to rely on written standards. *Stimmen* aims to diminish the role of a majority language for the users and actively encourages users to make recordings in more languages than their first. Hilton reports that while recordings in as many as 36 languages is possible, only a handful of users have recorded bilingually, i.e. in more than one language.

A possible approach to reach a wider audience than that possible with smartphones is to rely on web-based, rather than smartphone-based, applications. Web applications made with smartphone browser compatibility have the value of reaching users of computers with Internet access, which generally includes more people than those with smartphone access (OECD 2020). In this special issue Britain & Blaxter, and Chen & Meyers, report results collected through web applications, and Leemann, and Hilton, both argue in their contributions that future projects should consider whether web-based applications can be used, before starting the relatively costly endeavor of creating a smartphone application. Chen & Meyers indicate that participant retention could be more difficult in projects using web applications than in projects using native smartphone applications. Only one in five participants completed the entire experiment in their *Wordlikeness* study.

However, retention of participants is a problem also in other studies presented in this issue. Purschke and Hasse et al. note how it is difficult to keep users engaged longer than an initial short period of time. Suggestions to retain participants for the studies conducted using smartphones include regularly updating the applications, creating social media communities with updates that the participants can join, or to increase the use of gamification features.

4 (Data) ownership and storage concerns in smartphone research

One of the most successful ways to draw participants' attention to smartphone-based research projects is to seek collaboration with a media partner. Britain & Blaxter, Gilles et al., Hilton, and Leemann all note in their contributions to this issue how a media partner significantly increases visibility and user-base for crowd sourcing projects. However, such collaborations can come with pitfalls, especially with regards to application design and eventual data ownership. Hinskens et al. give an overview of the *Sprekend Nederland* project, a project initiated by the Dutch public broadcaster NTR. With its immense possibilities for visibility and promotion the value of a national media partner is obvious. *Sprekend Nederland* predictably collected some 17 million answers to questions in their questionnaire and attitude study. Hinskens et al. report, however, that production goals and deadlines influence the design and timing of collaborative projects, and that the focus of the project might have resulted in more standard-like speech than that another methodology may have elicited. Possibly more problematic, however, is that a media partner who owns a project could also raise ownership concerns over the data collected.

Ownership and access to the data is a concern raised by many of the contributions here. For many of the contributions the academic institution becomes the owner of the data collected. This is the case for *LinguaSnapp*, presented in Matras & Gaiser's contribution, for example, where the University of Manchester has ownership of the data collected. Some studies share their data and procedures in open access databases. All the speech recordings collected with *Stimmen*, for example, are instantly made available through the project home page stimmen.nl. Similarly, Chen & Meyers aim with their *Wordlikeness* database is to provide a platform where researchers can consult each other's data and thus create large scale comparative analyses. Purschke's *Lingscape* project is committed to transparent practices at every level of the research project. Yet, also he raises an ownership-related concern that all smartphone-based projects must consider: namely that the programming of applications and the web site front ends are generally done by a commercial partner, such as a web-design or app-design company. The actual code used for the creation of the smartphone tools, then, is generally not for researchers to share widely.

Another, complex, data issue in smartphone research that appears to be a long-term one is application maintenance and data storage. Leemann notes how the largest cost involved in smartphone-based projects can be the regular maintenance that is needed to have them function with the regular iOS and Android updates. Hilton points out that whilst the data collected through the *Stimmen* app is open access, the costs for data storage with her home institution is only covered for the total of 10 years. Purschke further adds that there are financial costs involved with the development of the project and its implementation, but also with meeting open access demands for storage, as well as for reporting of research.[1]

5 Ethical considerations in smartphone research

The final challenges we consider in this introductory chapter are those that concern ethics. All linguistic researchers have to make ethical considerations, and smartphone researchers are none the different. Yet, the ethical considerations in smartphone-based research are somewhat distinctive. Smartphones can record two types of data about its user: there is data that a person chooses to report, through applications built specifically for the purpose of the research, and data that is tracked automatically by the device (cf. Kaufmann 2019). In Europe the General Data Protection Regulation (GDPR) (European Parliament and the Council of the European Union 2016) has set restrictions, on top of those imposed by any national legislation, on the information that may be collected, stored and shared concerning human subjects in web-based research. One of the most

1 It should be noted here that network organisations such as CLARIN and META-SHARE have resources to make long-term storage of language resources possible for many European institutions and researchers.

important guidelines is the minimization clause that asks researchers to only collect and store data that is necessary to answer their research questions. That means that in many cases, the collection and storage of the data that is automatically tracked by the device, such as the device's GPS location, is deemed as unethical, unless it is necessary information for answering the research questions at hand.

The studies in this special issue all provide some discussion of ethical considerations taken in smartphone-based research. The studies have divergent concerns, however. Investigations of linguistic landscapes have, for instance, quite specific challenges concerning image data. Matras and Gaiser discuss how they have opted for registered users for their application *LinguaSnapp*. The users undergo some training before using the application. The project depends on identifiable information, such as car registration plates and human faces, being removed from images before uploading. Furthermore, images are only shared for research purposes with other, registered, users. Purschke discusses how *Lingscape*, also a linguistic landscape tool, has opted to make its users anonymous, which again has consequences for possibilities to motivate and retain people. The approach minimizes the adjustments to data that must be done to upheld privacy, however. Finally, in the contribution by Miley et al. the difficulty in ensuring user's privacy is discussed. The use of smartphones in the recording of voices for clinical purposes can be highly problematic in that respect. The authors suggest that future work using smartphones could consider extracting acoustic parameters without storing the raw data itself, as a possible solution.

6 Where does the future take us?

On the topic of future work the opportunities for smartphone-based research are immense. With the right customization there are vast possibilities for community engagement, opportunities for mapping highly complex, or superdiverse, multilingual contexts, and for making questions concerning language part of the daily reality of smartphone users through augmented reality. We identify here some timely developments that we see in the field of smartphone linguistics.

First off, the microphones built in to smartphones already provide recordings of great quality, and cameras evolve as rapidly. The opportunities for the field of clinical linguistics to monitor the properties of voices and faces through smartphones are extensive. A timely question concerns whether there are properties of our voices that can give away signs of neurological conditions, or even, whether there are certain voice characteristics that are associated with particular viral infections.

In the last decades, a development can be seen towards a democratization of the scientific endeavor through an increase in participatory research. Citizen science means engaging communities in different, or all, steps of scientific studies, as co-creators of research questions, as collectors of data, as data analysts, or as teachers or disseminators of results (cf. Bonney et al. 2016). While smartphones are not citizen science tools as such, their many opportunities for customization means they can easily be used for collaboration between the academic and the public researcher. The opportunities for data collection is dwelled upon at large in this special issue, but there are also further opportunities for joint work in understanding the data that has been collected, processing it, and in education, or the outreach stages, of research projects.

In the future augmented reality and eventually tools that display, or even place us in, a virtual reality will become commonplace. The possibilities for generating different social situations and physical environments for users of such tools can lead to a revolution in social sciences research. Research of contextual variation of language, (virtual) conversation analysis, and the effects of cues from our surroundings on language perception will see an array of new opportunities. Virtual reality tools that are standalone, or that can be connected to smartphones and computers, may also provide a haven for research done in a remote capacity due to the restrictions to travel and close contact for fears of viruses.

References

Bird, Steven, Florian R. Hanke, Oliver Adams & Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–5.

Bonney, Rick, Tina B. Phillips, Heidi L. Ballard & Jody W. Enck. 2016. Can citizen science enhance public understanding of science? *Public Understanding of Science* 25. 2–16.

Bucheli Berger, Claudia & Elvira Glaser. 2002. The syntactic Atlas of Swiss German dialects: Empirical and methodological problems. In Sjef Barbiers, Leonie Cornips & Susanne van der Kleij (eds.), *Syntactic microvariation*, vol. II, 41–74. Amsterdam: Meertens Institute Electronic Publications in Linguistics.

De Decker, Paul & Jennifer Nycz. 2011. For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics* 17(2). 7.

De Vries, Nic J., Marelie H. Davel, Jaco Badenhorst, Willem D. Basson, Febe De Wet, Etienne Barnard & Alta De Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication* 56. 119–131.

European Parliament and the Council of the European Union. 2016. Regulation (Eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119. 1–88.

Horn, Alexander. 2018. Can the online crowd match real expert judgments? How task complexity and coder location affect the validity of crowd-coded data. *European Journal of Political Research* 58(1). 236–247.

Kaufmann, Katja. 2019. *Mobile methods: Doing migration research with the help of smartphones*. The SAGE Handbook of Media and Migration, 167–179.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Lafourcade, Mathieu, Alain Joubert & Nathalie Le Brun. 2015. *Games with a purpose (GWAPS)*. Hoboken: John Wiley & Sons.

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain & Elvira Glaser. 2016. Crowdsourcing language change with smartphone applications. *PLoS One* 11(1). e0143060.

Lind, Fabienne, Maria Gruber & Hajo G. Boomgaarden. 2017. Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures* 11(3). 191–209.

Newzoo. 2020. Top countries by smartphone users. http://newzoo.com/insights/rankings/top-countries-by-smartphone-penetration-and-users/ (accessed 6 November 2020).

OECD. 2020. Internet access. https://data.oecd.org/ict/internet-access.htm (accessed 6 November 2020).

Shing, Han-Chin, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daume, III & Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 25–36.

Silver, Laura. 2019. Smartphone ownership is growing rapidly around the world, but not always eqully. https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ (accessed 6 November 2020).