
Article:
Almbark, Rana, Hellmuth, Sam orcid.org/0000-0002-0062-904X and Brown, Georgina (2023) Working with Public Involvement Coordinators to support remote collection of high quality audio speech data. Laboratory Phonology. ISSN 1868-6354

https://doi.org/10.16995/labphon.10541

Working with Public Involvement Coordinators to support remote collection of high quality audio speech data

Rana Almbark*, Department of Language and Linguistic Science, University of York, United Kingdom, rana.alhusseinalmbark@york.ac.uk

Sam Hellmuth, Department of Language and Linguistic Science, University of York, United Kingdom, sam.hellmuth@york.ac.uk

Georgina Brown, Department of Linguistics and English Language, Lancaster University, United Kingdom, g.brown5@lancaster.ac.uk

***Corresponding author's email address:** rana.alhusseinalmbark@york.ac.uk

Word Count: 12369 (11459 excluding appendix/references)

Funding information:

"The Dialect Variation in the Levant (DiVaL) project was funded by an ESRC Impact Acceleration Account 2019: York [ES/T502066/1] sub-award to the second author."

Abstract

In 2022 we planned speech data collection with speakers of Syrian and Jordanian dialects to inform an updated Syrian Arabic dialectology in response to sustained displacement of millions of Syrians. The pandemic imposed remote data collection, but an internet-based approach also facilitated recruitment with this highly distributed speech community. Their vulnerable situation brings barriers, however, since most prospective participants have limited internet data and rarely use email. We collected self-recorded short audio files in which participants read scripted materials and described pictures. Three platforms were tested: Gorilla, Phonic and Awesome Voice Recorder (AVR, smartphone app). Gorilla/Phonic offer stimulus presentation advantages, so were piloted thoroughly, but the audio quality obtained was not suitable for phonetic analysis. AVR yields full spectrum wav files but requires participants to submit files by email, so we recruited local fieldworkers, or Public Involvement Coordinators (PIC), to support participants with recording and file submission. We asked PICs and 20% of participants about their experience of working with us, through surveys and interviews. The results confirm PIC fieldworker involvement was crucial to the success of the project which generated high quality audio data, suitable for phonetic analysis, from 134 speakers within three months (Almbark, Hellmuth, & Brown, forthcoming).

Keywords: Syrian Arabic, remote data collection, fieldworkers, audio quality

1. Introduction

The COVID-19 pandemic affected research as it affected other aspects of life. The various restrictions that came with the pandemic forced researchers in linguistics, as well as in other fields, to design and conduct research remotely (Tiersma, et al., 2022). This need led to a revolution of innovative methods in remote field work. Typically, remote methods require online data collection via the internet, often using specifically designed online platforms. The advantages that come with remote methods can be tempting, as recruitment can be quicker and more targeted than in-person methods. Additionally, remote methods save the time, physical effort and monetary costs involved in travelling to conduct in-person data collection. Importantly, remote methods provide an opportunity for much wider participation in research, reaching underrepresented groups (De Decker & Nycz, 2011, p. 50; Tiersma, et al., 2022).

Reaching underrepresented communities does not come without challenges, however. For example, remote methods are ideally designed for the typical subjects of research in most fields, such as psychology and behavioural sciences; these are usually Western/American undergraduate students, who are equipped with financial and knowledge resources that facilitate conducting internet-based research experiments. Using this population, typically described as Western, Educated, Industrialised, Rich and Democratic (WEIRD) (Henrich, Heine, & Norenzayan, 2010), as a reference for other communities and countries is increasingly recognised to be misleading and the 'WEIRD' term itself does not capture the full range of differences between Western vs. non-Western populations (Gregdowney, 2010). Our key focus here is the fact that remote methods could be challenging to apply in communities with more limited resources compared to those Western and American undergraduates (Shepperd, 2022).

In this paper, we present a case study based on our experience in the Dialectal Variation in the Levant (DiVaL) project, in which we implemented remote data collection methods with non-Western communities, where some of the target population will have primary education level only, and many will have limited access to financial and technical resources. In DiVaL, our aim was to collect a corpus of speech recordings from up to 100 speakers each of a range of Syrian and Jordanian dialects of Arabic, to be recruited in Jordan. The outcome of testing a range of remote methods in this project confirmed many of the expected challenges that are specific to these communities, but also highlighted new issues, and the vulnerability of our target population of displaced speakers of Syrian Arabic dialects only amplified these challenges. Our solution to these challenges was to adopt a model of working with local fieldworkers, or 'Public Involvement

Coordinators (PICs)', to adopt a role name and role description used by the UK National Institute for Health and Care Research (NIHR, 2022). Our aim in this paper is: i) to explain the rationale for adopting this method of working; ii) to describe our methods in detail; and iii) to demonstrate from the results of a simple follow-up evaluation with both participants and fieldworkers, that the involvement of these local fieldworkers was pivotal to the success of our data collection enterprise.

In section 2, we briefly describe the aims and rationale of the DiVaL research project, whose data collection we discuss, before reviewing recent literature on the challenges of remote data collection in linguistic research and beyond. Section 3 describes the pilot study in which we trialled use of web-based platforms for our data collection, in particular explaining why we chose to reject their use due to audio data quality issues. Section 4 describes the approach used in our main data collection, which relied on collaboration with a team of PIC local fieldworkers. In section 5 we report the results of a follow-up survey of both participants and PICs, supported by semi-structured interviews with the three PICs, and argue for the critical role that working with PICs played in successful data collection. The paper closes with a brief discussion of the merits of this approach, including recommendations for other researchers on best practice in working with PICs in remote data collection.

2. Background to the study

2.1. Context: the DiVaL project

Dialectological descriptions of Arabic dialects are scarce and mainly focus on describing the linguistic features of single national dialects, usually those of capital cities. Therefore, any description of the various varieties within these nations is a contribution to fill this gap to enrich our knowledge of the variation within and across each nation. This is specifically relevant due to displacement of people in several regions in the Arab world such as Syria, Yemen, and Libya. It is vital to gain full or updated descriptions of the varieties of displaced people to capture these descriptions before any shifts occur due to contact with other dialects or varieties.

Typically, dialectological studies collect face-to-face data in the field where researchers tend to spend time in the local community to be familiar to the community members and to carefully recruit participants (Eckert, 2000; Cheshire, 1982). Similarly, phonetic studies involve collecting carefully designed experimental data face-to-face using high quality recording equipment and, in some cases, these recording sessions are conducted in phonetic labs. Carefully planned and

designed settings are crucial to work with phonetic data, for acoustic analyses where full spectral information is required.

The aim and purpose of the DiVaL project is to provide an updated description of local regional dialects of Syrian Arabic (northern Levantine) with regional dialects of Jordanian Arabic (southern Levantine), for comparison. The existing literature documents rich dialectal diversity even within the relatively small geographical area of Syria, based on data collection in the 1980s (Behnstedt, 1997). An updated dialectology is needed in light of the sustained displacement of millions of Syrians from the 2010s onwards.

Our original target was ambitious: to create a corpus of speech samples from a large number of Levantine speakers (up to 100 Syrians and 100 Jordanians) covering a wide range of places of origin in Syria and Jordan. A large number of speakers was key to generating representative linguistic maps for comparison to prior dialectological atlases. The planned stimuli consisted of scripted tasks (reading sentences and reading a folk story twice), a semi-spontaneous task (translation of sentences from Standard Arabic into their vernacular dialect) and a spontaneous speech task (three picture descriptions). These tasks were designed to generate speech data which can be used to examine a range of linguistic features (phonological, grammatical and lexical) in Syrian and Jordanian dialects.

In addition, it is important to note that our aim from the start was to generate a corpus for open access publication, and specifically with the UK Data Service, which strongly recommends that audio data for deposit be acquired in a lossless format. Our aspiration, in any case, was that the data that we obtained, at this pivotal point in the evolution of these dialects in diaspora, should be suitable for all range of types of linguistic analysis, including acoustic phonetic analysis. In particular, recent comparative work on Arabic dialects points to fine-grained dialectal differences in the acoustic properties of fricatives (Brown & Hellmuth, 2022), which can only be reliably detected in lossless audio data with full spectral information.

To capture descriptions of a wide range of dialects within Syria and Jordan, we needed to collect speech recordings from a large number of participants with a representative number of speakers from each dialect. Due to the COVID-19 pandemic, data collection was planned from the outset to be collected remotely, with Syrian and Jordanian participants who are resident in Jordan. Accordingly, we explored existing remote data collection methods to determine the best approach to use, for these non-Western communities, where we expect to have some participants with only

primary education level, and many participants with limited resources, e.g. having no internet or mobile phones.

2.2. Remote data collection in linguistics and beyond

Using remote data collection methods requires basic to advanced technology and literacy skills. Additionally, several methods require internet data to participate. Such remote methods can be manageable in communities where people own the technology and have good literacy and technology skills (i.e. can use smart devices and applications). However, Hilton and Leemann (2021, p. 4) highlight the challenge of owning smartphones and the ability to use them by different communities. For example, adults in communities such as in Mexico, own significantly fewer smartphones than in South Korea. They also highlight the variation in internet strength, cost, and availability between these communities. However, in communities where these skills and advances are minimal or do not exist, then remote methods can be very challenging. Thus, the lack of technology literacy and the lack of access to smart devices and internet may restrict the participation of such communities using remote methods. Accordingly, adaptation is required for the remote methods to work with populations with limited access to technology, and limited technology and literacy skills, to avoid excluding communities with low-income, for example.

The quality of speech audio recordings is another challenge that comes with remote methods for phonetic data, which depends entirely on the specifications of the smart devices and the sharing methods/settings used. De Decker & Nycz (2011, p. 50) examined the effect of four devices on vowel space measurements: “a Roland Edirol R-09 (WAV format) recorder, an Apple iPhone (lossless Apple m4a), a Macbook Pro running Praat 5.1 (WAV) and a Mino Flip video camera (AVI converted to AIFF). The Mino Flip file was then uploaded to YouTube and subsequently downloaded (MP3)”. Their main finding was that different devices altered the vowel space by “lowering along the F1 dimension and a widening of the space along the F2 dimension”. They concluded that recordings done by Macbook Pro and iPhone are sufficient for vowel analysis. However, the quality of the recordings from Mino and its YouTube download are not suitable for vowel analysis. Sanker et al. (2021) is another example of a work that compares the effects of different hardware (and software) on speech analysis. They similarly demonstrate variation in the resulting vowel spaces (among other speech features such as Centre of Gravity, segmental duration and F0) from the recorded speech signal using a range of tools. With a more targeted focus on formant measurements in remote data collection settings, Zhang et al. (2021)

demonstrated that recordings obtained via Zoom are likely to lead to less accurate formant measurements (in comparison to Awesome Voice Recorder (Newline Ltd, 2020)).

Smartphone applications have also been developed and used for remote data collection including text, speech, and imagery data. Hilton and Leemann (2021) reviewed a collection of papers/projects that used smartphones to collect their data in different subfields of linguistics. Like online platforms, the quality of the data collected may not be comparable to that collected using traditional methods, as some smartphone apps apply audio compression, and thus their data cannot be used for all types of phonetic analysis. Furthermore, smartphones are mainly used by young to middle-aged sections of the population, and more by the educated. Thus, a smartphone app-based approach introduces some risk of restricting participation to these groups and thereby potentially excluding older members of the target population and/or those with limited or no education.

Hilton and Leemann (2021, p. 4) report difficulties with participant retention, particularly in web-based applications, where participants do not complete the research tasks. Leemann, Jeszenszky, Steiner, Studerus, & Messerli (2020, p. 1) suggest a supervised remote data collection approach using a smartphone application recording session which is supervised by a researcher via Zoom (video conferencing tool). However, supervised sessions can be impractical, even when done remotely, as they are time consuming for the researcher, especially when a large number of participants is needed for the study. Also, these sessions require careful planning and timing between the researcher and the participant, so can be less flexible compared to unsupervised sessions. Another example of a supervised recording session approach is suggested using Cleanfeed (Hills & Bakos, 2022), which is an online studio for live audio and recording production. Although Cleanfeed produces high quality audio recordings in WAV format, which is suitable for phonetic analysis, the Cleanfeed recording session set up is not straightforward: an interview invitation has to be initiated by the researcher and shared with the participant using their name/code and email. Using this method requires collecting the participant's consent, background information, and email address in advance of the recording session. For data collection that requires presentation of text and images, Cleanfeed suggests using zoom video features alongside Cleanfeed (in a supervised session), which is not practical as it requires the participant to have two devices – one to record on and one to view the stimuli – and both of these devices stream the information in real time. Overall, this makes the task demanding for the participant, and presupposes a strong

internet connection and large data capacity. Additionally, this method is time consuming for the research team as it requires a dedicated member to run and download the recording sessions.

In a review of adjustments made in the field of medical research in response to the pandemic, Tiersma et. al. (2022) suggest a list of strategies to overcome the challenges that come with quantitative and qualitative remote research, based on previous literature and their own ongoing clinical research. For example, they suggest providing training (written/video instructions, or live assistance over the phone) to staff and participants who struggle with technology. Additionally, they suggest providing synchronous real time support by a team member to administer the data collection process (2022, p. 6). Real time support brings the advantage that the researcher can ensure data collection follows the study protocols, but requires full engagement and flexibility from a dedicated team member to administer the data collection at a time that suits the participant. The authors also stress the importance of motivating participation by building rapport with participants, recommending strategies such as "communicating clearly and confidently, and providing adequate emotional and technical support" (2022, p. 7). They also suggest developing relationships with the family members of prospective participants who may struggle with technology, so that the family member can be the guide to walk them through the data collection process.

2.3. Bridging the gap

To sum up, this review of remote data collection shows its potential advantages, seen most clearly to date in Western, educated, and rich communities. However, these methods come with great challenges if they are to be used in non-Western communities: low literacy skills may lead to difficulty in dealing with technical jargon and use of email, and such communities may also struggle due to limited resources, e.g. having limited internet access and/or not owning mobile phones. The challenges to obtaining high quality audio data for input to acoustic phonetic analysis are only higher, as we shall show below.

The suggestions of Tiersma et. al. (2022), in the context of medical research during the pandemic, focus on providing hands-on support to participants to enable their participation. Inspired by this suggestion, and to enable remote data collection in our target Syrian and Jordanian communities in Jordan, we lay out here the solution we employed, which was to work with local fieldworkers. Working with a local team is by no means a radical solution, since the importance of ensuring a prominent role for local language experts in language description and documentation has long

been acknowledged (Dimmendaal, 2001). Working with local fieldworkers to support remote data collection is somewhat more novel, and we highlight here the specific value of local fieldworker support to remote data collection of high quality lossless audio data, suitable for phonetic analysis.

Our conception of the role of our local fieldworkers takes inspiration from policy and practice on public involvement in medical research. In medical research, the specific role of a Public Involvement Coordinator (PIC) is both identified and encouraged. The UK National Institute for Health and Care Research (NIHR) defines public involvement as: "(p)ublic involvement in research is research carried out 'with' or 'by' members of the public rather than 'to', 'about' or 'for' them." (NIHR, 2021). For the DiVaL project, our PICs, who are members of the target population communities, worked 'with' us to facilitate the remote data collection; to cite the NIHR definition, their role involved "liaising with and supporting members of the public during the study [...] engaging with specific communities or being a bridge between researchers and members of the public." (NIHR, 2022). Additionally, working with local PICs whose interests overlap with those of the research project and the research team, means that the PICs gain specific benefits that meet their own career needs, such as gaining experience of ethical procedures, and training in using apps and software for research purposes. These skills proved invaluable to our local PICs, and, as we shall see in section 5.2.2, they emphasised the importance of acquiring them to improve their CVs.

PICs are thus local members of the public who support research participants and work with the research team – in this case – to support remote data collection. In the next section, we describe our pilot study which explored the potential use of existing web-based platforms for remote data collection for the DiVaL project data in Jordan, the results of which led to a change of approach, and in turn to our decision to enlist the help of local fieldworker PICs.

3. Pilot study: web-based solutions

To find a platform that can be used to collect speech recording data remotely for our DiVaL project in Jordan, we surveyed the range of web-based platforms available at the time and identified two platforms that allowed participants to provide audio responses to on-screen stimuli (images or text) and where the audio responses could be recorded in wav format for later analysis. We report in this section the results and the challenges that we faced during our pilot testing of these platforms.

3.1. Pilot implementation of web-based solutions

In the first round of piloting, two remote data collection platforms were tested: Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) and Phonic (Phonic Incorporated, 2022). Both platforms are web-based but participants can use a smartphone to perform the recording tasks (Hilton, 2021). In each case, participants received a web link to the pre-designed experiment, within which they navigated through an introduction, a project information sheet, a consent form and a background questionnaire, before performing four recording tasks. Both platforms allow a choice of interface language (Arabic in our case), which meant that pilot participants with limited or no English language skills were able to participate and follow the instructions independently.

Audio recordings from participants are directly recorded and stored on the server of the respective platform, for the researchers to access and download once available. In the pilot study, we included an additional short survey at the end of the sequence of tasks, to collect feedback from participants on their experience using the web-based platforms. We tested Gorilla with a total of 5 participants (all females; two aged 26-35 years and two aged 36-45 years) and Phonic with a total of 7 participants (2 females aged 18-25 years, and 5 males, two aged 18-25 years, one aged 26-35 years and two aged 36-45 years). These tests were performed over a period of approximately six weeks (due to trialling different settings within each platform to tackle some of the issues identified below). To the best of our knowledge, all participants took part using a mobile device (smartphone). For Gorilla, we can report that all used Android devices (OS 10 or 11) and viewed the experiment via Chrome; the Phonic platform only reports device type (i.e. mobile, as opposed to desktop or tablet), and this information is missing for two participants.

3.2. Pilot study results

In the post-task survey, participants rated Gorilla as very easy to use (average score of 1 out 10, where 1 is ‘very easy’ and 10 is ‘very difficult’); Phonic was also rated as easy to use but not as easy as Gorilla (average 2.7 out of 10). Overall, both platforms were commented on by participants as smooth to navigate, interactive, providing clear presentation and requiring minimum effort.

Although Gorilla and Phonic both afford a smooth presentation experience, the participants also commented on some issues that affected their performance or even their participation. For example, Gorilla did not give the option to re-record or listen back to a recorded sample; this was

problematic when participants made an error and wanted to re-record. Phonic, on the other hand, provided both re-record and playback buttons, which participants were able to use, but the recorded files took a long time to upload to the server before moving to the next item. During this loading time, some participants either exited the task attempt thinking that it had failed and thus that they needed to restart, or they just did not want to continue as the overall task was taking too long.

Despite these issues, both platforms appeared to show great potential for our purposes, in that they each collect audio responses to stimuli presented via a robust interactive web-based platform. In addition, Phonic came with the considerable potential advantage of offering the option of creating an automatic speech recognition (ASR) derived transcript of the content of each audio file, which could save transcription time later in the project workflow.

Unfortunately, however, when checking the quality of the audio recordings, wav files obtained from both platforms showed effects of filtering or distortion of the signal at some stage in the processing workflow. Figures 1-2 show a spectral slice taken towards the mid-point of a fricative [ʃ] in wav files recorded during our pilot study, via Gorilla and Phonic, respectively. In each case, there is no spectral information above 8KHz.

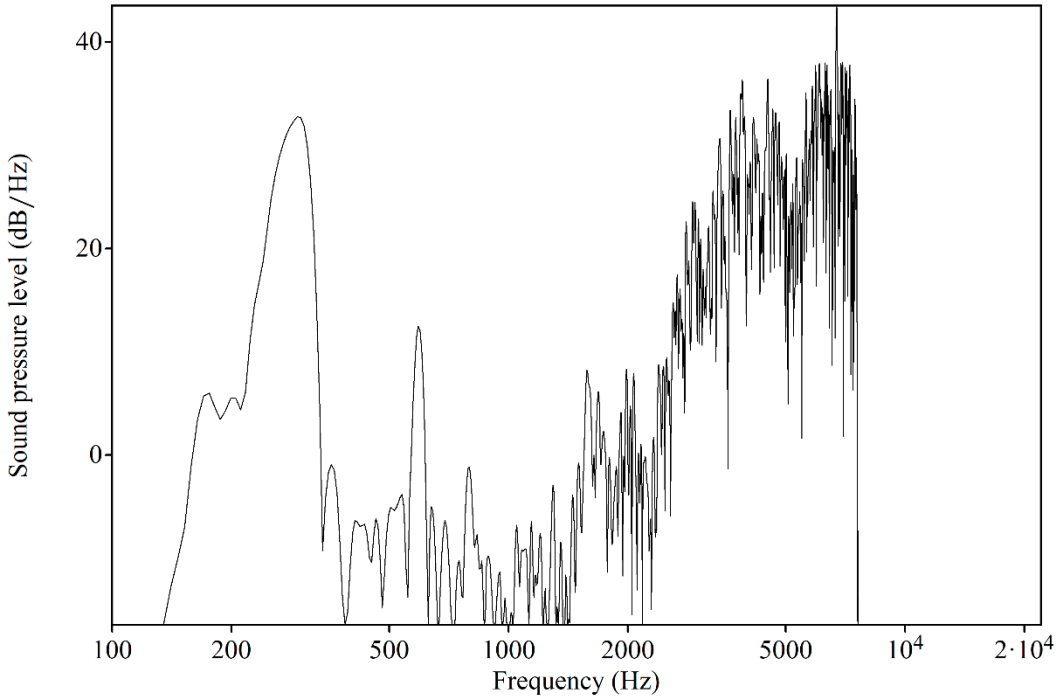

Figure 1: Spectrum taken in a fricative [ʃ] in a 44.1KHz 16bit wav file recorded during a pilot study using the Gorilla web-based platform, showing loss of spectral information above approx. 8KHz.

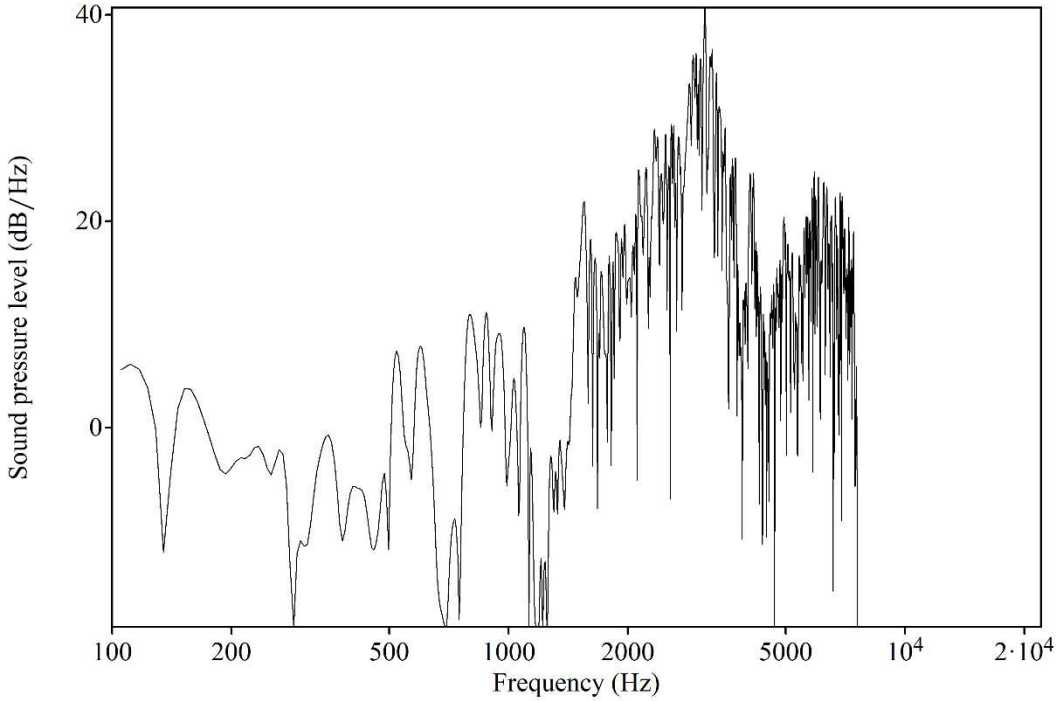

Figure 2: Spectrum taken in a fricative [ʃ] in a 44.1KHz 16bit wav file recorded during a pilot study using the Phonic web-based platform, showing loss of spectral information above approx. 8KHz.

Compression algorithms are usually applied to audio recordings to reduce their file size, which alters the spectral properties of the audio recordings. The wav files we obtained via Gorilla and Phonic in our pilot retained a sampling rate of 44.1KHz but routinely displayed this loss of full spectral information. Files without full spectral information are not suitable for acoustic phonetic analysis of sounds with high frequency information, such as fricatives, though could still be suitable for other types of analysis, such as inspection of formants in vowels (De Decker & Nycz, 2011). Our project brief was to create an open access corpus for use across a wide range of linguistic analysis types, and we were reluctant to compromise on this aim, by settling for a data collection technique which could not guarantee lossless audio recordings.

After discovering this issue, we invested a considerable amount of time exploring possible solutions to the loss of spectral information issue. We would like to acknowledge the patient and helpful technical support offered by the development teams at both Gorilla and Phonic, during this pilot phase. Nevertheless, at that time no solution to this issue could be found. We therefore reluctantly made the decision that neither platform was suitable for use in our project.

3.3. The alternative: self-recordings using a simple audio recording app

Our fallback plan was to have participants record themselves via the smartphone app Awesome Voice Recorder [AVR] (Newkline Ltd, 2020), which was emerging at the time as a viable alternative recording method (Zhao & Chodroff, 2022). AVR is an audio recording app only, however; it does not offer the option to present project information and stimuli nor to collect consent form and questionnaire responses. In an AVR-based workflow then, while AVR can be used to record the participant's audio responses, other functions must be implemented using alternative methods. To collect informed consent and participant metadata, ahead of the recordings the participant was sent a link to a Google Form which presented the project information sheet followed by the consent form, then a language background questionnaire. This link was shared with pilot study participants by email or by WhatsApp. Then, ahead of the recording session, we shared the task instructions and stimuli with pilot participants as a pdf file (again via email or WhatsApp) to be displayed during recordings on a second device (e.g. second smartphone). The audio files recorded locally using AVR were then sent by participants to the researchers by email, using the AVR file sharing menu. The workflow is summarised in Figure 3.

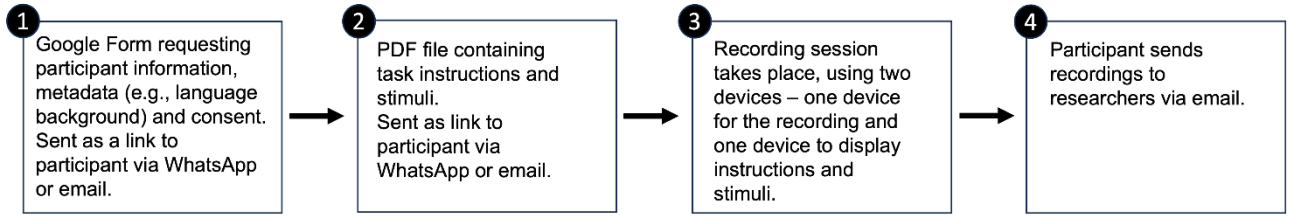

Figure 3: Summary of DiVaL data collection workflow using Awesome Voice Recorder.

We collected further pilot data via this AVR-based workflow with a subset of the pilot participants who had previously participated in trials of Gorilla and Phonic.

In this pilot phase for AVR, as for Gorilla and Phonic, pilot participants provided feedback on their experience of using the AVR app and associated workflow. The main comment was that this method required more technical involvement from the participants, rather than focusing on the stimuli and language-based tasks themselves. On the other hand, inspection of the quality of the recordings obtained via AVR showed that it provides lossless files containing full spectral information, as shown in Figure 4. Crucially, this full spectral information was only present in audio files sent to the researchers as an email attachment; any audio files sent via a WhatsApp message showed loss of spectral information. A reviewer notes that it is possible to share media files with full resolution via WhatsApp if they are sent as a Document attachment. We have trialled this method post-hoc, to simulate if it could have solved the issue. We found that this workaround requires participants to follow several additional steps before sharing the file with the research team, which we suggest means that they would be even more likely to need the support of our PICs, not less. Our post hoc test of sharing a sample .wav file recorded in AVR as a Document within WhatsApp to another WhatsApp account worked, but the file was received with the .m4a extension rather than .wav and there was still some loss of spectral information (no information beyond 16KHz). Overall, even if we had known about this option at the time of testing, we think that adopting a lossless sharing workaround via WhatsApp would have risked losing participants because of the number of additional steps it introduces to the process.

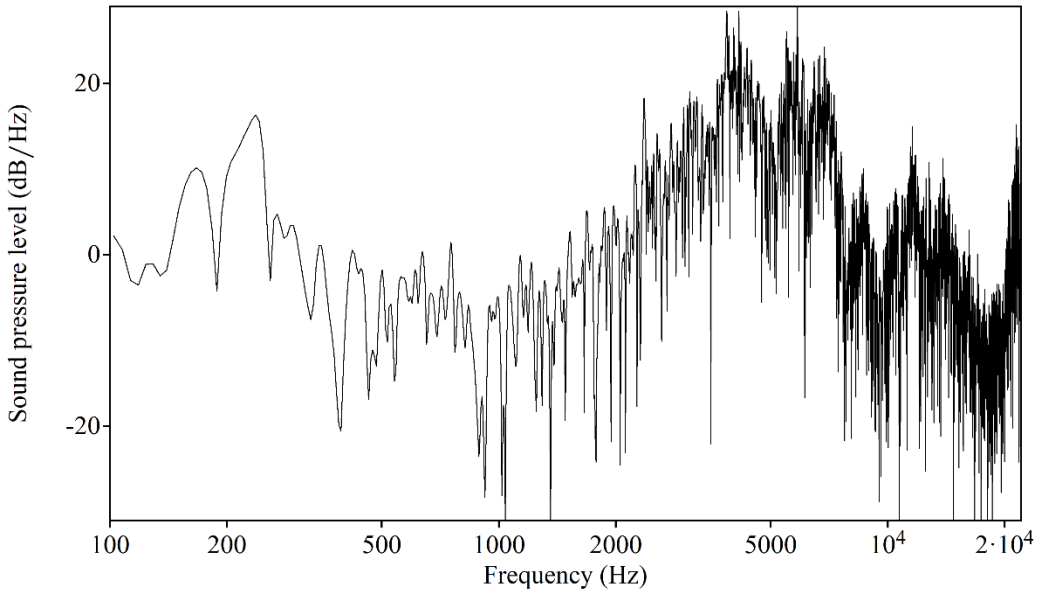

Figure 4: Spectrum taken in a fricative [ʃ] in a 44.1KHz 16bit wav file recorded during a pilot study using the AVR smartphone app and shared by email, showing full spectral information above 8KHz.

Table 1 provides a summary of the competing features of the different recording platforms and workflows that we piloted for use in DiVaL.

Table 1: Summary of key factors considered when comparing data collection platforms for DiVaL.

	Gorilla	**Phonic**	**Awesome Voice Recorder**
Possible to include participant information sheet, ethics form and questionnaire within the data collection interface?	Yes	Yes	No
Language of interface	Can change interface language to other languages	Can change interface language to other languages	English only
Possible for participants to re-record or re-listen to data items during data collection session.	No	Yes	Yes

	Gorilla	Phonic	Awesome Voice Recorder
Number of devices needed to complete the data collection session.	1	1	2
Data transfer and storage	Directly recorded and stored on app server	Directly recorded and stored on app server	Requires storage on local device and manual transfer to researchers by email
Spectral information	Limited to 8kHz	Limited to 8kHz	Full spectral information
Automatic speech recognition capability (i.e. automatic generation of transcripts)	No	Yes	No

3.4. Interim summary

The pilot phases of our project showed that – despite the many potential advantages of the two web-based platforms that we trialled – only a workflow based on participant self-recording using AVR provided a way to ensure we obtained lossless audio recordings with full spectral information intact.

A self-recording workflow via AVR carried a number of disadvantages though. The overall workflow was now rather fragmented, with different tasks to be carried out using different tools (Google Form + pdf of stimuli sent via WhatsApp + AVR to record + recording returned by email). Our target participants are in general more likely to have reduced digital literacy and/or access, so a workflow involving a sequence of different tools only enlarged the risk of discouraging target participants. Within the sequence, perhaps the most important barrier was the need for every participant to have access to their own personal email address. Shepperd (2022) explains the particular barriers to having a personal email address for low-literate Arabic-speaking participants in internet-based research, due to the fact that email addresses are obligatorily configured in the roman alphabet. In our experience, most potential participants had an email address, since an email address is usually required to set up a smartphone. This means that all participants who had a smartphone in our experience also had an email address, but many of them rarely used the email account, in practice.

We could in principle have avoided these issues by developing a bespoke web-based platform or app. A bespoke application could retain the advantageous presentation and delivery attributes of Gorilla and Phonic, but also obtain recordings with full spectral information (even if this meant that performance of the platform or app during data upload was slow). Similarly, it might now or in future be possible to work with the Gorilla, Phonic or other developers to amend the available platforms to be able to accommodate the collection of recordings with full spectral information.

In our case, we did not have sufficient time or resources to further pursue a technical solution, and both the cost and sustainability of bespoke solutions are acknowledged in the literature (Hilton & Leemann, 2021). Instead, we opted to tackle these issues by recruiting local fieldworkers to support our project participants through the different stages of data collection, as we will outline in the next section. In so doing, we found that introducing local fieldworkers to the remote data collection workflow brought advantages beyond those of the technical specifications of the recordings.

4. Our solution: working with PICs to support participant self-recording

4.1. From challenges to solutions: recruiting local fieldworker PICs

To review, a data collection workflow using the AVR recording app comes with technical advantages (ability to obtain lossless audio) but also challenges: participants need to be able to use a smartphone, download the AVR app, adjust the settings within the app, and share the recordings with the research team by email. Participants with limited or no technology literacy will find these technical details challenging. Furthermore, the app itself is a voice recorder, and it comes with no visual and presentation options, so a second device is needed to present the project information or collect any consent and background information from the participants; similarly, it does not allow presentation of text or image stimuli at the same time as collecting the audio responses/recordings.

Our decision to use an AVR-based workflow meant that only the participants who owned smartphones, and those with good technology and English skills would be able to work independently to follow any instructions provided by the research team. Thus, this decision might guarantee good quality recordings, but would come at the expense of reduced overall participation.

To facilitate recruitment of a sufficient number of participants from the target regions and dialects needed to meet our project aims, we recruited three local fieldworkers – Public Involvement Coordinators (PICs) – through the family and friends of the first author. Our expectations of the local PICs were to provide tailored support to our participants throughout the data collection process, i.e. providing internet, technology, literacy, or English language support as required (Tiersma, et al., 2022).

All PICs were recruited via local family and friends of the first author in Jordan. The criteria we used to recruit them included: PICs expressing interest in Arabic dialects, having considerable knowledge and familiarity with the different Syrian and Jordanian varieties, having a good network of friends and family to recruit a sufficient number of participants from the target varieties. We also focused on working with PICs who have good to advanced literacy and technology skills. The PICs also needed to have a mobile phone and a good internet connection.

We recruited PICs who resided in Jordan so that they would be able to assist or train participants as required on a case by case basis. For example, they provided language support to those with limited literacy in English as well as in Arabic; they provided technological support in explaining and filling out the Google Form to collect consent and background information, and in downloading the app, setting it up, making the recordings alongside the stimuli, and sharing the recordings with the researcher. The PICs were fully briefed about the objectives of the project and specifically of the data collection, so they could provide potential participants with full and accurate information. Initially, the PICs themselves took part as participants in order to familiarize themselves with the procedures involved in the data collection process. Then, further training provided PICs with a list of inclusion criteria to use to identify potential participants. Our work with the local PICs facilitated recruitment of both suitable and a sufficient number of participants, since they were able to recruit participants known to them directly or via networking, which means that the participants' linguistic background is confirmed. The PICs were also trained in how to run the various elicitation tasks in the AVR-based workflow. This training equipped them with technical skills in using the various web-based tools, such as downloading the AVR app, setting it up, and using emails to share files, so as to enable them to provide participants with one-to-one support in performing the tasks on their own devices.

As in the AVR pilot, we used a Google Form to present project information and collect both informed consent and participant background information. Our PICs shared the Google Form link

with the participant via WhatsApp in advance of doing the recordings. Then, the stimuli and written task instructions were sent in a pdf file to the participants via WhatsApp. The stimuli were to be displayed on a second device while recording the stimuli using AVR on a first device. We opted to use two devices after informal conversations with potential participants in our target communities (Syrian and Jordanian) about the likelihood of there being more than one smartphone or device owned by a single person or within a single household. The feedback was that most households have more than one smartphone, so it was possible to assume availability in our target communities of two devices for use during data collection.

To support the participants in downloading and using the app, our PICs shared a video with them via WhatsApp showing full instructions on how to download the app on their phone, adjust the settings of the app to the needs of the researchers (mainly changing two settings: file format to WAV, and channel to mono), record (including pause and stop options), and share the recordings with the research team via email. Using AVR required considerable technical involvement on the participant's side and thus technical support from the PICs was essential. For example, some participants struggled in downloading, setting up or using the app, so our PICs guided them step by step either over the phone or in person. This was caused in most cases by the language barrier; the AVR interface is provided in English only with no option for other languages, which was a barrier for participants with limited or no English knowledge.

Similarly, as noted in 3.3 above, a critical step was to share the recorded data with the research team via email within AVR (to avoid file compression). Our PICs shared the project email address with participants in advance, which they were asked to copy and paste in the 'To:' field in the email sharing screen within AVR on completion of recordings. Several participants needed step by step guidance to share the files by email, which was given verbally by our local PICs over the phone.

Finally, as well as AVR placing additional demands on participants, this approach was also demanding on the researcher side, as it requires accurate manual file naming and file management to keep track of recordings and to link them to participant metadata responses via the Google Form. Checks on correct file naming at the point of upload were therefore critical, and this was another area where the PICs provided invaluable support.

4.2. Results: outcome of data collection

With the described methods and adjustments to enable remote data collection methods using AVR, we collected recordings from 134 participants: 52 (21 males and 31 females) Jordanians and 82 (33 males and 49 females) Syrians. The total number of participants successfully recruited and recorded is less than our (ambitious) initial aim of up to 200 speakers in total, but this is not surprising considering the challenges during data collection using AVR described above. Each participant was asked to record seven files (yielding a potential total of 938 files) as listed in Table 2. The actual recording time took around 8 to 10 minutes, but with the addition of filling in the ethical consent form and background questionnaire, as well as the time needed to work with the PIC to explain the process as described in Figure 3 above, each participant took half an hour to an hour. These seven long recordings were then shared with the research team by email. The long audio files were segmented into short files within each task (up to a potential total of 6968) as shown also in Table 2.

Table 2: DiVaL corpus audio files processed for each participant, by task type.

#	*Recording task*	*code*	*Coding of processed short files*	*Total per task*
1	Grammatical sentences	gs	gs01 … gs03	3
2	Reading sentences	rs	rs01 … rs10	10
3	Story repetition 1	st-rep 1	st01-rep1 … st18-rep1	18
4	Story repetition 2	st-rep 2	st01-rep2 … st18-rep2	18
5	Picture description 1	pd1	pd01	1
6	Picture description 2	pd2	pd02	1
7	Picture description 3	pd3	pd03	1
	Total			52

The DiVaL corpus encompasses speakers from multiple localities in both Jordan and Syria, as shown in the map in Figure 5, which visualises each speaker's reported place of origin. For the purposes of analysis, each speaker is also coded according to a set of linguistically relevant

subgroups within each country. Table 3 shows the split of speakers by dialect subgroup and gender.

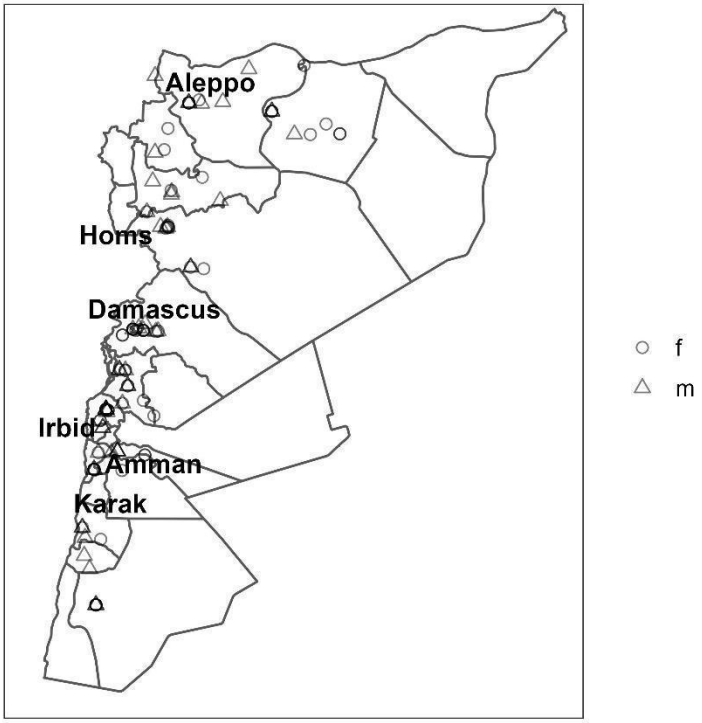

Figure 5: Reported places of origin of DiVaL corpus participants

Table 3: Localities and subgroups in DiVaL corpus

Subgroup	*Grouped localities*	*F*	*M*	*Total*
sy-neast	Al-Raqqa/Al-Hasakah	10	4	14
sy-nwest	Aleppo/Idlib	9	7	16
sy-homha	Homs/Hamah	10	10	20
sy-damas	Damascus + surroundings	12	6	18
sy-south	Daraa	8	6	14
jo-rural	Irbid/Al-Ramtha/Ajloun	9	7	16
jo-urban	Amman/Al-Salt/Al-Zarqa/Al-Azraq	14	7	21
jo-south	Karak/Al-Tafilah/Ma'an	8	7	15
Total		79	54	134

Table 4: Number of DiVaL corpus participants by age group, level of education and nationality

Age group	*Jordanians*	*Syrians*	*Total*
18-25	35	8	43
26-35	13	35	48
36-45	4	28	32
46-55	0	10	10
56-65	0	1	1
Total	52	82	134

Education	*Jordanians*	*Syrians*	*Total*
Primary	2	14	16
Intermediate	6	24	30
Secondary	7	21	28
Graduate	31	21	52
Postgraduate	6	2	8
Total	52	82	134

Data were collected from speakers in age groups from 18-65 years at time of recording. Table 4 presents the number of Syrian and Jordanian participants in each age range and by their reported highest level of education. The majority of participants were aged between 18-45; there is just one participant in the oldest age group (56-65 years). Most Jordanians reported a graduate level of education (60%), whereas the Syrians' level of education was more distributed. Some participants, particularly Syrians (17%), had primary level education only, which means their literacy skills are likely to be very limited; we can thus infer they are likely to have drawn heavily on the support provided by fieldworkers. Furthermore, this reliance on the support provided by fieldworkers is likely to hold also of the participants with intermediate and secondary education levels: even at intermediate education level, some participants may still struggle with literacy in both Arabic and English in general, and some participants with secondary level education may struggle with rapid decoding of English text, which is needed for independent use of email and the AVR app itself.

The data were collected over a three month period between June and August 2022. A total of 6885 files were processed for inclusion in the corpus. Overall, there were 2% missing files in the Jordanian data and 0.7% in the Syrian data, with just 2.7% missing files in total, as shown in Table 5.

Table 5: DiVaL corpus, missing files for Jordanian and Syrian data.

	Jordanian data			*Syrian data*		
Task	*Expected*	*Actual*	*Missing*	*Expected*	*Actual*	*Missing*

gs	156	156	0	246	246	0
rs	520	509	11	820	813	7
st-rep 1	936	934	2	1476	1476	0
st-rep 2	936	899	37	1476	1454	22
pd1	52	51	1	82	82	0
pd2	52	51	1	82	81	1
pd3	52	51	1	82	82	0
Total	2704	2651	53	4264	4234	30

The unscripted portions of the data are currently undergoing manual transcription and the full corpus (audio files plus accompanying transcripts) will be available for researchers on an open access basis via the UK Data Service shortly (Almbark, Hellmuth, & Brown, forthcoming). Acoustic analysis of scripted portions of the corpus, focussing on vowel formant properties, confirms the suitability of the corpus for its original purpose, of facilitating comparison with prior dialectal descriptions (Hellmuth, Almbark, Lucas, & Brown, 2023).

In the next section we turn to evaluation of the research data collection experience from the perspective of both participants and local fieldworker PICs, and in particular to gauging the contribution of the PICs to the success of the data collection enterprise.

5. Evaluation

5.1. Methods

We invited all DiVaL participants to evaluate their experience working with local PICs by responding to a short web-based survey composed of 11 Likert scale or ranked answer questions, and three open questions, presented in Arabic; English translations of the questions are provided in the Appendix. The survey was expected to take 10-15 minutes to complete and was implemented using Google Forms, which is easy to use on a smartphone. The questions covered four main topics: their motivation to participate, their knowledge of technology and of working with the fieldworker PICs, the project information provided, and their experience with the AVR smartphone app. Participants who completed the survey were entered into a prize draw as incentive for participation. As with the main data collection itself, the PICs helped us in promoting the follow-up survey; they shared the survey link with participants via WhatsApp and explained

the aim of the survey; they also helped with filling in the survey, but only when necessary to minimize influence on the results. The survey results were anonymous and no identifying information was collected.

The working methods chosen for this project were also evaluated from the viewpoint of our three local PICs and the research team. First, our PICs filled out a Google Form survey similar to the participants', but with additional questions about whether payment was a motivation to take on the role, and eliciting suggestions to increase and ease participation in future research. The PICs were also asked about the clarity of the instructions they were given to pass to and work with the participants. Additionally, they were asked open questions to describe how they identified and contacted the target participants and how they checked their linguistic background. The PICs were also asked to describe any challenges they faced and how they dealt with them. The same topics covered in the survey were explored in more detail and with examples in a semi structured interview, which was conducted via Zoom with each PIC. The interviews lasted around 30 minutes and were audio recorded to assist later analysis. The Google Form surveys were designed and presented in Arabic script, and the interviews were conducted in Arabic by the first author.

5.2. Results

5.2.1. The participant perspective

In total, 27 of the participants responded to our follow-up survey, which represents 20% of the total number of participants who took part in the project and provided speech recordings. While 20% is a good response rate, a caveat is that, in the interest of keeping the survey short, we did not re-request demographic information from follow-up survey respondents. In interpreting the survey results, therefore, we acknowledge that the subset of participants who responded likely fall in the younger end of the overall age range of participants and/or may be 'self-selecting' in that they are more willing to engage with a further online survey, or held generally more positive attitudes to the project and their participation in it, and thus opted to take part. A summary of the participant survey responses is provided in Table 6.

Table 6: Mean and standard deviation of ratings from respondents to the Participant Follow-up Survey (N = 27) on a Likert scale (1 = strongly disagree; 5 = strongly agree), plus average ranking of the three sources of project information (ranked from least useful to most useful; scored 1-3).

	Survey questions	***Mean***	***SD***

Motivation	Advancing knowledge	4.6	0.7
	Sharing data online	4.3	0.8
	Working with a British University	4.5	0.9
	Working with university researchers	4.6	0.9
	Working with trusted acquaintances [the PICs]	4.8	0.5
Information	DiVaL website provided sufficient information	4.4	0.7
	Ranking of the three information sources:		
	1. Website videos	1.96	1.0
	2. Participant Information Sheet (via Google Form)	1.74	1.1
	3. Website text [average ranking out of three]	1.67	1.0
Technology and working with PICs	Own knowledge of technology	4.4	0.8
	How easy was it to find and download the AVR app?	4.2	1.1
	How easy was it to set up the app?	4.3	0.8
	How easy was it to share the data from the app by email?	4.2	1.0
	I would have been able to record the data and share it without the help of the field assistant.	3.0	1.5
	Positive experience working with PICs	4.8	0.5

The survey results show that the participants found all of the suggested reasons for participation highly motivating, but they rated working with local PICs highest (mean: 4.8/5, with low SD). This corresponds to the importance of familiarity and networking in field work, which is known to be a significant factor in in-person linguistic field work (Eckert, 2000; Cheshire, 1982).

The respondents were asked to rank the usefulness of the information they received about the project, from least useful to most useful, which were scored from 1 to 3. The three sources of information were: the project website text and video recorded information, respectively, and an information sheet embedded in the background questionnaire presented via a Google Form. This information sheet laid out the aims of the project, what the participants would have to do, and what will happen to the data collected. The results rank the video information (recorded by a member of the research team in Arabic and English), as slightly more informative than the Google Form information sheet and website text. However, the ranking of all three information platforms was similar, and overall we infer that the participants received sufficient briefing from our PICs during recruitment. We infer from this also a general preference for receiving information from a person (that is, the PICs).

This subset of participants reported having a very good level of technology skills, with knowledge of using internet, phones, phone apps, and sharing files within apps. This is to be expected in the younger population in general, and is likely to be true for the majority of DiVaL participants, whose age ranges from 18-45. In addition, the participants rated use of the AVR app in particular highly positively (mean: 4.2; SD 0.9, across the three questions: downloading, set up, file-sharing), which suggests that the respondents' struggled with the technology only to a limited extent.

Despite this high self-reported ability with technology, and positive experiences of working with the specific software that we used, the survey respondents reported much lower confidence – on average – in their hypothetical ability to have made the recordings and shared the files if working independently, without the support of the PICs (mean: 3.0; SD 1.5). To unpack this lower reported confidence on average, a histogram of all responses to this question is provided in Figure 6. The survey respondents are in fact somewhat divided on this question: 22% (6/27) are highly confident that they would have been able to participate in the project without the input of a local fieldworker, but another 26% (7/27) are confident that they would <u>not</u> have been able to participate without local fieldworker help; interestingly, an equally large group (26%; 7/27) are unsure whether they would have been able to take part or not without local assistance.

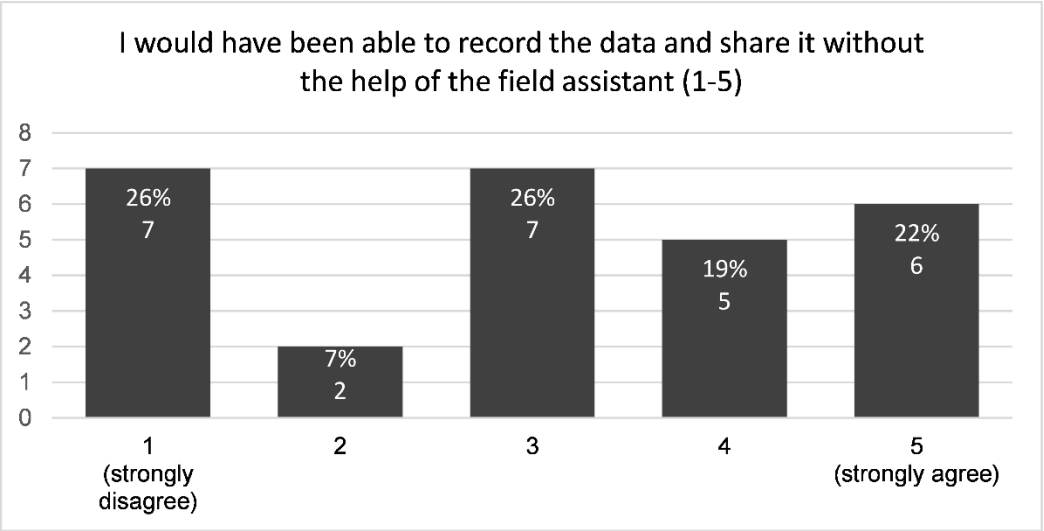

Figure 6: Histogram of responses to Participant Follow-Up Survey question about PIC assistance.

We interpret these results as strong evidence of the positive effect of working with local PICs for this type of research. To demonstrate this, let us extrapolate from this spread of responses and scale them up to our overall corpus size, to estimate how many participants we might hypothetically have been able to recruit if we had not made the decision to work with local PICs. In that hypothetical scenario, we might have lost up to a third of our eventual total of 134 participants (N = 45), since 33% of respondents in this follow up survey disagreed with the statement that they could have participated without PIC assistance. We might be reassured that we would still have obtained a good sized sample i.e. the remaining 67%. However, we suggest that we might also have lost a further quarter of our eventual total (N = 35), corresponding to the 26% of respondents here who were unsure about whether or not they would have been able to participate without assistance; this uncertainty might well be enough to discourage someone from participating in a purely online survey, however high their intrinsic motivation to do so. Crucially, all of the follow-up respondents reported that their experience of working with the PICs was positive (mean 4.8, SD 0.5); this still holds of the subset of 11 respondents who estimated they might have been able to participate without their help (mean 4.7, SD 0.6), so there is no negative impact of availability of PIC assistance for those who do not need it, or perceive that they do not.

In sum then, we interpret these survey responses as very strong evidence in support of the methodological choice to collaborate with local fieldworkers in support of remote data collection; the PIC approach facilitates participation by many individuals who might otherwise be excluded.

5.2.2. The local fieldworker 'Public Involvement Coordinator (PIC)' perspective

Table 7 presents the demographic background information of our three PICs, which shows diversity in their level of education, which was reflected in their responses. Table 8 reports responses by the PICs themselves, to the same survey questions as answered by participants, as well as some additional questions (as outlined in 5.1).

Table 7: DiVal Public Involvement Coordinators (PICs) demographic information.

PICs	*Age*	*Gender*	*Origin*	*Education*
PIC1	40	Female	Syrian	Graduate/Arabic
PIC2	25	Female	Jordanian	Postgraduate/Linguistics
PIC3	35	Male	Syrian	Secondary

Table 8: Mean and standard deviation of ratings from respondents to the Fieldworkers Follow-up Survey (N = 3), on a Likert scale (1 = strongly disagree; 5 = strongly agree), plus average ranking of the three sources of project information (ranked from least useful to most useful; scored 1-3).

	Survey questions	***Mean***	***SD***
Motivation	Advancing knowledge	5.0	0.0
	Sharing data online	4.3	1.2
	Working with a British University	5.0	0.0
	Working with university researchers	5.0	0.0
	Working with trusted acquaintances	4.3	1.2
	Compensation	3.3	1.5
Information	DiVaL website provided sufficient information	4.7	0.6
	Ranking of the three information sources:		
	1. Participant Information Sheet (via Google Form)	3.0	0.0
	2. Website videos	1.7	0.6
	3. Website text	1.3	0.6
Technology	Technology knowledge	3.7	1.2
	How easy was it to find and download the AVR app?	5.0	0.0
	How easy was it to set up the app?	4.7	0.6
	How easy was it to share the data from the app by email?	4.3	1.2
Instructions	Clear instructions about participant recruitment criteria	5.0	0.0
	Clear instructions about the Google Form used to obtain informed consent and elicit background questionnaire data.	5.0	0.0

	Clear stimuli instructions	4.7	0.6
	Clear AVR app instructions	4.7	0.6

The survey results from PICs show that they were motivated to help with the DiVaL project by all of the suggested reasons listed in the survey, but that they rated advancing knowledge of Levantine dialects (5/5) highest. We interpret this as due to their awareness of the linguistic benefits of providing updated descriptions and analysis of the target dialects, which is backed up from remarks made in the structured interviews. For example, PIC2 commented on the dialect situation in Jordan where speakers show a blend from several dialects "where Jordanians may try to speak like Syrians or Palestinians, or even Syrians try to speak like Jordanians. So, finding a pure accent has become difficult because of the dialect contact". PIC2 comments further about the importance of preserving the dialects which are in constant contact due to immigration of Syrians to Jordan; "it is very important to preserve the dialects, particularly with the immigration movement and dialect contact that happened in recent years". PIC1 suggests that this dialect contact means that: "our children, the new generation, are losing their accent completely. This is something we witness, especially in the [Syrian refugee] camp [in Jordan], where an originally Damascene person speaks Homsi, depending on their neighbours and environment". The influence of dialect contact also affects "even the adults/older generations, who are put in a position, where they have to modify their dialect to be fully understood (PIC1)".

The PICs report similarly high motivation from working with respected British universities and researchers (5/5), which is expected particularly for PIC2, who is a linguist and reported in the interview that: "I was encouraged to participate because it was a great experience for me and my career to work with a UK university and university researchers". PIC1, who is an educator, was also highly motivated to participate because: "It is also important for my CV that I worked with the University of York and university researchers". Only one of the three PICs said that the compensation to be received for the work was a motivating factor in taking on the role.

Surprisingly, the PICs' rating of their technology knowledge is on average lower than that reported by participants, which could be due to the PICs underestimating their knowledge. This interpretation is supported by the PICs' ratings for use of AVR, which show that they found it easy to work with (across the three AVR questions: mean 4.7, SD 0.7). In their interviews, all three PICs confirmed that they had no difficulty using AVR: "[The] AVR app was very easy to use, and I had no problem with setting it up (PIC3)"; "I myself found AVR very easy to work with in terms of download, settings, and sharing files (PIC2)"; "the app is generally easy to use (PIC1)". This technology knowledge enabled our PICs to provide support to the many participants who "found

AVR challenging, especially in sharing files by email (PIC2)". PIC2 reported that in order to help the participants they "used things like screen shots and video screen recordings to show them how to do the settings and sharing the files step by step". Similarly, PIC1 "explained to the participants what to do and how, using videos".

In contrast to the participants, the PICs found that the formal Participant Information Sheet, provided by the project team via the Google Form, was the most helpful information source, in comparison to the website videos and text. However, the PICs indicate that the website videos were useful "especially for people with poor reading skills, but for literate people, some found the information sheet in the google form helpful" (PIC1).

Regarding the information and training provided to the PICs themselves, the PICs found the instructions about the different elements of the data collection to be mostly very clear (across the four questions: mean 4.8, SD 0.4), but there was some room for improvement in the instructions about presentation of the stimuli and use of AVR. For example, in the interview, although PIC2 comments that the stimuli instructions were clear, they note that "some participants got confused and changed the target words, so I had to tell them to repeat the recordings without changing the original words". Similarly, PIC1 had to explain to the participants that "they only needed to adapt the sounds to their dialect without changing the original words in the text". It turned out that the goal of keeping to the same lexical items, while changing only their pronunciation, was challenging for participants, but our PICs were creative in their support. For example, PIC1 says that in order to support her participants "I sometimes had to record myself adapting the story to my own dialect, to give them an example of what needs and can be modified in the original text". Our PICs explicitly informed the participants that we are interested in the linguistic features of their (the participant's) own local accent, in order to minimise any risk of influence from the PIC's local accent.

Overall, our PICs demonstrate awareness of the fact that their role was critical in recruiting sufficient participants from the target dialects. For example, PIC2 comments that "being a local person, enabled me to focus on finding Jordanians, with confirmed Jordanian origins"; they continue: "as a local Jordanian, I was able to collect data from different cities and locations via networking, which would be challenging for a non-Jordanian person. Even though I am local, I still struggled recruiting participants from certain cities where I have no connections". We interpret these comments as further confirmation that collecting this data would have been an even bigger, and perhaps impossible, challenge, without the help of our local PICs.

Importantly, the PICs all describe how they checked and confirmed the linguistic background of the participants, so that the corpus sample is based on accurate and reliable dialect classifications. This process included using personal knowledge about participants: “when they are family or friends, I know their background. For example, one participant from [city X] was my brother’s friend who is from and lived in [city X], so I was sure about his dialect (PIC2)”. PIC2 also “used the names of the famous families of certain cities. For example, the family names of [xxx] and [yyy] are well known families in [city X]”. All our PICs also relied on their own experiences with the dialects: “I also have previous experience working with dialects, so I know the linguistic features of different Syrian dialects (PIC1)”; “I have also personal knowledge and familiarity with the different dialects in Jordan (PIC2)”; and “I used my experience with the different Jordanian and Syrian dialects to recruit participants and to check their dialects. For example, there are clear differences between the dialects in their sounds and vocabulary, which I can recognize (PIC3)”.

6. Discussion

From our perspective as researchers working on creation of the DiVaL corpus, we attribute our success in collecting the desired volume of speech data, and in a high quality audio format, almost entirely to the decision to adopt the PIC model and work with local fieldworkers.

Our initial maximum stretch target sample size was 200 speakers, and we obtained a sample of 134 speakers (67% of our stretch target). We are certain that our overall sample size would have been considerably less without the collaboration of PICs. Our PICs provided the participants with the (digital) literacy and technology support needed to perform the recording tasks; they also provided participants with required project information in different formats, and explained these whenever needed, supplementing them with creative instructions on how to perform each task as needed. In addition to the achievement of a desired participant sample size, working with PICs enabled us to collect a considerable amount of data in just three months, which is much shorter than would be possible using in-person methods. Crucially, for the aims of our research, working with local PICs enabled us to reach underrepresented groups that would have been challenging for us as researchers to find and collect data from, in any format, due to various reasons such as the conservative culture of a target sub-group, their age, or their remote places of origin or current location. Although the PICs undoubtedly boosted our data collection efforts in this context, it should be noted that the selection of PICs is likely to influence the demographic makeup of the resulting dataset. For example, PICs might be more likely to recruit and work with participants who fall into their own demographic groups (e.g. age groups). Therefore, an imbalance in PICs is

likely to lead to an imbalance in the dataset and this is likely to form part of the explanation for why our resulting dataset has a low number of older participants.

All of the web-based workflows for remote data collection that we tested required a certain level of digital literacy as well as ownership of suitable devices to be able to participate (Tiersma, et al., 2022). This was not an issue for our pilot participants who owned devices and had basic to advanced digital literacy. However, we anticipated that this would be a barrier for the main study remote data collection with some of our target participants, particularly displaced Syrians. This also means that socio-economically restricted communities, and certain sections of those communities e.g. older members, are likely to be disadvantaged. A related challenge was that participants needed to use their own internet data allowance to perform the required tasks on their smartphones; we had no budget to compensate them, and the default position of our institutional ethics board is to recommend that no financial compensation is offered to participants classified as 'vulnerable', to avoid any risk of coercion. It can be argued that this concept of risk is imposed by, typically WEIRD, academic reviewers managing institutional ethics boards; what can be considered as risky may be culturally and country specific, and in our case a financial compensation would arguably have been an appropriate contribution towards the participants' time and data allowance. Shepperd (2022) argues that ethical recommendations of this type fail to consider the specific situations of diverse groups labelled as 'vulnerable', and indeed display a Western, educated bias in assuming that people could or should participate in research solely in the interests of advancing science. Our participant survey results suggest that the advancement of knowledge about participants' own language was a motivating factor, in the absence of compensation, but we note also that the involvement of trusted acquaintances (our PICs) was even more important.

In face-to-face methods, researchers usually build or already have connections with people in the target community which makes recruiting enough participants possible. On the other hand, without these personal interactions, researchers can struggle with participant recruitment and participant retention. This can be particularly true when working with displaced or vulnerable communities. Shepperd (2022) points out the importance of cross-cultural awareness in participant recruitment. In the case of DiVaL, some prospective female participants in Syria or Jordan might be reluctant to participate because the research involves recording of their voices; and in general, people in this community may tend to avoid signing a consent form using their real name. Our awareness of these issues, and in particular the methodological choice to work

with PICs who are members of the community, helped ensure that participants had sufficient information about the project in advance, about how their data will be used and how their identity and recordings are kept anonymous. The task of explaining the aims and the settings of the project and data collection methods can be simpler and more straightforward in face-to-face data collection, compared to remote methods, as participants can easily ask, check, and clarify as the researcher is explaining the study procedures. Furthermore, the lack of personal interaction with participants could diminish researchers' control over key aspects of the research itself, such as ensuring participants' adherence to study procedures and validating the participants' background against inclusion criteria (Tiersma, et al., 2022). In our case, using local PICs, who worked directly with the participants as well as with the research team, allowed us to resolve the many challenges that come with remote data collection methods.

Overall then, our experience of working with local fieldworkers in the 'PIC' model, to support remote data collection, has been overwhelmingly positive. Moreover, we believe that working with PICs in this way could provide a way forward to improve access to populations with limited resources in data collection for fine-grained phonetic analysis. Many linguistic analyses will not require full spectrum audio data for the purposes envisaged at the outset of the research, but it is a missed chance to collect audio data that would allow the full range of phonetic analysis at a later date, while the opportunity presents itself to collect data - especially if working with speakers of remote or under-studied language varieties. The technical issues that we faced, with loss of spectral information in files recorded directly to an online platform, might well already have solutions or will be solved in the near future (for example, in an open source approach where the researcher can control the recording parameters, as outlined for different purposes by Vogt et al (2021)), but this would not remove the issue that participants would have to draw heavily on internet data allowances to take part. We therefore recommend that researchers should consider collecting data in a way that secures high quality full spectrum audio recordings, by adopting the approach outlined here, whereby remote data collection of high quality audio is facilitated through collaboration with local PICs.

Based on our experience in the DiVaL project, and on the reflections reported to us by the PICs and participants that we have worked with, we make the following recommendations for effective working with local fieldworker PICs in support of remote data collection for linguistic analysis:

1. Information: Provide local PICs with sufficient project information in various formats (video, text, website, leaflets), which they can use according to the needs of participants.
2. Instructions: Provide local PICs with detailed instructions for each element of the project workflow. Similar instructions should be prepared ready to be shared with the participants, preferably in various formats to suit the needs of different participants.
3. Implementation: Provide local PICs with the tools they need to implement the workflow, including the required technology, such as a mobile phone or tablet and an internet connection, as well as full training in use of all software and hardware, and in ethics procedures. This list can be adapted depending on the type of project and data needed.

We close by thanking our local fieldwork PIC colleagues again for their expert assistance to the DiVaL project. It was only by working with PICs, in the approach outlined here, that we were able to collect high quality audio data for fine-grained phonetic analysis with participants who often have limited digital access and/or literacy, but whose linguistic knowledge and heritage is no less worthy of recording for future generations and for ongoing detailed study.

1. Acknowledgements

We would like to thank all of our participants and, above all, our three local PICs: Sara Aljammal, Ranya Almobarak and Ahmed Almbark. We also gratefully acknowledge the support of the development teams at both Gorilla and Phonic.

2. References

Almbark, R., Hellmuth, S., & Brown, G. (forthcoming). *Dialect Variation in the Levant*. Retrieved from UKDS: https://reshare.ukdataservice.ac.uk/856484/

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-407. Doi: https://doi.org/10.3758/s13428-019-01237-x

Behnstedt, P. (1997). *Sprachatlas von Syrien.* Germany: Harrassowitz.

Brown, G., & Hellmuth, S. (2022). Computational modelling of segmental and prosodic levels of analysis for capturing variation across Arabic dialects. *Speech Communication* (141), 80-92. Doi: https://doi.org/10.1016/j.specom.2022.05.003

Cheshire, J. (1982). Variation in an English dialect: A sociolinguistic study. *Cambridge Studies in Linguistics London, 37*.

De Decker, P., & Nycz, J. (2011). For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics, 17*(2), 50.

Dimmendaal, G. J. (2001). Places and people: field sites and informants. In P. Newman, & M. Ratliff, *Linguistic fieldwork* (pp. 55-75). Cambridge University Press.

Eckert, P. (2000). *Linguistic variation as social practice.* Oxford: Blackwell.

Gregdowney. (2010, July 10th). We agree it's WEIRD, but is it WEIRD enough? Neuroanthropology Understanding the encultured brain. https://neuroanthropology.net/2010/07/10/we-agree-its-weird-but-is-it-weird-enough/

Hellmuth, S., Almbark, R., Lucas, C., & Brown, G. (2023). Vowel raising across Syria and Jordan in the DiVaL Corpus. *Proceedings of the 20th International Congress of Phonetic Sciences.* Prague.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences, 33*, 61-83. doi:10.1017/S0140525X0999152X

Hills, M., & Bakos, M. (2022). *Cleanfeed*. Retrieved July 25, 2022, from Cleanfeed: https://cleanfeed.net/

Hilton, N. H. (2021). Stimmen: A citizen science approach to minority language sociolinguistics. *Linguistics Vanguard*, 7(1). https://doi.org/10.1515/lingvan-2019-0017

Hilton, N. H., & Leemann, A. (2021). Using smartphones to collect linguistic data. *Linguistics Vanguard, 7*(1). https://doi.org/10.1515/lingvan-2020-0132

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). https://doi.org/10.1515/lingvan-2020-0061

Magistro, G. (2021). Speech prosody and remote experiments: a technical report. *arXiv preprint arXiv*. https://doi.org/10.48550/arXiv.2106.10915

Newkline Ltd. (2020). Awesome Voice Recorder [Android/iOS smartphone application]. Retrieved from: http://newkline.com/.

NIHR. (2021, April 05). *Briefing notes for researchers - public involvement in NHS, health and social care research.* Retrieved May 06, 2023, from

https://www.nihr.ac.uk/documents/briefing-notes-for-researchers-public-involvement-in-nhs-health-and-social-care-research/27371

NIHR. (2022, August 31). *Payment guidance for researchers and professionals [Version 1.3 - July 2022].* Retrieved from National Institute for Health and Research: https://www.nihr.ac.uk/documents/payment-guidance-for-researchers-and-professionals/27392#Public_involvement_coordinator_and/or_community_engagement_lead

Phonic Incorporated. (2022). Phonic [Computer Software]. Retrieved from https://www.phonic.ai/.

Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N. and Bowern, C. (2021). (Don't) try this at home! The effects of recording devices and software on phonetic analysis: Supplementary material. *Language*. 97. https://doi.org/10.1353/lan.2021.0079.

Shepperd, L. (2022). Including underrepresented language learners in SLA research: A case study and considerations for internet-based methods. *Research Methods in Applied Linguistics*, 1(3). https://doi.org/10.1016/j.rmal.2022.100031

Tiersma, K., M., R., Popok, P. J., Nelson, Z., Barry, M., Elwy, A. R., . . . Vranceanu, A. M. (2022). The Strategies for Quantitative and Qualitative Remote Data Collection: Lessons From the COVID-19 Pandemic. *JMIR formative research, 6*(4). https://doi.org/10.2196/30055

Vogt, A., Hauber, R., Kuhlen, A. K., & Rahman, R. A. (2021). Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice. *Behavior Research Methods*, 1-22. https://doi.org/10.3758/s13428-021-01686-3

Zhang, C., Jepson, K., Lohfink, G., Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *Journal of the Acoustical Society of America*. 149. 3910 - 3916. https://doi.org/10.1121/10.0005132.

Zhao, L., & Chodroff, E. (2022). The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1985-1990.

Appendix

What motivated you to take part in DiVaL?
The linguistic aims of the project such as advancing our knowledge of the Levantine dialects.
The plan to share the data online.
Working with a British university.
Working with university educators
Working with/through trusted acquaintances.
Technology and PICs
Rate your experience with technology such as mobiles and their applications.
I would have been able to record and share my recordings without the help of the local PIC.
I had a positive experience working with the local PIC.
Information

The project website provided me with sufficient information about the project. Which information we provided was more useful? The text or the video information on the project website, or the project information sheet in the background questionnaire google form? Order this information from the least to most useful.
AVR
How easy was it to find and download AVR? How easy was it to set up the settings in AVR? How easy was it to share the files in AVR? Describe your experience and challenges using the AVR smartphone app to record audio files, renaming files, and sharing the audio files with the research team