Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers

Thomas Reitmaier
Swansea University
Swansea, UK
thomas.reitmaier@swansea.ac.uk

Electra Wallington
University of Edinburgh
Edinburgh, UK
electra.wallington@ed.ac.uk

Ondřej Klejch
University of Edinburgh
Edinburgh, UK
o.klejch@ed.ac.uk

Nina Markl
University of Edinburgh
Edinburgh, UK
nina.markl@ed.ac.uk

Léa-Marie Lam-Yee-Mui
Vocapia Research and Univ.
Paris-Saclay, CNRS, LISN
Orsay, France
lamyeemui@vocapia.com

Jennifer Pearson
Swansea University
Swansea, UK
j.pearson@swansea.ac.uk

Matt Jones
Swansea University
Swansea, UK
matt.jones@swansea.ac.uk

Peter Bell
University of Edinburgh
Edinburgh, UK
peter.bell@ed.ac.uk

Simon Robinson
Swansea University
Swansea, UK
s.n.w.robinson@swansea.ac.uk

ABSTRACT

In this paper we develop approaches to automatic speech recognition (ASR) development that suit the needs and functions of under-heard language speakers. Our novel contribution to HCI is to show how community-engagement can surface key technical and social issues and opportunities for more effective speech-based systems. We introduce a bespoke toolkit of technologies and showcase how we utilised the toolkit to engage communities of under-heard language speakers; and, through that engagement process, situate key aspects of ASR development in community contexts. The toolkit consists of (1) an information appliance to facilitate spoken-data collection on topics of community interest, (2) a mobile app to create crowdsourced transcripts of collected data, and (3) demonstrator systems to showcase ASR capabilities and to feed back research results to community members. Drawing on the sensibilities we cultivated through this research, we present a series of challenges to the orthodoxy of state-of-the-art approaches to ASR development.

CCS CONCEPTS

• **Computing methodologies → Speech recognition**; • **Human-centered computing** → *Participatory design*; *Interaction techniques*; *Field studies.*

KEYWORDS

Text/speech/language, automatic speech recognition, mobile devices: phones/tablets

CHI '23, April 23 – 28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581385

ACM Reference Format:
Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Nina Markl, Léa-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2023. Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers. In *CHI '23: ACM CHI Conference on Human Factors in Computing Systems, April 23 – 28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581385

1 INTRODUCTION

Driven by advances in AI and cloud computing, large technology companies are increasingly integrating Automatic Speech Recognition (ASR) systems—those that turn speech to text—into products and services. These advances have expanded the range of languages able to access such services: Google's Cloud Speech-to-Text API, for example, currently supports 73 languages in 139 dialects and variants.[1] There are, however, many less-prominent languages that are unsupported, leaving communities unheard. Research across HCI and ASR has also expressed concern at the current direction of ASR development towards unsupervised approaches with extremely high data requirements [35, 45]. Such approaches to ASR development, even if aimed at broadening access to language technologies for smaller languages, bypass local language speakers [6] and have been criticised for getting *"caught up in an information-theoretic view of the problem"* [35]. Consider Facebook AI's 'wav2vec 2.0' ASR Framework, which is motivated by the desire to support the more than 7,000 languages spoken worldwide for which labelled training data (audio recordings annotated by transcripts) is hard to come by, but also requires 53,000 hours of unlabelled speech data [2]. In addition to bypassing communities, such unsupervised approaches with their attendant big data requirements are increasingly the exclusive provenance of large technology organisations that are able to afford the cost of training new language models (see [51]). One community of Māori speakers, for instance, have expressed a strong moral argument against 'big tech' and their approaches

[1] https://cloud.google.com/speech-to-text/docs/languages

to ASR development that cut out communities, accompanying this with a clarion call: *"we need to create better alternatives"* [12].

In this paper we present such an alternative approach, embodied in the form of a toolkit for community-engaged ASR development. The toolkit consists of a collection of hardware and software components to gather speech data in public settings on topics of community interest and to then involve community members in the manual transcription of that data. Community feedback on these activities as well as the generated speech and transcript data are further, conceptual tools that evidence how traditional ASR development pipelines and the metrics used to build/improve systems break down when confronted with diverse, multi-lingual and unconstrained language use. We leverage these conceptual tools to shed light on important future work on how language-models-in-the-making need to be modified and tuned in order to better support community needs and functions.

Finally, the toolkit also contains methodological components salient to the type of community partnerships that we advocate through this paper, namely the need to reciprocate and return to communities to demonstrate research results (see [53]) and feed back community responses from earlier data gathering.

In developing the toolkit we also respond to Steven Bird's impassioned call from within ASR research for *"researchers working on local languages to make a local turn, working from the ground up with speakers to identify new opportunities for language technologies"* [6, p.7817]. Bird prefers the term 'local languages' because it serves to remind us of *"local lifeworld and cultural area"* [6, p. 7820]. The more common 'low-' or 'under-resourced' language monikers are, in his view, barriers to understanding, and myopically focus on what a language lacks – the quantities (and forms) of data required for creating language technologies.

The term 'local languages' in its plural form also better represents the realities of many speech communities that *"have a repertoire of languages, each one playing a different role in the local linguistic ecosystems"* [6, p.7819].

Local, in the context of this paper, is Langa – an isiXhosa-speaking community we partnered with on the outskirts of Cape Town, South Africa. With the help of our toolkit we gathered 318 community-generated stories collected in public settings across Langa, totalling just under two hours of audio from voices that are marginalised and not currently reflected in existing datasets [3, 54]. The toolkit also supported five community members to comprehensively transcribe the collection of stories, which exhibit vibrancy, innovation, and mixing of language(s). The hardware and software components of the toolkit thus embody [26] the main HCI contribution of this work; namely, to address the technological barriers and work through more social/structural barriers to participation in order to engage communities of minoritised language speakers in the development of speech and language technologies suited to their ways of speaking.

Leveraging the data and insights the community engagement processes surfaced, we improved an isiXhosa ASR system from a character-error rate of 51.2 % to 27.7 %, but were also challenged to consider how we might re-imagine and further situate ASR evaluation processes and metrics in community contexts. Beyond the toolkit and spoken/transcript data, our research makes the following key contributions:

- to show how community engagement, facilitated through the toolkit, surfaces salient social and technical issues and opportunities for language technologies;
- to highlight how voice-based interactions can broaden digital participation in communities that are linguistically and economically marginalised;
- and to demonstrate the impact of speech and language technologies to Language Technology for All (LT4All) [28] and Information and Communication Technology for Development (ICT4D) [61] research communities.

2 BACKGROUND

To contextualise the paper, we begin by introducing the community we partnered with and their linguistic practices. Next, we outline related research on community engagement and speech data-collection systems. We then consider interface and platform innovations to crowd-source transcription that emphasise working with and positively impacting resource-constrained language communities. Finally, we discuss issues surrounding data that affect ASR systems specifically and AI systems more broadly.

2.1 Context and language

In this research we have collaborated with isiXhosa-speaking residents of Langa, a township on the outskirts of Cape Town, South Africa, that was established in 1927 as a result of segregation laws. The division and injustices of apartheid are still experienced in the township, whose population is predominately Black African (99 %) and isiXhosa-speaking (~92 %).[2]

isiXhosa is an indigenous Nguni language and one of eleven official languages recognised by South Africa's Constitution[3]. The Constitution rightly *"recognis[es] the historically diminished use and status of the indigenous languages"*; while most residents speak isiXhosa at home, the linguistic landscape in Langa's public spaces—advertising, business names, notices and newspapers—is dominated by English [15]. Residents of Langa draw on a linguistic repertoire of more than one language (e.g., English, Afrikaans, and/or isiZulu, amongst others). Particularly in Cape Town, English is dominant, so it functions as a vehicular language, typically used for commerce and education, which relegates isiXhosa to a more vernacular role, used for participation in the local lifeworld (see [6]). Of course, language use in general resists neat categorisation [8] and, like most urban Black South Africans, residents of Langa often switch from one language to another in their daily interactions, a process that residents refer to as 'mixing' and linguists refer to as 'codeswitching' [24].

In Langa, like in many urban contexts in Africa, the political pulse traditionally beats strongly along the urban pavement, bars, markets and taxi ranks [22]. South African media scholars have studied how such discourses are now also finding expression with and through digital media [16, 60]; however, also argue that high data costs strongly shape if and how those on low incomes are able to participate in online spaces. Platforms like WhatsApp that compress digital media (e.g., pictures, audio and videos) before sending them, or peer-to-peer protocols such as Bluetooth or WiFi Direct with

[2] https://www.statssa.gov.za/?page_id=4286&id=318

[3] https://www.gov.za/documents/constitution/chapter-1-founding-provisions

no associated data costs, are more popular than streaming services (e.g., YouTube) [16, 50]. Consequently, residents of Langa—and the specific ways they speak and mix languages—are poorly represented in online spaces. Because of this mixing and the more vernacular functions of isiXhosa, residents prefer to send voice messages rather than type out text on their mobile devices. Accordingly, previous research involving people from Langa has uncovered use-cases for ASR-driven innovations that would have an impact in contexts where media and linguistic ecologies favour asynchronous voice over text (see also [21]), for instance to find an old voice message just like an old text message—using search—or to 'listen discreetly' by reading an automatically-generated transcript rather than playing audio when out in public [45].

2.2 Data-gathering

To support ASR development, researchers have developed smartphone applications and other services designed to collect speech samples directly from users [1, 17, 33]. However, these record users as they read out a list of prompts, and are therefore not representative of how people speak in everyday life. In Langa specifically, this also does not account for how people mix languages. To more accurately capture speech in everyday life, Reitmaier et al. [45] have previously demonstrated how WhatsApp itself can be utilised as an 'unplatformed' [36] tool for collecting voice message samples from community members, leveraging users' familiarity with the messaging app and the authentic form of expression it represents – especially because it is typically used to communicate with friends, family, and community. However, such messages are generally highly personal and thus need to be handled delicately.

In Pakistan, Raza et al. have leveraged Interactive Voice Response (IVR) systems designed for communication and entertainment to collect spontaneous speech data and used that data to train an ASR system [44]. Especially when compared to the prompt-reading speech mentioned above [1, 17, 33], Raza et al.'s approach yields more natural and authentic data, and allows users to benefit from the process (e.g., communicating with others & entertainment). However, in Langa, residents normally avoid making voice calls as they are comparatively expensive.[4]

In India, Pearson et al.'s StreetWise systems compared the efficacy of smartspeakers powered by either human or machine intelligence that were deployed across public settings in Dharavi, a low-income settlement within Mumbai [42]. The project showcases the benefits of multi-sited public speech installations—residents submitted over 12,000 spoken queries across the different smartspeakers—and the authors have also released the hardware and software platform of the smartspeakers as an open-source toolkit.

Beyond ASR development, we are also inspired by case-studies and innovations that engage and gather feedback from community members. Golsteijn et al.'s VoxBox comes to mind, and the way it leverages engaging interactions (including speech) to collect opinions from people, especially in contexts where traditional surveys fall short [29]. Furthermore Xu et al. [63]'s research shows how onsite—rather than remote—methods are the most effective to survey hard-to-reach or under-served populations.

2.3 Transcription

To crowdsource transcribed audio content, HCI research has not only led to a series of mobile-friendly tools, such as Respeak [57], Recall [56], and BSpeak [58], but also demonstrated how user-friendly and accessible transcription tools can benefit and empower marginalised communities. However, these novel interfaces currently either ask the user to read out text [1, 17, 58] or require the existence of an ASR system for the language in question [56, 57].

The advantages of utterance-driven speech apps, such as BSpeak [58] and Woefzela [17], is that they alleviate the need to transcribe audio as a separate step: contributors provide an audio recording reading a predetermined transcript. However, read speech is generally slower-paced and also lacks natural intonation and prosody (changes in pitch or loudness) or coarticulation effects (where speech sounds are affected by those that precede or follow it) common to spontaneous or continuous speech. Such predetermined, utterance-driven speech apps also do not meet our goal of community engagement; that is, to gather and build upon conversations and responses to topics of community interest.

ASR-driven apps, such as Respeak and Recall, begin by breaking up longer and potentially noisy recordings into smaller chunks, then play these to the user, who is then asked to record themselves speaking what they heard in a quiet environment. A pre-existing ASR system then transcribes the recording, and the transcript is presented to the user, who can then either mark the task as complete or retry, for instance to listen again or speak more clearly to obtain a better transcript. However, high quality ASR systems are not available for many smaller languages, including isiXhosa. Furthermore, commercially available ASR systems currently do not support language mixing. That being said, we are inspired by how mobile interface innovations can be leveraged to situate the transcription tasks in marginalised community contexts and create new economic opportunities in the process. Relatedly, Reitmaier et al. have recently demonstrated some success in creating a modest ASR dataset by working with community members, but also acknowledge that they did not adequately support community-generated transcriptions [45].

Finally, web and desktop-based transcription software used by professional transcription service providers are comparatively complex and require bespoke hardware (e.g., foot pedals to control playback speed), significant transcriber training, and are targeted towards users with more advanced digital and media literacy, such as the ability to touch-type.

2.4 ASR development

Of course, context, language, data collection and transcription also figure heavily in the ASR development process. Here we draw on the ASR, HCI and sociolinguistic disciplinary perspectives of our research team to summarise and critique the current state of the art of ASR development.

Typical data-collection approaches utilised in ASR development see data as something that can be harvested or scraped online, turning data collections into a neutral, inert, and interchangeable substrate [14] which bypasses communities [12]. In our case, data-harvesting [46] (e.g., from YouTube) and crowd-sourcing through Amazon's Mechanical Turk platform [10], are not as effective, not

[4] Voice calls are more than ten times as expensive in South Africa as they are in Pakistan.

least because digital inequalities lead to poor online representation and in effect render the voices of people living in places like Langa 'digitally invisible' [60]. While state-of-the-art ASR systems [2] have been served well by building complex acoustic and language models using large data collections, languages that have less data available—so-called 'low-resource' languages in the ASR argot—are not well served by these approaches and consequently are left behind [14].

While there are existing isiXhosa ASR datasets containing read speech [3], and code-switched isiXhosa-English datasets of TV soap operas [54], voices of people from places like Langa are poorly represented therein. However, rather than dismissing these scarce language resources altogether, we are inspired by the late Gary Marsden's pragmatic design approach that emphasises leveraging the most out of existing resources [37], in our case by training and tuning an isiXhosa ASR system using both existing resources as well as data, feedback and provocations generated through engagement with a particular community. We also align ourselves with Joshi et al. [35]'s research in this regard and scholarship more broadly that advocates ground-up approaches to ASR development [6]. However, these are the exception rather than the rule. Community and context are critical differentiating factors here, for conventional approaches to ASR and language model development invest more heavily into deep learning models than data collecting or curation [46]. The danger in focusing too narrowly on the 'model-work' of ASR is that it configures the task of ASR—turning speech into text—into a mathematical problem to solve and optimise in the lab using objective metrics of error rates, and operating on existing datasets of speech data paired with 'gold-standard' human transcripts.

Another nuance that is unaccounted for, especially in the context of Langa, is that transcription and writing practices in minoritised, unstandardised and local language varieties can differ from the requirements of technologies designed for 'high-resource' standardised languages [6]. For instance, what constitutes an 'accurate' transcript might be particularly contested in language varieties which are not heavily standardised [9, 38]. Surfacing and accounting for such questions and community involvement, reciprocity, and 'data-work' more generally—undervalued and de-glamourised aspects of AI [41, 49]—are an important part of the critical alternatives to the current state of the art of ASR development.

In this paper and through our toolkit we consider the above implications through interdisciplinary collaboration and—critically—also involving and engaging community members at each step of our ASR development pipeline that is outlined in the following sections.

3 TOOLKIT OVERVIEW

The toolkit we have developed is designed to situate key aspects of the ASR development pipeline within community contexts. In this section we introduce hardware, software and conceptual components of the toolkit and describe how they were tailored to support our community engagements and research activities in Langa. While our research focuses on Langa and isiXhosa specifically, we also touch on more general traits that are likely to apply to other minoritised language contexts, or are useful beyond ASR development.

3.1 SpeechBox

To engage community members and collect data in-situ, we adapted the open-source hardware plans[5] of Pearson et al. [42]'s smartspeaker that was previously deployed in a resourced-constrained community in India. We were inspired by the way the design enabled community members with little technological experience to walk-up-and-use the system in everyday life without prior training or needing to install an app. The design is also simple and cheap to construct: within its utilitarian casing the device runs on a Raspberry Pi micro-computer with a mobile dongle for connectivity, and a speaker, microphone, small display, keypad and button for interaction (see Fig. 1).

Utilising this hardware platform the SpeechBox runs on bespoke client and sever software we developed that in essence (1) prompts users to record their perspectives on a topic of community interest, (2) gives users an opportunity to re-listen (and if necessary re-record) the story, before (3) confirming that they are happy to add their narrative to the public collection, which then uploads and stores their contribution on the server. In the next section we outline in more detail how we chose the topic, situated the precise interaction flow, and operationalised incentives and consent.

3.2 TranscriptTool

The next step in the ASR pipeline is to involve community members directly in the transcription of the community-generated data. For this task, we created the TranscriptTool, a bespoke mobile app (see Fig. 2) that is inspired by and draws on the findings of Vashistha et al. [57]'s Respeak. The key insights were: to break up longer audio segments in order to reduce cognitive load; to keep the original order of split audio segments so users can retain context while transcribing; to avoid clipping words when segmenting audio; to let multiple participants transcribe the same audio content to support normalisation; and, the ability to reject or skip over unclear tasks.

The main transcription interface attempts to focus the transcription task to its essential components, and consists of a large text area for the transcript, audio playback controls and enough space for the device's on-screen keyboard. A segmented timeline shows how the current audio recording is segmented into smaller parts of a maximum of five seconds. Audio recordings longer than five seconds are not themselves split; instead the playback of the recording pauses to minimise cognitive load and give users the opportunity to transcribe the current segment, re-listen if necessary, before moving to the next. The TranscriptTool also overlaps these segments by 750 ms to ensure that if a word is inadvertently clipped at the end of one segment it can be picked up in its entirety at the start of the next segment instead; this avoids the need to detect natural pauses that may not be possible to accurately identify if recordings are noisy. Transcripts are stored in a database on the TranscriptTool, which also logs how a transcript is built up over time—which audio segment of a longer recording was active when text was entered—in order to help capture time alignment data. The database is periodically synchronised with a server, which is also responsible for distributing audio recordings.

[5] https://github.com/reshaping-the-future/streetwise/tree/master/streetwise-hardware

3.3 Generated data

The data gathered and transcribed by community members using the above tools was furthermore an invaluable conceptual tool that surfaced and provided concrete data on more anticipated challenges, such as noisy recordings and mixed language use, as well as less anticipated challenges, such as considerable variability between transcribers of the same audio content. This suggests that isiXhosa has a lower degree of language standardisation in its written form (see [18]). Engaging with communities in public and working with less standardised languages calls into question terminologies and standards that ASR development ascribes to data, for instance of refined audio made in quiet environments and of gold-standard human transcriptions without variability (see [5]). We qualify these challenges as more or less anticipated, because even when we deliberately try to escape them we are nevertheless influenced by the models and mindsets of monolingualism and the gold-standard status of human transcriptions that are ingrained in the orthodoxy of ASR practice. So, seeing these challenges expressed and reflected in datasets helped us challenge the hegemony of such concepts and respond with new research trajectories.

3.4 ASR demonstrators

In order to feed back research results and outputs, a core tenant of community-oriented research [53], the toolkit also comprises two ASR demonstrators to ensure that community members are able to experience first hand how their voices and/or transcriptions have contributed to ASR development. This also enables them to feed back on and shape future development trajectories.

The first demonstrator was inspired by Reitmaier et al. [45]'s ASR probe, and is a simple Android application that lets users transcribe audio content that has been recorded on their phone. The second demonstrator closes the feedback loop that began with the SpeechBox and allows community members to use voice queries to search the corpus of contributions collected by the SpeechBox. Like the SpeechBox, it is embodied in the form of an information appliance, and uses similar components, but adopts a more minimal aesthetic like that of a commercially-available smart speaker (see Fig. 4). Here our ASR system is used to transcribe incoming voice queries, which are used as inputs to a search engine that is indexed with the corpus of collected stories.

4 STUDY I: FORMATIVE DESIGN AND PUBLIC ENGAGEMENT

Before leveraging the SpeechBox to collect speech data from people local to Langa on a topic of community interest, we first needed to ensure that it would be usable in context and that the chosen topic was of interest. In refining and deploying the SpeechBox, we also imagined how community organisations, political groups or local councils might gather opinions and reflections from people in Langa on important topics, whereby recorded speech could be a quicker and more convenient alternative to open-ended text responses typically used in surveys. We partnered with a community-liaison—a market researcher local to Langa who we have had a long-term relationship with—and together decided to configure the SpeechBox to collect user stories about experiences of the COVID-19 pandemic, a widely-discussed topic at the time of this research (in early 2022). That choice of topic was also inspired in part by rapid-response initiatives documenting people's experience of lockdowns elsewhere [32, 52] and in Cape Town specifically [25], though these projects themselves relied on remote data-collection methods. In subsequent conversations we also discussed appropriate incentive structures, refined isiXhosa interaction prompts and lastly organised a formative design workshop with participants from Langa. To meet our goals of reciprocity and community-engagement—central to our method—we also trialled the TranscriptTool and the mobile ASR demonstrator app at that workshop. We did this to engage participants on the overall ASR development process and showcase how ASR systems might be useful for people in Langa, in addition to obtaining more immediate feedback on the SpeechBox design and interaction flow.

4.1 SpeechBox interaction flow

Figure 1 shows the SpeechBox installed in an internet cafe in Langa. Pressing the glowing blue button starts the interaction, upon which the user is prompted (in isiXhosa) that the box is collecting stories about COVID-19, and that if the user decides to submit their story they will receive a R20 (~$1) airtime voucher, and that they should press the button again if they want to continue. If they press the button, the user is prompted that they can start their recording after a beep. After recording their story, the user is prompted to press one of the following keys on the keypad: the green button if they are ready to share their story; the red button if they want to start again; or, the number 0 if they want to listen to their story before they decide (which returns them to the same prompt after their story is played). If the user decided to share their story, they are asked to enter their phone number on the keypad in order to receive the airtime voucher. Finally, the user is thanked for their participation, reminded to expect an airtime voucher soon, and informed that they will also be sent an SMS with a link to a short optional survey for which they can receive an extra R20 airtime payment.

If no user input is received within a five second window after any prompt, the device resets itself and it is assumed the user did not want to share their story. Similarly, the recording phase has a timeout of 2 minutes, but users can also stop the recording by pressing the main button again. To avoid compromising user anonymity in case a device is stolen, our ethics review—completed for all studies in this paper—also required us to never store mobile numbers on the appliance itself, to store mobile numbers separately from the recorded stories on the server, to delete those numbers after airtime vouchers have been issued, and to delete local copies of the story after they have been uploaded to the server or if the user did not confirm they wanted to share.

4.2 Formative design and public engagement workshop

The workshop facilitator recruited 11 participants (7F, 4M) local to Langa and determined a locally appropriate participation incentive payment. Opening the workshop, we introduced ourselves and explained that we wanted to work with participants to think of ways to improve the effectiveness of speech systems in local languages, like isiXhosa. We explained that we had a initial isiXhosa ASR system (introduced later), but that we need to improve it by (1) gathering data via and improvements to the SpeechBox and (2) transcribing that data using the TranscriptTool. Finally, we wanted to (3) showcase the current status of the ASR system using the mobile demonstrator app.

4.2.1 *SpeechBox trial.*

We then introduced the SpeechBox system and demonstrated how it works. In two groups of five and six participants, respectively, we asked them to each have several turns at using the SpeechBox. Each group interacted with their own SpeechBox, and a member of the research team observed their interactions and could answer questions (if necessary). The community liaison moved between the two groups to support. In total participants recorded 37 stories (about three per person).

Although all participants were L1[6] isiXhosa speakers and recorded their stories in isiXhosa, we noted many instances of code-switching with English words. We also observed some interactional issues that related to the system not correctly recognising button presses, for instance because a prompt was still playing during which the system was not listening for button presses. Some button presses were also missed, for instance when users were pressing and holding the main button while recording their story and then releasing the button when they were finished.

Participants wanted English signage to accompanying the SpeechBox during its deployment, demonstrating how the linguistic landscape in Langa is dominated by English [15]; we had only created isiXhosa signage. Several people wanted those in power to listen to their stories, exemplified by one participant's plea: *"we've been through a lot – we are suffering, we are unemployed, we need support"*. When asked, participants imagined that younger generations (18–45) and *"people who went through the most during Covid"* would be the most likely to use the device and felt that *"not many older people go into the internet cafe"* and older generations would need someone to talk them through using the device, a process called 'intermediation' [48].

4.2.2 *TranscriptTool trial.*

Next, we turned our attention to the TranscriptTool and outlined how ASR systems are trained not just on speech data, but that they also need text data so the system can learn how sounds relate to text. We explained that transcription is typically done by experts on desktop machines, and invited participants to experiment with us to see if it could be done on their phones. We then asked participants to install the TranscriptTool app on their own phones. This was a time-consuming process, as participants struggled to complete the installation process. Some mobiles had severely cracked screens (see Fig. 2), making navigation in unfamiliar apps such as the Google Play store difficult, while other mobiles were out of battery or had run out of space. By clearing space, charging phones and demonstrating the precise steps required to install the app, nine participants (82 %) were ultimately able to install the app. The remaining two were ultimately unable to install the TranscriptTool because their mobiles were running much older and unsupported versions of Android.

We preloaded the app with nine transcription tasks. Eight were short recordings (2.2 to 5.2 s) from van der Westhuizen and Niesler [54]'s soap opera corpus, split between mixed isiXhosa and English (n = 4) and solely isiXhosa (n = 4). The ninth task was 120 s long and contained all of the audio prompts from the SpeechBox, which were recorded in isiXhosa but also contain some commonly-used English terms (e.g., 'button' or 'airtime') and even some mixed words, such as 'iCovid' (= isiXhosa noun prefix + *Covid* [24]).

Participants completed all of the tasks, but also uncovered some basic usability issues that we subsequently addressed. These mostly pertained to navigation and the ergonomics of listening and typing. Participants wanted a clearer way of completing a task—a straightforward 'done' button—as well as a better indication of *"where you are within the list"* of tasks. On this point they also suggested that the app could automatically take the transcriber to the next task after they have completed the current one. There were several suggestions to allow controlling the playback speed to make it slower and facilitate listening and writing at the same time.

Participants had no trouble switching between typing isiXhosa and typing English. After explaining that it would be useful for ASR developers to know which words are in isiXhosa and which are in English there were a range of suggestions for how this could be achieved. One participant suggested that the app could use bold/colours to indicate an English word. However, others felt that enclosing English words in brackets would be more visually appropriate mechanism.

Finally, we asked participants to imagine they had been hired to complete a series of similar transcriptions and polled—by show of hands—whether they would prefer to complete the tasks on their mobile phones or on a PC/laptop. A majority of participants (9; 82 %) chose a mobile phone as the preferred device.

4.2.3 *ASR demonstration trial.*

In the concluding part of the workshop we focused on our mobile ASR speech-to-text demonstrator app so that users could experiment with ASR capabilities and limitations on their own devices. Here only five participants (~45 %) were able to install the app without compatibility issues, as it required a minimum version of Android 7.0 (late 2016). We had chosen this configuration based on usage figures reported in the Android Studio IDE, which claims 92.7 % of devices would be supported. However, these figures are likely weighted very differently across the globe, and do not account for mobiles that remain largely offline and typically 'sideload' apps, using software such as SHAREit[7], rather than obtaining them through Google Play.

The ASR demonstrator used a baseline isiXhosa ASR 'hybrid' system. We trained the model on 50 hours of read speech from the NCHLT South African speech corpus [3]: a dataset that has previously been used in the isiXhosa ASR literature (cf. [7, 34, 55]) and used the smaller isiXhosa-English TV soap opera dataset [54] for testing and tuning. When evaluated on the soap opera data, this baseline model achieved a word error rate (WER) of 93.6 %, and a character error rate (CER) of 51.2 %[8]

In the workshop, we contextualised the metrics of the ASR system and explained that it would likely make lots of errors. In groups of two (and one group of three) we asked participants to record and transcribe messages to their group partner(s) on three different topics: organising a party together; informing each other about a community activity happening in the area; and, chatting about

[6] The first language that a person has been exposed to from birth.

[7] https://shareit.ushareit.com/

[8] See Section 6.1 & 7 for a more detailed discussion on ASR development and metrics.

work. The partner(s) would then try to gauge the content of the recording from the transcript before listening to the recording. They would then respond with their own recording and repeat the process.

One group reported that they could not make sense of any of the transcripts. The remaining groups said that the ASR recognised some things correctly and that they could get the gist of most of the messages, because they also knew the context from our instruction. Common words, like 'enkosi' ('thanks') would normally be correct, but the ASR also transcribed some English words (e.g., a shopping item, "Hennessy Whiskey" for the party scenario) as incomprehensible isiXhosa. That being said, the ASR system performed slightly better than we had expected and the metrics might have suggested. We thanked participants for their feedback on all aspects of the workshop and reiterated that their SpeechBox and TranscriptTool use and feedback would help reduce ASR errors in future.

4.2.4 Discussion: SpeechBox and TranscriptTool refinements. In response to participants' feedback during the workshop we improved how button presses are interpreted on the SpeechBox appliance, especially during the recording phase in order to accommodate both the 'press to begin and press again to stop' as well as 'press while recording and release to stop' behaviours observed. We also ensured that the system was listening for button presses even if a prompt was playing, for instance so users could start entering their phone number instead of having to listen to the entire prompt first. As recommended, we designed an English version of the SpeechBox signage, and also refined the precise language of each prompt.

We addressed issues and implemented changes to the TranscriptTool that participants uncovered and suggested: to support task navigation (e.g., automatically navigating to the subsequent task upon current task completion) and improving the ergonomics of listening (e.g., controlling playback speed). We also implemented a rejection/flagging feature so users could identify recordings that are (1) blank or contain only background noise, (2) are inappropriate, or (3) were recorded by someone sounding under-age. Finally, we added a final step after the user marked the task as completed, namely to rate the task difficulty as easy, medium or difficult.

5 STUDY II: DATA COLLECTION AND TRANSCRIPTION

In order to engage community members and collect responses on their experiences of COVID-19, Study II began with a deployment of the SpeechBox system. We then utilised the TranscriptTool to involve community members in the transcription of the community responses collected by the SpeechBox.

5.1 Data collection

We deployed the SpeechBox at two locations in Langa selected by the community liaison: a spaza shop[9] and an internet cafe. The spaza shops sells items for as little as R1 (~$0.06) and the internet cafe also provides essential printing and photocopying services (for instance to apply for jobs or access government programmes). In this instance too, the community liaison determined an appropriate amount of compensation for the shop owners for hosting the SpeechBoxes for a two-week period, as the devices would need to be charged, take up some counter space, and it was unclear if and how the SpeechBoxes would affect turnover (i.e., by drawing people in or causing disruptions).

Over a period of two weeks the SpeechBoxes recorded 209 stories at the spaza shop and 109 stories at the internet cafe, or 318 stories in total. We promptly paid out airtime incentive payments throughout the deployment. As explained earlier, the SpeechBox also sent out a link to a short, mobile-friendly online survey with the offer of a further airtime incentive payment. However, we only received 33 responses out of the 318 survey links that were distributed. At times there were issues keeping the devices operational, either because they had internet connectivity issues or shopkeepers had taken the devices offline. However, it quickly became clear that publicly-installed devices broadened digital participation and were more accessible, especially when compared to the online survey and the mobile apps we had attempted to install during earlier workshops.

5.2 Transcription

The audio of the recordings reflected the messiness and busyness of social life in Langa writ large, especially compared to the two existing corpora of South African languages: the read speech of the NCHLT corpus [3] and to a lesser extent the soap opera corpus [54], both discussed earlier. So, we leveraged the TranscriptTool to situate that critical process in the community so that the people transcribing speech data are familiar with emerging language surrounding contemporary topics (e.g., 'iCovid' or 'iLockdown') and the social use of language in place more generally [24]; that is with accents, dialects, vernaculars, and mixed language-in-use [19]. Communities also retain more sovereignty over their data [5] and can play curatorial roles to, for instance, flag content that could portray the community in a negative light.

More broadly, we were interested in studying the challenges of working with novice transcribers and, similar to previous research, explore how mobile interface innovations could extend digital participation in this new form of digital work to marginalised communities and to support marginalised languages for which there are few existing options in the market [1, 11, 56].

Turning to the task itself, we decided to remove five recordings that were less than three seconds in length, as these did not contain any meaningful content, leaving a total of 350 recordings to be transcribed: 37 from the workshop and 313 from the deployment (see Fig. 3). We listened to a random sample of recordings with the community liaison, and noticed that recordings often contained background noise or a second speaker, which we suspected would necessitate users listening multiple times in order to accurately transcribe audio. As a result, we made the decision to pay users R20 (about $1.20) per minute of transcribed audio to account for the increased difficulty and time required to transcribe.

The community liaison recruited six users to participate in the transcription study. Two of those users were also part of the earlier workshop and thus had some familiarity with the TranscriptTool. One user withdrew before the main study began, and another had to withdraw after completing two-thirds of the study, citing phone difficulties. Following best practices surrounding 'crowdsourcing', we asked the community liaison to create a WhatsApp group chat for

[9] An informal convenience shop business in South Africa

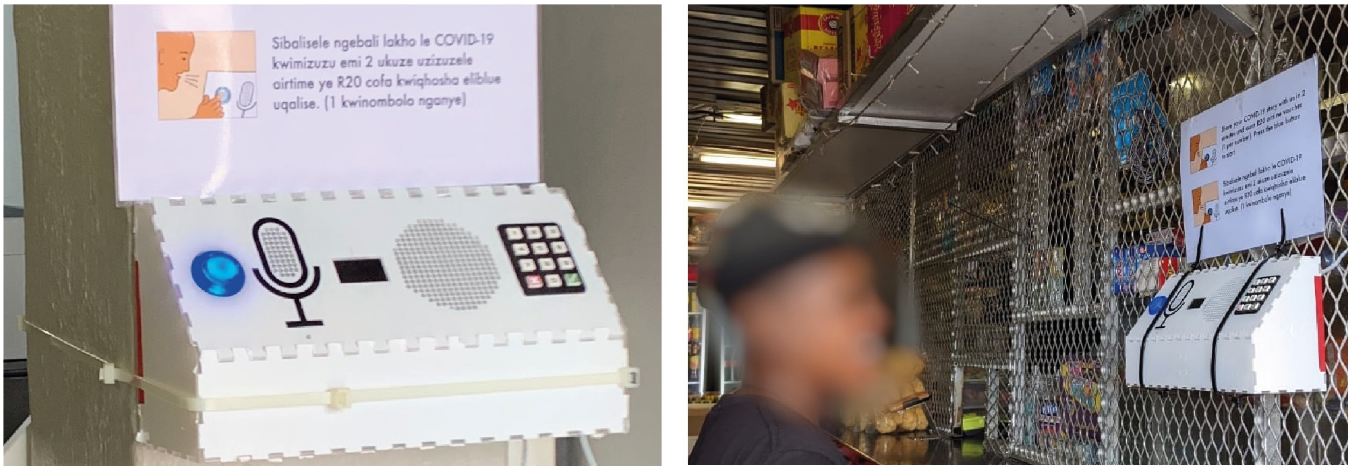
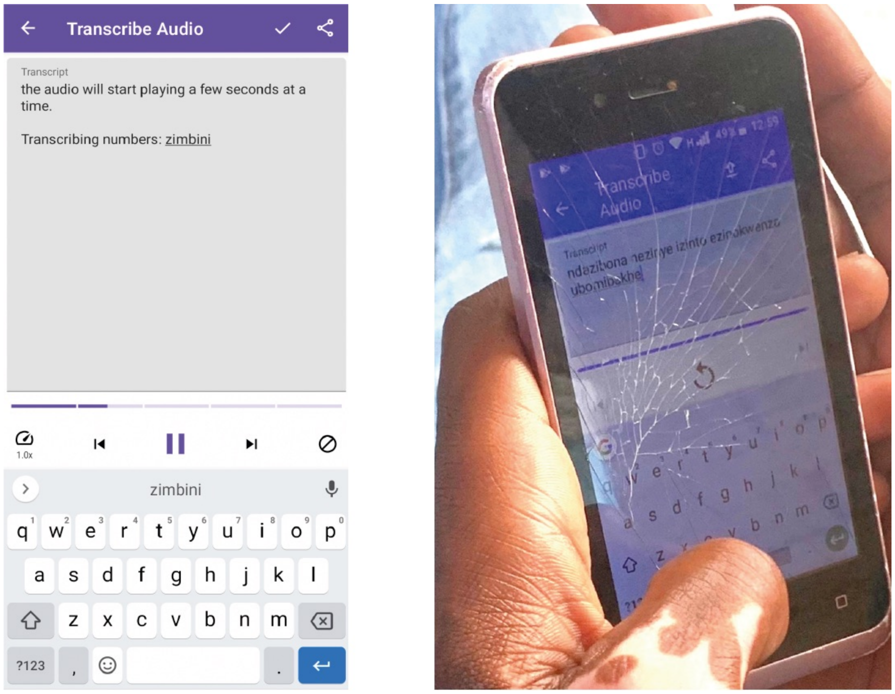

Sibalisele ngebali lakho le COVID-19 kwimizuzu emi 2 ukuze uzizuzele airtime ye R20 cofa kwiqhosha eliblue uqalise. (1 kwinombolo nganye)

Figure 1: The SpeechBox system as it was deployed in an internet cafe (left) and spaza shop (right) in Langa over a period of two weeks to gather spoken community responses on their experience of COVID-19.



the audio will start playing a few seconds at a time.

Transcribing numbers: zimbini

ndazibona nezinye izinto ezimakwenzo ubomi bokhe

Figure 2: The main TranscriptTool interface to involve community members in the transcription of collected speech responses: (Left) as a screenshot taken from the instruction video and (right) installed on a participant's phone with a cracked screen.

the participants so they could communicate with one another [30]. We also created a video tutorial—like Abraham et al. [1]—outlining how we wanted users to undertake the transcription tasks using the tool:

- Asking users to transcribe as accurately as possible,
- Transcribing all words in whatever language they were spoken, and spelling out numbers in the language they were spoken in (e.g., 'ten' vs. '10');
- Showing users how to repeat a segment, move to the next or previous segment and control playback speed;
- Reminding users that segments overlap slightly so no words are missed, but to only transcribe overlapping words once;
- Explaining how to flag content that is blank, inappropriate, or created by minors, and how to mark a task as complete and rate its difficulty

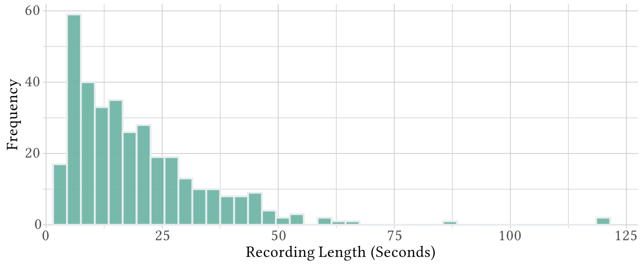


Figure 3: Recording length distribution graph, showing that users preferred to record shorter stories of around 5 to 25 s and that all recordings were shorter than the cut-off of 120 s.

- Showing how to navigate between tasks and where to find new tasks, on-going tasks (partially completed tasks), and completed/flagged tasks;
- Informing users where to direct questions/suggestions, and thanking them for their participation.

Finally, we asked users to install the app on their phones and explained that it was preloaded with sample transcription tasks, which participants could use to familiarise themselves further with the UI.

5.3 Task distribution, adaptations, and results

Our goal was that over the course of the transcription task, each audio recording would be transcribed by three users, that users would need to complete one batch of 18–20 minutes of recordings before taking on the next, and that all users would transcribe roughly the same amount of data by the end of the study.

The reader with an interest in theoretical computer science might recognise in this goal a formulation of the NP-complete knapsack problem. So, we had to use approximations and heuristics to achieve task assignments that were somewhat balanced (+- 2 minutes within each assignment batch) across the corpus of 350 recordings (114 min total, 19.6 s average, 3.1 s min., 120.1 s max.). However, we only learned about one of the withdrawn participants after assigning the first batch of tasks to users, and much later in the study had to redistribute uncompleted tasks of the second withdrawn participant to other users, some of whom had already transcribed those uncompleted tasks earlier and could therefore not transcribe them again.

Finally, as we noticed some variation in the transcripts received from users, we also created 19 new tasks containing recordings from the soap opera corpus [54], which we included as part of the second (10 additional tasks) and third (9 additional tasks) batch of tasks so we could draw comparisons to published 'gold-standard' transcripts. At the end of the study each participant had transcribed between 41 minutes and 69 minutes of audio. Most (324) received transcriptions from three users, however, some (26) were only completed by two participants.

Consider the example in Table 1 taken from the soap opera corpus [54]. Notice how in this shortened example transcriptions exhibited variability on the spelling of the first word 'kuthiwa' –"*It is said*". Some users made a spelling mistake here and dropped the 'i' and/or 'h'. The word segmentation chosen by participant 2—with an apostrophe instead of a space—in this example is also interesting. Participant 2 has clearly gone through extra steps to access the apostrophe on the symbol character map of the keyboard, as opposed to the '[space]' key, which would have been much simpler to type and illustrates that participants were completing the task diligently and comprehensively.

The variations in word spelling and word segmentation also highlight how orthography in isiXhosa is less codified. In other words, the representation of a spoken utterance in writing allows for more individual choice and variation. Participant 2's utilisation of apostrophes and elision of vowels is a great example of this: they clearly attempt to approximate the prosody/phonology of quick isiXhosa speech (i.e., the way isiXhosa words can bleed into each other). Such variations raise important questions that we flag here. Should the transcription choices like those made by participant 2 should be deemed 'errors' or 'mistakes'? After all, string-matching algorithms would mark participant 2's transcript as different from the 'gold-standard' reference transcript provided by [54]. Further, what does it mean given that this variation is so inherent and apparent throughout *all* the transcripts collected during our study? We shall pick up on this discordant relationship between transcript/orthographic variability and generally-accepted notions of 'gold-standards' again in Section 6.1 and Section 7.

To better understand how participants gained familiarity with the TranscriptTool, what strategies they utilised, and the challenges they faced, after they had completed all of the tasks we invited them to join us for a workshop to explore those topics.

5.4 Participant workshop

Only two of the four transcribers who completed the task were able to participate in the workshop, as the others were on prolonged family visits to the Eastern Cape Province, the traditional homeland of the amaXhosa (Xhosa people). As a result, we adapted the workshop to also include three inexperienced participants and an additional person who had previous experience of the TranscriptTool in Study I. The community liaison recruited the six participants and we followed similar introductions, ethics/consent, and incentive structures to previous sessions. We paired each of the three inexperienced users with the two experienced participants and the one semi-experienced participant from Study I, and asked the more experienced participants to train their partners. The most experienced transcribers showed finesse and high levels of experience with the transcription process and the user interface, and assisted novices during the process far beyond simply showcasing the interface.

We then asked four participants to transcribe additional recordings while we separately interviewed the two participants who had taken part in the main transcription study. Both had enjoyed being part of the bigger picture – creating language technologies that would benefit their community in future. They felt it was easier to transcribe clusters of stories that were around the same topic, but found it demotivating to transcribe audio recordings where people, in their view, were *"messing about"* with the device. The participants also reported limiting the amount of tasks per day and found they could transcribe between 10 and 20 recordings at a time

Table 1: Transcript variability of a sample taken from the soap-opera dataset [54].

User	WER	CER	Transcript
[54]	–	–	kuthiwa abantwana abaninzi … especially boys are more prone to
1	73.6	34.9	kuthwa abantwa abaninzi … ingakumbi amakhwenkwe babandlongondlongo
2	48.2	19.8	Kuthw'abantwan'abaninzi … especially boys are more pairing to …
3	51.5	23.5	kuthiwa abantwana abaninzi … especially boys are more p to …
4	41.3	15.7	kutwa abantwana abaninzi … especially boys I'm more prying to …
5	57.0	15.2	Kuthwa abantwana abaninzi … especially are more prone to …

before it became tedious. Another strategy was to tackle more difficult recordings (e.g., with background noise or multiple speakers) in particular places, for instance in a quiet room or with the help of family members who would listen with them to identify what was said. Particularly if the recording picked up multiple speakers, participants skilfully transcribed the whole storyline, rather than simply the loudest speaker.

There were further suggestions to improve the process and the app: participants wanted a way of adding comments to the transcription to, for instance, indicate changes in speakers. With the existing app they occasionally improvised methods to achieve this – for instance, when an exasperated person recorded a story about being tired of COVID-19, one of the participants added a transcription comment in brackets "(mama in background telling him to mention umsebenzi [jobs])" whereas the other participant focused on the main speaker. The two disagreed about the place of inappropriate language: one saw it as a part of how people speak and, given the goal of building ASR systems that reflected real isiXhosa, wanted this reflected in the dataset and so transcribed it; the other flagged any instances of profanity (as per instructions). Finally, participants thought the video tutorial with the instruction was helpful, but found it difficult to reference as they only had one device, so they could either watch the video *or* open the TranscriptTool. One participant therefore wrote down the instructions on a piece of paper. Both agreed that it would be better if the instructions could be referenced within the app.

6 STUDY III: DEMONSTRATION

In Study III we closed the feedback loop by giving community members access to the repository of stories they had created earlier using this as a way to experiment with the ASR system they helped develop in the process. For this iteration we again adopted an information appliance-based approach, but streamlined the aesthetics and interactions of the device, drawing inspiration from contemporary smart speaker design. We piloted the design in a community workshop, deployed the devices in public over three weeks, and asked community members to trial the system in context and reflect with us on the design of the system.

6.1 The ASR system

Informed by the new datasets and findings reported in the literature [18, 45] showing that speakers code-switch between many South-African languages, not just between isiXhosa and English, we decided to make the ASR system multilingual. The model was designed to recognise five languages: isiXhosa (xho), isiZulu (zul), Sesotho (sot), Setswana (tsn)—all belonging to the same Southern Bantu family of languages—as well as English (eng).

We built a hybrid ASR system using the Kaldi toolkit [43], with a model consisting of an acoustic model using a grapheme lexicon and a language model. Similar hybrid systems have been successfully used to recognise a diverse range of Indic languages, which also exhibit code-switching [20], providing evidence that training on multiple languages (especially if these languages are acoustically/linguistically similar) is more robust than training on one language alone.

Although integrating the Covid Stories data into our training pipeline would have been ideal, for reasons we unpack in Section 7, we did not use these recordings for training/testing directly and instead used the 'test-set' component of the South African soap opera (SO) dataset [54], which covers all five of the languages listed above. We reasoned that the SO dataset would be a closer match in domain and style to the type of speech the ASR system would be used to recognise during our deployment.

We initially built the acoustic model component of this system by pretraining on the NCHLT corpora [3], followed by fine-tuning on the (train-set proponent of the) South African Soap Opera (SO) dataset [54]. Although NCHLT is a large dataset that covers all five languages of interest, there is a big domain mismatch between NCHLT and SO, because NCHLT contains read speech and SO contains conversational speech from TV programmes.

To alleviate this mismatch and to better suit the model for the COVID-19 stories use-case, we instead pre-trained the acoustic model on 500 hours of speech from the the British English Multi-Genre Broadcast (MGB) corpus [4], which was also fine-tuned on the training part of SO. Note that, unlike the NCHLT corpus, the MGB dataset contains only *English* speech. Nevertheless, given its matching domain, we found pre-training on this corpus to yield better results. This model achieved a WER of 58.1 % and CER of 27.7 % when evaluated on the SO test set. For the language model component, we concatenated text data from all five languages, using the data-sources listed in Table 2.

Given the success and popularity of large pre-trained self supervised models for low-resource ASR speech recognition (e.g., wav2vec 2.0 [2]) and to compare our system to one based upon a self-supervised model, as a control experiment we built the system detailed in Table 3. This used the pre-trained self-supervised model XLSR-53 [13]—a multilingual version of wav2vec 2.0 [2] (trained on 53 languages)—to extract hidden representations as input features for a small acoustic model trained on the SO training data.

Table 2: The six datasets used to build our final multi-lingual ASR system.

ID	Dataset	Languages	Domain	Use
1	MGB	eng	Multi-Genre Broadcast	Pre-training Acoustic Model
2	SO (train)	eng, xho, zul, sot, tsn	(Scripted) Conversational	Fine-tuning Acoustic Model
3	MGB Text	eng	Multi-Genre Broadcast	Training Language Model
4	NCHLT Text	eng, xho, zul, sot, tsn	Multi-Genre	Training Language Model
5	Scraped News	xho, zul	News	Training Language Model
6	SO (test)	eng, xho, zul, sot, tsn	(Scripted) Conversational	Testing System

Table 3: The six datasets used for our control experiment. Datasets 3–6 were identical to those shown in Table 2 (not repeated here). Differences in dataset sources and uses are bolded for clarity.

ID	Dataset	Languages	Domain	Use
1	**XLSR-53**	**53 languages**	**Multi-genre**	**Feature Extraction for Acoustic Model**
2	SO (train)	eng, xho, zul, sot, tsn	(Scripted) Conversational	**Training Acoustic Model**
3–6			*(As in Table 2)*	

This control model achieved a WER of 54.0 % and CER of 25.8 %. Importantly, though these error rates are lower than those achieved by our MGB-based model, this improvement is only slight. Recall from Section 1 our concerns regarding these large-scale self-supervised approaches to low-resource recognition. It is thus encouraging that we can reach similar performance levels with a system which does **not** rely on self-supervised approaches that bypass communities. Moreover, self-supervised models—like XLSR-53—cannot be used for real-time speech recognition, as they need access to a complete utterance recording in advance. As such, our MGB-based system (see Table 2) is more suitable than the control system (Table 3) for deployment.

6.2 The ASR demonstrator appliance

The ASR demonstrator appliance, which utilises the above ASR system, is shown in Fig. 4. Compared to the SpeechBox, the enclosure is much smaller and designed to look like a smart speaker. The keypad is replaced with two rating buttons (+/-). An LED ring light mounted around the main button and behind the semi-transparent front panel creates a diffusing effect, and replaces the original version's display.

On the software side, we integrated the demonstrator appliance with the ASR system using bidirectional streaming protocols[10] to transmit audio requests and concurrently receive transcript responses, so users can expect a result shortly after they finish speaking.

We also used the improved ASR system to re-transcribe all audio recordings that were not flagged as blank, inappropriate, or underage. Following von Holy et al.'s research on accessing digital libraries through search in resource-scarce South African languages including isiXhosa [59], we indexed and queried COVID-19 story transcripts as is (without stemming or n-grams) using the Apache Lucene[11] open-source search software.

When users walk up to the device, pressing the main button initiates an interaction flow. The user is greeted by a prompt, explaining that the device contains stories collected from Langa about COVID-19 and that they can press and hold the button again and tell the device what they are interested in. While the device is recording, the ASR system transcribes the audio data, which is not otherwise retained on the device or on the server. The LED ring light visualises the microphone signal, to indicate to the user that the device is actively listening. Given the high WER we decided not to surface transcripts, and instead used it to query the search engine. We posited that having both the query and the recorded stories transcribed by the same ASR system would lead to more relevant results than using those created by the human transcribers. For instance, during tests with the community liaison it emerged that the ASR system would often transcribe the spoken term 'covid' incorrectly as the similar-sounding isiXhosa word 'khopi'. However, this would not negatively effect search results as it would transcribe consistently across the query and the indexed documents so the search engine could still match on the term itself. To further account for variations in spelling we used fuzzy queries that allow for small number of one-character changes when matching terms (for instance, when removing (flack ⟶ lack), changing (dog ⟶ fog), or inserting (dog ⟶ dogs) a character, or transposing two adjacent characters (act ⟶ cat)).

While awaiting a response, the LED ring light acts as a spinning indeterminate progress indicator. Once results are received, the appliance plays the search result audio and prompts the user to rate the relevance of the story on a scale of 1–5 using the plus or minus buttons. At any time the user can press the main button to record a new query. If no result is returned the user is also invited to record a new query. If any device prompt is not responded to within a short timeout, the device resets and waits for the next button press which restarts the process.

[10] https://grpc.io/

[11] https://lucene.apache.org/

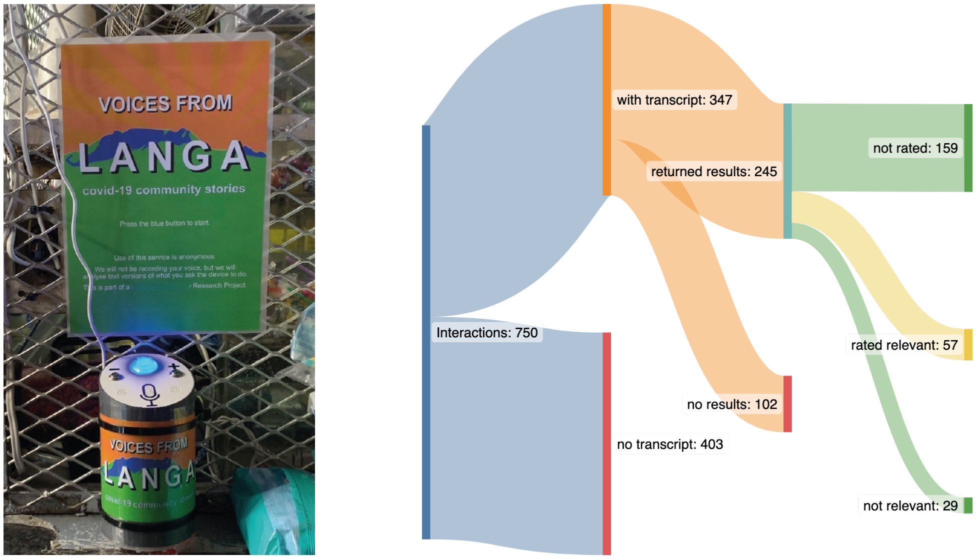


Figure 4: The ASR demonstrator as it was deployed in one of five shops in Langa over a period of three weeks (left) and a breakdown of recorded interactions on the devices (right). Around a third of queries generated a search response (i.e., an oral community story about COVID-19 that we could match to the incoming voice query). Few participants chose to respond to the post-interaction rating question, but of those who did, 66 % rated the result as relevant.

6.3 Pilot workshop

The community liaison recruited six L1 isiXhosa speaking residents of Langa (3M/3F; aged 25–41) to participate in the pilot workshop and again determined a locally appropriate compensation amount.

Recurring power blackouts[12] contributed to degrading network latency on the demonstrator's 4G internet connection and negatively impacted ASR system performance during the workshop. As a result, we had to make do by 'hard-coding' responses on a predetermined set of topics that we could switch between during the study. To determine these topics, we iteratively and collaboratively went through the transcribed corpus with the community liaison to find recurring themes, settling on the following: *jobs*, *money*, *sickness*, *school*, and *medicine*.

Once participants arrived, we introduced the larger research project and went through ethics and consent. We showed the SpeechBox and explained how it had been recording people's experiences of COVID-19 in Langa earlier in the year. Turning to the ASR demonstrator appliance, we explained how people can now access those stories using their voice.

We then asked participants to split into two groups of three people and trial the appliance. We asked both groups to first find stories related to *jobs*, then on the topic of *money*, and finally about *sickness*. Participants asked questions such as *"how did people make money during covid"* and *"how did the government give their country help with money"*. While the hard-coded responses did not address the question directly—one was about the cost of airtime and another on how someone spent money during COVID-19—they did approximate how search engines often return partially-matching or less-relevant results.

Participants also critiqued the scaled rating mechanism (1–5 stars) and suggested a simpler binary rating prompt (+ or -), which also simplified the interaction.

We asked participants where to install the devices in Langa, who of their friends and family might try the device, and why. Participants mentioned that places that people pass through, where groups already gather (e.g., spots with free WiFi), or where they might be waiting (e.g., for a minibus taxi) would likely stimulate usage. But they were also wary of putting the device, which also contains sombre stories, in places where people want to socialise, such as a tavern. In this sense the device might be more useful, in future, installed in a library or community centre so people could remember and learn about what had happened at the height of COVID-19. They thought elders might be curious to learn about the stories within the device. Younger people might be curious because of the similarity to Alexa-style smart speakers that they have seen in movies but not yet used. And children would likely be disappointed because they would want the device to play music.

6.4 Deployment

We deployed the ASR demonstrator devices over a period of three weeks in five locations: four spaza shops and an internet cafe/print shop. Figure 4 gives an overview of the 750 interactions that were initiated on the devices. Of these, 403 (~54 %) generated no transcript. As we intentionally did not log audio, we have no way of knowing if this was because users were not speaking, the ASR system did not recognise any words that were spoken, or internet connectivity was causing interference. The remaining 347 interactions (~46 %) generated a transcript. Of these, 102 (~29 %) did not return a matching search result from the corpus, leaving 245 transcripts that returned search results (~71 %), which were played back on the devices. Most

[12] https://www.capetown.gov.za/loadshedding/

of these (159 or ~65 %) were not rated by users, who likely walked away; however, 57 stories (~23 %) that were played back were rated as relevant by users and the remaining 29 stories (~12 %) rated as not relevant.

6.5 Observational study and workshop

Three days into the deployment we conducted an observation study and workshop for which we asked the community liaison to recruit 13 participants who had not used the device before. We began by asking participants to individually use the ASR demonstrator appliance in a nearby spaza shop, explaining that we would like to observe and video them before reconvening the workshop and discussing their experience. This gave us the opportunity to observe how novice users would approach the device in context and the video served as a reminder of what happened during group discussions. All participants consented to take part.

One of the more interesting insights that this method surfaced was that people did not know what to do after listening to the initial prompt. To be sure, being video-recorded will have contributed to the uncertainty that participants experienced. However, group discussions revealed that the initial prompt was harder to understand because it used 'deep Xhosa', as it is spoken in the traditional homeland of the amaXhosa in the Eastern Cape Province, and not the urban variety that is spoken in Langa (see also [18]). The prompt started with "this device *contains* stories …" and participants suggested that changing it to "this device *has* stories …" would make it more understandable.

We also discussed why certain queries failed to return results, and with the help of the community liaison later interrogated a specific scenario that the group discussions brought up: learning how people's work and jobs were affected. Consider these three snippets about jobs and work taken from the corpus: (1) "iCovid 19 indithathele *umsebenzi*" – "Covid 19 has taken my *job*"; (2) "abantu baphelelwa *yimisebenzi* …" – "people lost their *jobs*"; and (3) "*andisebenzi*" – "I don't *work*". Notice how the emphasised words share the common *suffix* 'sebenzi', whereas the English translation of the first two emphasises words share the common *prefix* 'job'. Querying a translated corpus for 'job' would have returned two search results, whereas the search engine—not tuned to isiXhosa—could only ever return one result, depending on which term is searched for: umsebenzi, yimisebenzi, or andisebenzi. This also helps contextualise why 102 transcribed queries did not return a result (see Fig. 4).

7 USING THE DATA

Study II focused on gathering and transcribing isiXhosa speech recordings. Subsequently, Study III discussed our building and deployment of an isiXhosa ASR system. Importantly, this system did *not* directly use recordings that were gathered during Study II as an explicit ASR training set. Here then we discuss the alternative value of such data for developing ASR systems and why attending to specific qualities of that data, surfaced through our community engagements, caused us to pause and reflect, rather than immediately integrate the data in the ASR pipeline used in Study III. This is why we position community-engaged data as an invaluable conceptual tool to uncover pressing issues surrounding evaluation metrics and illustrate how future research trajectories could respond to these.

isiXhosa/English Transcript	English Translation
i*story* sam *about* i*covid* … iye ndaphelelwa ngumsebenzi *due* i*covid and i lost* uMakhulu wam oye wagula ngesaquphe senditsho uba … *even* nakwi *community* ithi abantu balahlekelwe zizihlobo zabo	My *story about covid* ... I lost my job *due to covid and I lost* my grandmother who fell ill suddenly ... *even* in the *community*, people have lost their friends

Table 4: Original transcription and translation of a story about loss endured during COVID-19 showcasing fluid and unconstrained code-switching between isiXhosa (upright text) and English (italic text).

7.1 Community-engaged COVID-19 stories & transcripts

Because the design of the Speechbox and its deployment encouraged wide participation in a community that is linguistically and structurally marginalised, the collection of spoken story contributions users made are a rare and useful resource that differs from existing corpora of isiXhosa [3, 54]. The dataset features (unscripted) conversational speech, often draws on different languages and sometimes multiple speakers, and overall more accurately reflects the way Langa residents speak in day-to-day life. In fact a salient feature of the dataset, and of language-use in Langa more generally, is fluid and non-domain-specific code-switching. In other contexts, code-switching tends to be more domain-bound – see for instance [27, 31], or contain only a single switch point, such as the example from the Soap Opera dataset shown in Table 1. In comparison, in the transcript of a story of loss endured during COVID-19 shown in Table 4, the speaker uses English connectives ('and' & 'even') amidst isiXhosa phrases within the same sentence. It also shows intra-word code-switching examples ('iStory' & 'iCovid'), but interestingly 'community' is mentioned without the isiXhosa noun prefix 'i'. Notice, too, how the speaker once uses isiXhosa utterances for 'I lost' when speaking of losing their job, and shortly thereafter uses the English phrase 'I lost' when speaking about loss of a family member. The frequency and unconstrained nature of this switching (i.e., English usage is not just constrained to content words of a particular domain), and the intra-word examples, necessitates more sophisticated handling than simply running an English and isiXhosa ASR system concurrently. Improving the code-switching capabilities of our ASR system—e.g., by modelling likely switch points between languages or how particular topics (jobs, money, medicine) influence code-switching—is thus an integral next step.

A further feature of the dataset is the variability of transcriptions, discussed in Section 5.3, which are not particularly amenable to integration into traditional ASR development pipelines without at least significant normalisation. But, more troubling for ASR orthodoxy, this variability evidences the notion that we cannot assume that 'gold-standard' transcripts exist, particularly for vernacular languages—such as isiXhosa—that do not necessarily have a well-defined (or widely used) written standard [6] and where such standards are entwined with colonial encounters [18].

Previous research on crowdsourced transcription—albeit in settings where languages have more defined written standards and norms—have utilised algorithmic approaches (e.g., multiple string alignment and majority voting) to determine the best transcript estimate across multiple differing transcript candidates [57]. However, in our case this normalisation would, in effect, create an artificially imposed 'standard', that would risk circular ASR development as subsequent model iterations would be based on this, and reinforce that imposed 'standard'.

Our decision to engage with community members directly, rather than employing external linguists or second language speakers, to transcribe stories affords us the opportunity to draw on local (socio)linguistic expertise and local writing and transcription practices. By comparing multiple transcripts from several annotators, we can better understand and respond to variations in spelling, lexis, and interpretation.

These variations and lack of written standards, have important implications for how we develop and evaluate ASR systems, for conventional ASR approaches rely on both 'gold standard' transcripts and 'standard' metrics (e.g. WER/CER), which have considerable flaws even in 'high-resource' settings [23, 40]. Consider that, on average, there was a 63.9 % disagreement rate between annotators at a word level, and a 30.4 % disagreement rate at a character level. Note then that we would infer very different things about our ASR system's quality and what required improving depending on how we normalise transcripts. The dataset therefore reveals, how this standard approach to ASR development and evaluation breaks down in the context of Langa. The problem of meaningful and contextualised evaluation is, of course, not limited to 'low-resource' languages or speech technologies but an increasingly pressing issue in almost all domains of machine learning [14].

7.2 Next steps: community-engaged evaluation

The collected speech samples and accompanying transcripts surfaced fundamental problems with standard metrics to evaluate ASR systems for minoritised languages and less standardised languages. In fact, attending to transcript variations reminded us of the challenges we experienced tuning our ASR models-in-the-making, whereby different models would split longer utterances into words at different points, much like the annotators of Table 1. Model tuning is a highly iterative process whereby design decisions are evaluated and compared using standard metrics to select/pursue the model which achieves the lowest WER. As traditional ASR development pipelines rely on the notion of a 'gold-standard' transcript, they fail to acknowledge that variability in word splitting is a feature of language-in-use in Langa. In other words, the orthodox evaluation methodology penalises systems for 'mistakes' that an isiXhosa speaker in Langa may not recognise as such.

Likewise, the tuning of code-switching needs to be carefully balanced. For instance, some models would switch into English too aggressively, leading to isiXhosa utterances to be incorrectly 'transcribed' as the closest matching English word(s), or omitted altogether, as was the case with the intermittent model used in Table 5. We also occasionally saw the inverse: transcribing isiXhosa when English was spoken. Such errors may also necessitate additional forms of evaluation to probe if users feel differently or more strongly about errors involving English being output instead of isiXhosa.

Table 5: An intermittent model incorrectly handling code-switching in the Covid Stories data by outputting English (italic) instead of isiXhosa.

Reference	ASR Output
… nidlala ngathi nina ndizawuthini ndizawuthini ndizawuthini	indlela *is a teen queen to michael mcintyre*

Given the limitations of current evaluation methodology then, for future work we propose to engage community members in the evaluation and tuning of ASR-models-in-the-making. This includes presenting community members with a series of transcripts of the same audio but generated by differently tuned ASR models and asking for feedback on the transcription 'errors' that matter more to them and what a 'good'—or 'good enough'—transcript looks like to them. Situating this further aspect of the ASR development pipeline in situ would guide design choices and create more appropriate systems.

8 DISCUSSION, CONTRIBUTIONS AND CONCLUSION

We started this research endeavour with the observation that currently available ASR systems do not support isiXhosa. We also argued that state-of-the-art approaches to ASR development, with their big data and computational requirements [51], are placed out of reach of all but the largest companies, and bypass the very communities that are currently unheard [12]. However, for the baseline isiXhosa system we developed early on in this programme of work, using more traditional approaches to ASR development, we could only utilise existing datasets. These consisted of a larger, read-speech dataset [3] for training and a smaller, code-switched isiXhosa-English TV soap-opera dataset [54] for testing. The decision to leverage the more 'in-domain', soap-opera dataset for testing resulted in poorer WER and CER metrics than splitting the larger read-speech dataset into training and testing subsets. However, that decision also gave a preview of the types of speech—faster paced and mixed—it would be confronted with in Langa, and allowed us to tune and augment the model accordingly. In the community engagement workshop participants confirmed the shortcomings of the ASR system as indicated by its metrics, but also found instances where the ASR system performed well enough for them to get the gist of a recorded message from its transcript. Participants also indicated that an ASR system for 'voice typing' would be tremendously useful in their context. It is clear that for traditional approaches to succeed in Langa, and unlock this use-case, better data is required and more of it.

In search for such data, that in our case reflects the ways that people speak and mix languages in Langa, we designed a toolkit for community-engaged ASR development. We tailored the SpeechBox with the community liaison, decided on the much discussed topic of COVID-19, and surfaced and addressed usability issues through a formative design workshop. Deploying the SpeechBox in two public locations, residents recorded 318 responses. In contrast, only about 10 % of those respondents completed an online survey that would have earned them a further incentive payment. Workshops revealed further difficulties installing mobile apps, which often had cracked screens, not enough storage to install an additional app, and degraded battery life (see [62]). Compared to surveys and apps, we contend that the SpeechBox is a more effective and accessible tool to gather responses from people that otherwise face barriers to digital participation. The contents of the audio responses also reflected the messiness and dynamic language of social life in Langa writ large that was missing from existing datasets. Here audio contents consisted of mostly isiXhosa mixed with occasional English terms or phrases (e.g., 'iCovid', 'iLockdown', or 'iMask').

To involve community members in the transcription of collected audio data, we created the TranscriptTool mobile app to support and scaffold the core audio transcription task. Through a formative design workshop we again surfaced and addressed usability issues, then enlisted community members to transcribe the corpus of recordings. Interviews with participants and examination of the transcribed dataset itself revealed that participants completed transcript tasks of often noisy recordings diligently and comprehensively. This contrasts with previous speech-based research in Langa that did not support transcription through bespoke tooling and where transcripts often omitted words and only paraphrased what was said [45]. We therefore contend that the TranscriptTool was effective in supporting participants from marginalised communities in comprehensively transcribing audio content. The tool also extends HCI/ASR scholarship on crowdsourced transcription because it does not depend on an existing ASR systems (e.g., [57]) or support only read speech (e.g., [1]).

However, participant transcripts also exhibited variability, particularly in regard to word boundaries, a phenomena we repeatedly encountered in training and tuning our ASR systems. This helps explain why participants in earlier workshops were able to make sense of some messages from ASR transcripts alone, and why character error rates were generally better than the word error rate metrics would suggest. The ASR development methodology/pipeline we leveraged throughout this research has previously achieved good results in competition[13]. Such competitions are a common organising force within many domains of AI—including NLP—to drive the field forward [39]. However, engaging directly with community and context, as we have done through our research and with the help of our toolkit, rather than indirectly through competition, also forced us to confront core assumptions of ASR orthodoxy. While community/user engagement—outside of usability labs—has become a central tenant of contemporary HCI research [47], this is not (yet) the case in ASR research. So the community-generated dataset, in all its messiness and variability, represents a further, conceptual part of our toolkit. For the dataset gives concrete evidence of the wicked problems ASR developers face when modelling languages that do not appear to have a well-defined standard [6], where language speakers have limited practice with the written norms of the language, and where those written norms are intertwined with colonial encounters and have not kept abreast with the vibrancy and innovation of the language [18], particularly as it is spoken and mixed in public spaces in Langa.

We therefore propose further research to engage community members in the tuning process of ASR development, to weigh in on the choices and compromises that ASR developers make, and inform decisions and identify possible consequences. The toolkit we introduce, document, and open-source[14] here, and the community connections and trust it facilitated, will underpin those future research trajectories, and with the help of the toolkit we hope that others will join us, and engage with diverse language communities, in this challenging area of research. Which brings us to the final component of our toolkit: the ASR demonstrator appliance that we tailored and deployed to give community members the opportunity to query their collection of stories of people's experiences of COVID-19. In this sense, the ASR demonstrator also serves as a methodological reminder that reciprocity and feeding back research results (see [53]) is a cornerstone of community-engaged research.

ACKNOWLEDGMENTS

We would like to thank Minah, study participants, and all SpeechBox users and transcribers in Langa for their contribution to this work. All source code data is available at https://github.com/FITLab-Swansea/UnMute-Toolkit. This work was supported by Engineering and Physical Sciences Research Council grant EP/T024976/1.

REFERENCES

[1] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2819–2826.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]* (Oct. 2020). arXiv:2006.11477 [cs, eess]

[3] E. Barnard, M. H. Davel, C. Van Heerden, Febe De Wet, and J. Badenhorst. 2014. The NCHLT Speech Corpus of the South African Languages. In *4th International Workshop on Spoken Language Technologies for Under-Resourced Languages*. St Petersburg, Russia.

[4] Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, et al. 2015. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 687–693.

[5] Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3504–3519. https://doi.org/10.18653/v1/2020.coling-main.313

[6] Steven Bird. 2022. Local Languages, Third Spaces, and Other High-Resource Scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7817–7829. https://doi.org/10.18653/v1/2022.acl-long.539

[7] Astik Biswas, Ewald van der Westhuizen, Thomas Niesler, and Febe de Wet. 2018. Improving ASR for Code-Switched Speech in Under-Resourced Languages Using Out-of-Domain Data.. In *SLTU*. 122–126.

[8] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things out: Classification and Its Consequences*. MIT Press, Cambridge, Mass.

[9] Mary Bucholtz. 2000. The Politics of Transcription. *Journal of Pragmatics* 32, 10 (2000), 1439–1465. https://doi.org/10.1016/S0378-2166(99)00094-6

[10] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, USA, 1–12.

[11] Manu Chopra, Indrani Medhi Thies, Joyojeet Pal, Colin Scott, William Thies, and Vivek Seshadri. 2019. Exploring Crowdsourced Work in Low-Resource Settings.

[13] A top-placed award at a multi-lingual and code-switched challenge for low-resource languages in India [20].

[14] https://github.com/FITLab-Swansea/UnMute-Toolkit

In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[12] Donavyn Coffey. 2021. Māori Are Trying to Save Their Language from Big Tech. *Wired UK* (April 2021).

[13] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979* (2020).

[14] Kate Crawford. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press, New Haven.

[15] Andiswa Mesatywa Dantile. 2015. *Language in Public Spaces: Language Choice in Two IsiXhosa Speaking Communities (Langa and Khayelitsha).* Ph. D. Dissertation. University of Stellenbosch, Stellenbosch, South Africa.

[16] Indra de Lanerolle, Marion Walton, and Alette Schoon. 2017. *Izolo: Mobile Diaries of the Less Connected.* The Institute of Development Studies, Brighton.

[17] N. J. De Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. De Waal. 2011. Woefzela - An Open-Source Platform for ASR Data Collection in the Developing World. (Aug. 2011).

[18] Ana Deumert. 2010. Imbodela Zamakhumsha – Reflections on Standardization and Destandardization. 29, 3-4 (Nov. 2010), 243–264. https://doi.org/10.1515/mult.2010.012

[19] Ana Deumert. 2014. *Sociolinguistics and Mobile Communication.* Edinburgh University Press, Edinburgh.

[20] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, et al. 2021. Multilingual and code-switching ASR challenges for low resource Indian languages. *arXiv preprint arXiv:2104.00235* (2021).

[21] Vittoria Elliott and Bopha Phorn. 2021. Fifty Percent of Facebook Messenger's Total Voice Traffic Comes from Cambodia. Here's Why. https://restofworld.org/2021/facebook-didnt-know-why-half-of-messengers-voice-traffic-comes-from-cambodia-heres-why/.

[22] Stephen Ellis. 1989. Tuning In to Pavement Radio. *African Affairs* 88, 352 (1989), 321–330.

[23] Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare Voss, and Frauke Zeller. 2013. Automatic human utility evaluation of ASR systems: does WER really predict performance?. In *Proc. Interspeech 2013.* 3463–3467. https://doi.org/10.21437/Interspeech.2013-610

[24] Rosalie Finlayson, Karen Calteaux, and Carol Myers-Scotton. 1998. Orderly Mixing and Accommodation in South African Codeswitching. *Journal of Sociolinguistics* 2, 3 (1998), 395–420. https://doi.org/10.1111/1467-9481.00052

[25] Fiona Anciano, SJ Cooper-Knock, Mmeli Dube, Mfundo Majola, and Boitumelo M. Papane. 2020. Cape Town Lockdown Diaries. https://capetownlockdown.wordpress.com/.

[26] Alan Galey and Stan Ruecker. 2010. How a Prototype Argues. *Literary and Linguistic Computing* 25, 4 (Jan. 2010), 405–424. https://doi.org/10.1093/llc/fqq021

[27] Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* 1850–1855.

[28] Esther Sánchez García and Michael Gasser. 2021. Stochastic Parrots: How NLP Research Has Gotten Too Big. *Science for the People Magazine* 24, 2 (2021).

[29] Connie Golsteijn, Sarah Gallacher, Lisa Koeman, Lorna Wall, Sami Andberg, Yvonne Rogers, and Licia Capra. 2015. VoxBox: A Tangible Machine That Gathers Opinions from the Public at Events. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '15).* Association for Computing Machinery, Stanford, California, USA, 201–208. https://doi.org/10.1145/2677199.2680588

[30] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt, Boston.

[31] Gualberto A Guzman, Jacqueline Serigos, Barbara Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the second workshop on computational approaches to code switching.* 12–20.

[32] Lauren Hall-Lew, Claire Cowie, Catherine Lai, Nina Markl, Stephen Joseph McNulty, Shan-Jan Sarah Liu, Clare Llewellyn, Beatrice Alex, Zuzana Elliott, and Anita Klingler. 2022. The Lothian Diary Project: sociolinguistic methods during the COVID-19 lockdown. *Linguistics Vanguard* 8, s3 (2022), 321–330. https://doi.org/doi:10.1515/lingvan-2021-0053

[33] Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mike LeBeau. 2010. Building Transcribed Speech Corpora Quickly and Cheaply for Many Languages. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010).* 1914–1917.

[34] Christiaan Jacobs and Herman Kamper. 2021. Multilingual transfer of acoustic word embeddings improves when training on languages related to the target zero-resource language. *arXiv preprint arXiv:2106.12834* (2021).

[35] Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities. https://doi.org/10.48550/arXiv.1912.03457 arXiv:1912.03457 [cs]

[36] Daniel Lambton-Howard, Robert Anderson, Kyle Montague, Andrew Garbett, Shaun Hazeldine, Carlos Alvarez, John A. Sweeney, Patrick Olivier, Ahmed Kharrufa, and Tom Nappey. 2019. WhatFutures: Designing Large-Scale Engagements on WhatsApp. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–14.

[37] Gary Marsden. 2008. Toward Empowered Design. *Computer* 41, 6 (2008), 42–46.

[38] James Milroy. 2001. Language Ideologies and the Consequences of Standardization. *Journal of Sociolinguistics* 5, 4 (2001), 530–555. https://doi.org/10.1111/1467-9481.00163

[39] Melanie Mitchell. 2019. *Artificial Intelligence: A Guide for Thinking Humans.* Pelican Books, London.

[40] Andrew C Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition. In *INTERSPEECH-2004.* 2765–2768.

[41] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[42] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. StreetWise: Smart Speakers vs Human Help in Public Slum Settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300326

[43] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society.

[44] Agha Ali Raza, Awais Athar, Shan Randhawa, Zain Tariq, Muhammad Bilal Saleem, Haris Bin Zia, Umar Saif, and Roni Rosenfeld. 2018. Rapid Collection of Spontaneous Speech Corpora Using Telephonic Community Forums. In *Interspeech 2018.* ISCA, 1021–1025. https://doi.org/10.21437/Interspeech.2018-1139

[45] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3491102.3517639

[46] Anna Rogers. 2021. Changing the World by Changing the Data. *arXiv:2105.13947 [cs]* (May 2021). arXiv:2105.13947 [cs]

[47] Yvonne Rogers. 2012. HCI Theory: Classical, Modern, and Contemporary. *Synthesis Lectures on Human-Centered Informatics* 5, 2 (May 2012), 1–129. https://doi.org/10.2200/S00418ED1V01Y201205HCI014

[48] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. 2010. Intermediated Technology Use in Developing Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10).* ACM, New York, NY, USA, 2583–2592. https://doi.org/10.1145/1753326.1753718

[49] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445518

[50] Alette Schoon. 2016. Distributing Hip-hop in a South African Town: From the Digital Backyard Studio to the Translocal Ghetto Internet. In *Proceedings of the First African Conference on Human Computer Interaction (AfriCHI'16).* ACM, New York, NY, USA, 104–113. https://doi.org/10.1145/2998581.2998592

[51] Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The Cost of Training NLP Models: A Concise Overview. *arXiv:2004.08900 [cs]* (April 2020). arXiv:2004.08900 [cs]

[52] Betsy Sneller, Suzanne Evans Wagner, and Yongqing Ye. 2022. MI Diaries: ethical and practical challenges. *Linguistics Vanguard* 8, s3 (2022), 307–319. https://doi.org/doi:10.1515/lingvan-2021-0051

[53] Fiona Ssozi-Mugarura, Edwin Blake, and Ulrike Rivett. 2017. Codesigning with Communities to Support Rural Water Management in Uganda. *CoDesign* 13, 2 (April 2017), 110–126. https://doi.org/10.1080/15710882.2017.1310904

[54] Ewald van der Westhuizen and Thomas Niesler. 2018. A First South African Corpus of Multilingual Code-Switched Soap Opera Speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA), Miyazaki, Japan.

[55] Ewald van der Westhuizen and Thomas R Niesler. 2019. Synthesised bigrams using word embeddings for code-switched ASR of four south african language pairs. *Computer Speech & Language* 54 (2019), 151–175.

[56] Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–13.

[57] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1855–1866.

[58] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[59] Andreas von Holy, Alon Bresler, Osher Shuman, Catherine Chavula, and Hussein Suleman. 2017. Bantuweb: A Digital Library for Resource Scarce South African Languages. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists on - SAICSIT '17*. ACM Press, Thaba 'Nchu, South Africa, 1–10. https://doi.org/10.1145/3129416.3129446

[60] Marion Walton. 2014. Pavement Internet: Mobile Media Economies and Ecologies for Young People in South Africa. In *The Routledge Companion to Mobile Media*, G. Goggin and Larissa Hjorth (Eds.). Routledge, London, UK.

[61] Frederick Weber, Kalika Bali, Roni Rosenfeld, and Kentaro Toyama. 2008. Unexplored Directions in Spoken Language Technology for Development. In *IEEE Spoken Language Technology Workshop*. 1–4. https://doi.org/10.1109/SLT.2008.4777825

[62] Susan Wyche, Tawanna R. Dillahunt, Nightingale Simiyu, and Sharon Alaka. 2015. "If God Gives Me the Chance i Will Design My Own Phone": Exploring Mobile Phone Repair and Postcolonial Approaches to Design in Rural Kenya. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 463–473. https://doi.org/10.1145/2750858.2804249

[63] Xuecong Xu, Xiang Yan, and Tawanna R. Dillahunt. 2019. Reaching Hard-To-Reach Populations: An Analysis of Survey Recruitment Methods. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 428–432. https://doi.org/10.1145/3311957.3359447