Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource 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

Dani Kalarikalayil Raju
Studio Hasi
Mumbai, India
daniel@studiohasi.com

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

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

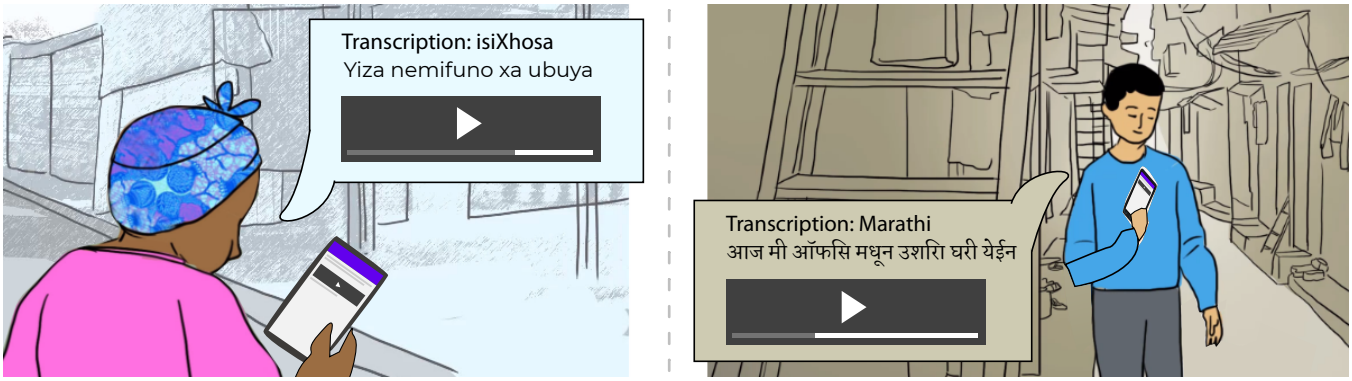

Figure 1: Artist's representation of the Automatic Speech Recognition systems we developed and field-tested in partnership with two communities in South Africa and India to transcribe isiXhosa (left) and Marathi (right) voice messages.

ABSTRACT

Automatic Speech Recognition (ASR) researchers are turning their attention towards supporting low-resource languages, such as isiXhosa or Marathi, with only limited training resources. We report and reflect on collaborative research across ASR & HCI to situate ASR-enabled technologies to suit the needs and functions of two communities of low-resource language speakers, on the outskirts of Cape Town, South Africa and in Mumbai, India. We build on long-standing community partnerships and draw on linguistics, media studies and HCI scholarship to guide our research. We demonstrate diverse design methods to: remotely engage participants; collect speech data to test ASR models; and ultimately field-test models with users. Reflecting on the research, we identify opportunities, challenges, and use-cases of ASR, in particular to support pervasive use of WhatsApp voice messaging. Finally, we uncover implications for collaborations across ASR & HCI that advance important discussions at CHI surrounding data, ethics, and AI.

CCS CONCEPTS

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

KEYWORDS

Speech/language, automatic speech recognition, mobile devices

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29–May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3491102.3517639

CHI '22, April 29–May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517639

1 INTRODUCTION

Researchers in the field of Automatic Speech Recognition, or ASR, are increasingly turning their attention towards so-called low-resource languages.[1] At the same time companies like Facebook and Google are developing their own ASR approaches (e.g., Facebook's wav2vec [66]) and language models (e.g., Google's Cloud Speech-to-Text [31]).

Such 'unsupervised' approaches no longer require the accompaniment of 'gold-standard' transcriptions but depend on large amounts of training data instead. As critical scholarship reveals, such approaches raise salient questions surrounding data privacy and data collection practices [87]. For instance, unsupervised approaches often rely on web-scraping techniques to collect and augment training data; these work on the problematic assumption that such language data is interchangeable – regardless of where it is found [13]. Furthermore, training Natural Language Processing (NLP) models on such big datasets comes with substantial additional costs: cf. the cloud computing bills that run into the hundreds of thousands of dollars to train, test, and tweak AI models [67], and the associated climate [72] and rare earth mineral [13] impacts that such large-scale computations require.

While the current state-of-the-art in 'low-resource' ASR methodologies is seeing burgeoning interest and rapid development, we worry that these are motivated by the intellectual challenge [8] and are increasingly distant from users and communities themselves—far removed from the specific needs and functions of the speakers of such low-resource languages—and are being developed upon datasets that do not adequately represent the ways in which people speak and communicate. Given HCI's 20-year history of working with marginalised and minority users in the Global South [26], in this paper we develop a HCI perspective to collaborate with ASR researchers and language experts: to leverage their advances, but also to contextualise, situate and develop use-cases and datasets for ASR in partnership with two communities of minority language speakers.

Looking beyond technical challenges and using the lenses afforded by post-colonial computing frameworks [25, 38] we can interrogate what it means that Google's Cloud Speech-to-Text service now supports three languages spoken in South Africa—Afrikaans, English, and isiZulu—and raise critical questions, such as why these three and not one of the country's other eight official languages? In a post-colonial setting such as South Africa, with legacies of racist ideologies and eugenic science that justified colonial and apartheid rule, such questions rarely have neutral answers. For the twin projects of colonialism and apartheid entrenched various forms of inequality that, sadly, persist to this day: racial, economic, digital, but also linguistic inequality. It is great to see South Africa's second most widespread language supported: isiZulu, which is predominately spoken in the economically prosperous Gauteng province. However, isiXhosa, which is mostly spoken by people from the less prosperous Eastern Cape, is technically the third most widespread language, rather than Afrikaans, which next to English is the second main language spoken by White South Africans of Dutch, French, and German descent [54].

Our research is motivated by addressing this class of gap and, in this instance, developing a baseline isiXhosa ASR language model; but, just as importantly, we uncover the challenges, opportunities, and use-cases for such an ASR system in one particular marginalised South African community. We are also mindful of HCI scholarship that identifies text-input as a barrier to digital participation for low-resource language speakers in a different setting: Marathi-speaking 'emergent users' in India [20]. The Devanagari script of Marathi means that its speakers are often forced to transliterate their messages into the Latin script, or to install custom keyboards, which presents its own set of challenges such as more complex UI hierarchies [82]. What role could ASR-enabled technologies play here?

To address these questions, we engage with communities directly and report and reflect on interviews and co-design workshops involving data-collection exercises and technology probes. Through these diverse activities we identify opportunities, challenges, use-cases, and implications for collaborations both across our fields of research and crucially through involving community partners in an isiXhosa-speaking township in the outskirts of Cape Town, South Africa and with Marathi-speaking residents of an informal settlement in the heart of Mumbai, India.

Through our research we identify a novel use-case for voice technologies in such communities—voice message transcription—and demonstrate pervasive and creative use of WhatsApp voice messaging. Finally, we develop prototypes of baseline Speech Recognition systems and field-test these to further engage user communities. We review user perspectives collaboratively, and consider how challenges—particularly surrounding data collection and transcription—could be addressed from both ASR and HCI perspectives.

2 CONTEXT

We begin our inquiry by introducing the communities and localities we partnered with, namely Langa and Dharavi.

2.1 Langa, Cape Town, South Africa

With about 50,000 residents Langa is the oldest township in Cape Town, a peripheral locality in the Cape Flats that was set aside for nonwhites during apartheid rule. As the provenance of the place itself is rooted in discrimination and racism, Langa is classified as a previously disadvantaged area and, sadly, the issues of structural inequalities, poverty and crime established during apartheid not only persist to this day, but often (though no longer exclusively) fall along racial lines and exhibit a strong spatial dimension [85].

While Langa is inhabited largely by first-language isiXhosa speakers, the linguistic landscape of public spaces—local newspapers, advertising, business names, government and community notices, etc.—is dominated by English and Afrikaans [14]. Co-designing and carrying out formative and summative evaluations of novel technologies in Langa over the past ten years we have also come to know and appreciate the hospitality, good humour and honesty of its community members. As COVID-19 travel restrictions and lockdowns rendered co-located design activities impossible, we

[1] Here and elsewhere in this contribution we adopt the ASR term 'low-resource languages' to draw attention to the dearth of publicly-available data and to build bridges to ASR researchers working in that area.

relied on the relationships and trust we have formed over many years as well as on local facilitator embedded in the community to support unfamiliar, technologically-mediated design workshops.

2.2 Dharavi, Mumbai, India

Our team has similar longstanding and equally formative experiences of partnering with community members in Dharavi, which we again relied on as we pivoted to remote interviews and technology deployments. In addition, our team also consists of a researcher (and author of this paper) who lives in a nearby suburb in Mumbai and is a fluent Hindi speaker.

Dharavi, colloquially referred to as "the largest slum in Asia," is a neighbourhood in the centre of Mumbai, right next to the financial centre of India, and therefore occupies some of the most expensive and sought after real estate in the world [40, p.264]. Situated on initially undesirable marshland, Dharavi attracted migrants and settlers from across India, who worked hard and reinvested their savings into improving their housing and locality without gaining any clear legal title [10, p.47]. Over generations, settlers developed thriving fishing, tannery and textile industries, as well as associated supply chains and satellite businesses. With the land surrounding Dharavi now formally developed, Dharavi has become extremely densely populated, estimated to be around 20 times the population density of London, or 500 times the density of Miami [10]. The migrant history of Dharavi is reflected in its cultural and linguistic diversity. Less than half of Dharavi's residents hail from Maharashtra and speak Marathi, the official language of the State; about 30 % of residents came from the Gangetic plains and 20 % from Tamil Nadu, so "Marathi, Hindi, Tamil, and Urdu are the most commonly spoken languages in Dharavi" [64, p.45]. In this context, Hindi triumphs over Marathi (and English) as a unifying language that is "more or less understood by Dharavi's entire population and facilitates day-to-day communication" [64, p.45].

3 BACKGROUND

Research at the intersections of AI and HCI is a hot topic at CHI, eliciting critiques [13] and agenda-setting positions from some of the most prominent commentators and visionaries of our field [35, 68]. While we are mindful of these critical perspectives, and advance critical debates surrounding AI within our field and society writ large in our conclusion, we turn first and foremost to the work of scholars local to the communities we partnered with to situate our contribution. In Cape Town, the work of the linguist Ana Deumert and media scholar Marion Walton shows particular depth and breadth. And in Mumbai, we draw upon the pioneering work of Devanuj and Joshi to make technology more accessible to 'emergent users' and who identified text-input as a distinct barrier to digital participation [20].

3.1 Sociolinguistics & Materiality

Deumert's book on "Sociolinguistics and Mobile Communication" is particularly relevant as it focuses on digital communication and multi-lingual data and is grounded in ethnographically-informed case studies [18]. Given its unique emphasis on African perspectives and datasets, it is no surprise that creativity and inequality emerged as key themes of her research. Of course, drawing attention to creative and vernacular forms of design-in-use are pervasive and celebrated qualities of ethnographic HCI research [73]. On this point, Deumert advises "to look carefully at the kinds of creativities we see in new media environments in order to understand the possibilities for novelty as well as the constraints within which writers/speakers operate" [18, p.170].

Given her focus on African case-studies, Deumert furthermore points out that "the issue of inequality [. . .] matters whenever we write about mobile communication" [18, p.172]. Studying, theorising and intervening to address issues surrounding inequality are, of course, dominant genres within the kindred fields of ICT for Development (ICT4D) [36] and HCI for Development (HCI4D) [16] research that often surface and address questions surrounding the *material* conditions of digital access. Such a material perspective remains important in recognising issues that occur 'after access', but are nevertheless shaped it [23]. Here, South African media scholar Marion Walton operationalises the apposite term 'pavement internet' to critique the metaphor of mobile platforms, which in her view are not flat but highly unequal and render some user groups 'digitally invisible' [81]. Walton shows how high data costs associated with accessing streaming video platforms are a barrier to participation for economically marginalised users. Comments on streaming video content, such a videos associated with breaking news stories, often include requests to "plz watsapp it 4 m @ [phone number anonymised]"; this shows that such users "wanted to pass it to their own WhatsApp contacts or smuggle it via Bluetooth through the cracks of the pavement internet" [81].

HCI commentators are increasingly recognising that how information is materially represented shapes how it can be put to work, and that consequently the "material arrangements of information [. . .] matter significantly for our experience of information and information systems" [24, p.4]. Particularly for mobile telecommunication applications, it is a mistake to treat the text, voice and multimedia messages that people send as a purely immaterial digital form made of bits that encode information; that is, from an information theory perspective as articulated by Claude Shannon, the prevailing mythology of the digital realm. After all, as Richard Harper explains, "communication acts are not to be thought of as, say, a transfer of information [. . .] but as acts that alter the moral fabric of the relationship between the senders and the receivers" [34].

Such performative values, which are part and parcel of all communication acts, are important to consider. They come into play when we consider an additional form of inequality that Deumert reveals and that the HCI4D and ICT4D research communities are less familiar with: linguistic inequality.

Here, Deumert demonstrates that only few languages shape the linguistic diversity of virtual spaces in general. And concretely, there is a dearth of isiXhosa content online, and the laudable initiatives that aim to increase isiXhosa content online, often use a form of language that "is not a representation of a real, existing *language-in-use*"; in effect such initiatives can reproduce rather than challenge global inequalities [18]. In the Global North and in venues like CHI or Interspeech we may be keen to recognise, support and celebrate linguistic diversity and multilingualism, but can easily make the assumption that "people will necessarily want to access material in their first or 'native' language" [18, p.75]. To

avoid locking already marginalised people into one scale-level—the local [74]—Deumert's words of warning are critical to take to heart, even if they are difficult to hear:

> *However, in an increasingly global and interconnected world, where most people speak more than one language, such monoglot ideologies have little currency. This is particularly true for postcolonial societies where everyday multilingualism is the norm, and where the process of becoming literate is usually linked to acquiring proficiency in the former colonial language* [18, p.75].

The canonical text situated at the intersections of language, culture, identity, and coloniality is "Decolonising the Mind" by the great Kenyan writer and scholar Ngũgĩ wa Thiong'o [55]. Through his book, Ngũgĩ makes the impassioned argument that African mother-tongues are vehicles of culture and identity and to 'decolonise the mind' African literature ought to be written in African mother-tongues:

> *Language is thus inseparable from ourselves as a community of human beings with a specific form and character, a specific history, a specific relationship to the world* [55, p.15].

Our research is situated at the intersection of these contrasting perspectives, which we use as lenses to navigate a difficult and contentious terrain. But most of all, we empathise with people who navigate through these issues as they go about living their lives. And ultimately, we take our cues from them.

Research on SMS messaging practices in South Africa furthermore illustrates this dialectic. Poly-lingual users prefer texting and accessing online information in English, making use of predictive text as well as frequently abbreviate English words [19]. Less frequently, the same users send isiXhosa messages as a matter of pride and principle. Compared to English messages, these are harder to type without predictive-text dictionaries and in the context of sociolinguistic norms that discourage abbreviating isiXhosa [19]. These norms and choice of language not only resonate with Ngũgĩ's perspectives, but also point to the performative values of communication, namely those going beyond conveying information, for which an abbreviated English message would have sufficed.

3.2 WhatsApp, Voice Messaging & Text-Input Barriers

Before the arrival of widespread and low-cost messaging apps, unabbreviated isiXhosa messages often required message content to be spread across multiple SMS messages, which came at additional cost when compared to abbreviated English messages conveying the same information. Applications like WhatsApp have since drastically reduced costs, provided richer capabilities such as voice, photo, and video messaging, and ultimately transformed telecommunication practices across the Global South. For instance, in South Africa WhatsApp is now the most popular mobile app, with 58 % of all mobile phone owners using it [71]. Such statistics are also reflected in a diary study of less-connected mobile users across South Africa, where de Lanerolle et al. found that "the instant messaging application WhatsApp was by far the most frequently used Internet-based application" [15]. Widespread WhatsApp usage in South Africa is also reported among refugees [80] and financially-excluded entrepreneurs [42]. Further, an ethnographic study of chat at work in India and Kenya that, like our research, is situated across Sub-Saharan Africa and the Indian sub-continent, reveals widespread usage of WhatsApp [53]. In India this comes as no surprise, considering that there are almost half a billion Indian WhatsApp users[2]. Furthermore, WhatsApp is credited with removing barriers to digital participation in India specifically [4].

Of these varied contributions situated in the Global South, only de Lanerolle et al.'s diary study of less-connected South African users [15] and Balkrishan et al.'s Indian WhatsApp study [4] report that some users use WhatsApp for voice messaging. However, other than such passing references, research has not unpacked specific practices or motivations.

It is, however, generally accepted that text entry in Indian languages poses challenges [4]. Even looking back to the days of physical keyboards, typing in such languages was complex due to the structure of Indic scripts and large number of characters involved [41]. There have been advances in creating dynamic keyboards for mobiles, with good success. For instance, in trials, the Swarachakra keyboard [50] has proven to be an improvement over the standard Indian script keyboard layout (InScript) in terms of speed, accuracy and user ratings. However, research has shown that for some users, bad experiences using Indic language keyboards on phones in the past had led to them giving up and returning to the standard QWERTY Latin-script model [4]. Furthermore, the entry-level barrier of such Indic scripts is particularly prevalent for novice users [20, 30].

4 LANGA WORKSHOPS

Given the context and background discussed above, we now report on two co-design workshops we conducted in Langa. The aims of the Langa Workshops were two-fold: to explore and develop potential use-cases for an isiXhosa ASR system; and, to collect more representative data of isiXhosa 'language-in-use' [19]. Compared to unsupervised ASR approaches that require large datasets, we were interested in exploring whether smaller amounts of 'in-domain' data could instead be utilised for ASR development. That is, data which more accurately reflects the style, speed, locale or topic of conversation, and is therefore invaluable for testing and evaluating language models during development.

We partnered with an experienced workshop facilitator and translator, local to Langa, with whom we have a long history of collaboration. The facilitator recruited participants, ensured that design activities followed local COVID-19 restrictions, and fed back on workshop plans, in particular to keep discussions and activities focused. This latter suggestion also matched the appeal from the ASR researchers on our team: to identify and focus on particular "domains of speech" and to, at least initially, avoid more general-purpose and open-ended speech-driven Intelligent Personal Assistant (IPA) application areas, such as those found in Google's Assistant or Apple's Siri.

[2]https://techcrunch.com/2021/01/11/youtube-and-whatsapp-inch-closer-to-half-a-billion-users-in-india/

Table 1: Workshop 1 isiXhosa participant demographics.

ID	Gender	Age	ID	Gender	Age
M1	Male	26	F1	Female	39
M2	Male	41	F2	Female	38
M3	Male	24	F3	Female	21
M4	Male	33	F4	Female	43
M5	Male	22	F5	Female	20
M6	Male	21	F6	Female	36

4.1 W1: Exploring Use-Cases & Data-Collection

Before the first Langa workshop, the HCI, ASR and facilitator team collaboratively developed two initial use-cases and discussion points that leveraged a user-engagement and data-collection probe, as outlined below.

The first use-case centred around a simple IPA system to support isiXhosa voice reminders. We hoped that this focused use-case would present an interesting but also delineated 'domain of speech'—times, people, places and activities—and convey to participants some of the core capabilities of an isiXhosa ASR & NLP system: the ability to transcribe spoken language and respond to well-structured spoken commands.

The second use-case explored a simple information retrieval example whereby an ASR-enabled system could surface isiXhosa mobile media content in response to spoken queries. We posited that this use-case might reveal a speech domain of entertainment and play [28], as well as shedding light on some of the isiXhosa media content that participants store, access, and share on their mobiles (cf. [81]).

Finally, we also wanted the workshop to provide space for participants to share their own ideas and potential use-cases.

4.1.1 Participants. Twelve (six male; six female) isiXhosa-speaking residents of Langa (see Table 1) and the nearby township of Khayelitsha participated in the workshop. All participants were recruited by the local facilitator using the following criteria: first-language isiXhosa-speaker, mixed age, mixed gender, active user of WhatsApp. After obtaining their consent, participants were asked to engage in discussions and activities using the WhatsApp Probe described below. Participants received a data bundle and were compensated an appropriate amount for their time, which was determined by the facilitator.

4.1.2 WhatsApp Data-Collection & Engagement Probe. We were inspired by Lambton-Howard et al.'s notion of 'Unplatformed Design' that "frames existing platforms as material, with material qualities" and that platforms like WhatsApp "can be appropriated and designed with" [46]. Not only does this notion resonate with the material perspective we developed earlier, but it allowed us to leverage the fact that WhatsApp is already installed and used by many people, especially in South Africa and India, thus lowering the barrier to participation. It is furthermore an authentic form of expression – it is what people use to communicate with their friends, family, and community. There is also precedent for using WhatsApp as a data-collection and engagement platform, such as in Vuningoma et al.'s research with refugees in South Africa [80] or Kaur et al.'s research with pregnant women in India [43].

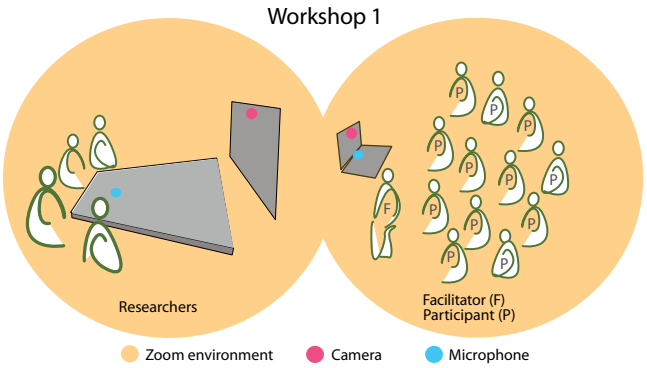

Figure 2: The remote setup used during Workshop 1, with researchers shown on the left and workshop participants and facilitator to the right.

Given the familiarity and authenticity of WhatsApp as a platform, we were eager to explore if the probe would be a viable alternative to more traditional approaches to collecting speech data, such as in-person or in-studio recordings that are increasingly difficult to conduct due to the COVID-19 pandemic. More novel approaches to collecting speech data, such as from interactive voice forums [63], were infeasible at scale because of the high cost of voice calls in South Africa. Crowd-sourced approaches have furthermore previously been shown to be better suited to gathering speech data on delineated topics (e.g., recording survey responses, repeating pre-defined sentences, saying isolated digits [1, 62]).

4.1.3 Setup. Within our workshop the local facilitator matched six pairs of participants and created a WhatsApp group for each pair named Group 1–6 respectively. In addition to the participant pairs, each group also included the facilitator and a researcher. For the data-collection and engagement activities, we asked participants to create isiXhosa voice recordings and share media items to the group. As part of the consent process, and again before each activity, we reminded participants that all content shared within the WhatsApp group would be transcribed, analysed and used to test and train isiXhosa ASR models.

Due to ongoing COVID-19 travel restrictions, the workshop was conducted through a Zoom teleconferencing link (see Fig. 2) to connect co-located researchers in one location with co-located workshop participants and facilitator in Langa. Both researchers and facilitators were responsible for adhering to COVID-19 regulations in their respective places. At the beginning of the workshop we also obtained participant consent and permission to record the workshop. Participants were encouraged to respond in their choice of either isiXhosa or English, and so the facilitator also translated participants' comments during the workshop. Finally, an isiXhosa-speaking note-taker also audited the workshop, with whom we subsequently shared the Zoom recording. After the workshop we exported all group chat content, from which we extracted, transcribed, and translated voice recordings.

4.1.4 Activities. The first two design activities centred around the voice reminder and media retrieval use-cases. Each began with a discussion about participants' current practices, followed by a hands-on activity using the probe where participants would alternate driving the interaction using a voice message shared to their WhatsApp group and with their paired partner subsequently generating a simulated appropriate system response, again using a voice recording. This is an adapted, participatory twist to Wizard of Oz approaches commonly used to showcase and demonstrate novel voice user interfaces [12]. For the data-collection side, we hoped that pairing queries—for media items or to remind the user of something—with simulated responses would generate a useful, multi-speaker dataset, that was recorded on participants' own devices and therefore exhibits similar audio fidelity and noise to those an ASR system would be confronted with when deployed.

For the *reminder* use-case, we began by discussing what tools—if any—participants currently use to remember things, to reflect on what sorts of things they need to remind themselves of, and in what language. We then asked participants to use their respective WhatsApp groups to create isiXhosa reminders of things they presently want to remind themselves of, or have needed to remind themselves of recently. For instance, "remember to bring an umbrella tomorrow as it will be raining". The paired participant was then instructed to respond as an IPA would, restating the voice reminder: "OK, I'll remind you tomorrow to bring an umbrella". Participants were asked to alternate roles for the next 20 minutes. Before taking a mid-morning break, we asked all participants to create a perfect-as-possible transcription of the most interesting voice note that their partner created, listening as many times as necessary.

For the isiXhosa *mobile media* use-case, we asked participants how they find isiXhosa content to look at or listen to on their mobiles; what type of content they access (e.g., pictures/memes, audio, videos/GIFs, etc.); and, how they access, store, and share their content (e.g., through links and streaming, or by downloading, uploading, 'Bluetoothing,' etc.). After learning about their current practices, we asked participants to share examples of isiXhosa media (e.g., a music video from popular hip-hop artist Zanzolo) in their respective groups and for the receiver to then craft a corresponding spoken language query that might return the media item as a result: "may I have Zanzolo music videos". We again asked participants to alternate roles for the next 20 minutes.

Before breaking for lunch, we asked participants to think about situations where they used voice recordings and voice messaging or wish they had recorded them. After lunch, we engaged in lively discussions on this and other topics, where participants were also encouraged to envisage scenarios and use-cases of their own.

4.1.5 Results & Reflections. For the reminder use-case, participants reported using a variety of digital and physical tools to help them remember things. On the mobile this was mostly through built-in phone functions (e.g., contacts, calendar, notes) and bespoke reminder and todo apps. One participant described herself as "old school" as she still makes use of paper diaries and not her mobile device to set reminders. Another shared how he records his daily reminders in bullet-points and not full sentences. For example, he would create a calendar entry titled "gym" at a specific time to remind himself to work out. Eleven of the twelve participants shared that they prefer recording their daily reminders in English; however, participants also agreed that there are times when they use both English and isiXhosa to record their reminders.

One participant shared that she records her family events and meeting reminders in isiXhosa as it allows her to thoroughly articulate and show respect to the meaning of the event (such as a funeral or circumcision ceremony) set to take place. However, the same participant also stated that it is easier to record community meeting notes in English. In contrast, another participant articulated his clear preference for English:

> *Putting a reminder on my diary or phone in isiXhosa means I'm still living in the past. I'm young and I need to move with the times, so my reminders need to be in English. I'm imagining if it says that I'm going to the Eastern Cape in isiXhosa that's kind of embarrassing. I cannot do it in isiXhosa, because I'm young.*

These discussions demonstrate the contentious terrain that language represents in post-colonial and post-apartheid South Africa (see Sections 2 and 3). The clear preference for English, even for those who sometimes created isiXhosa reminders, also meant that participants did not complete the data-collection exercise that followed in the ways that we anticipated they might. Most participant pairs ended up simply chatting with one another via voice messaging after sending and responding to a few voice reminders. Here we benefited from our 'unplatformed' [46] WhatsApp probe and the authentic and familiar form of expression it presented to participants. These messages consequently yielded a rich dataset, which the facilitator and note-taker later transcribed and translated into English so all of the research team could review content. An example of one such message is shown in Table 2, demonstrating creative language use by multilingual speakers, who use elements of both English and isiXhosa syntax[3] and phonology[4] as they converse with one another, a process referred to as code-switching [69]. However, compared to the simpler recordings we were expecting, such as "remind me to bring an umbrella," these conversational recordings also proved more difficult to transcribe by the workshop note-taker in order to form part of the ASR test dataset (see Section 6).

The media retrieval discussion revealed that cost of access remains a barrier to online participation. For instance, participants mentioned finding and sharing funny content on their Facebook feeds and specific local or isiXhosa-speaking groups. At the same time, the set of practices that surround Walton's apposite term 'pavement internet' (cf. [81]) remain prevalent. Participants mentioned taking screenshots of jokes or memes—taking the content off the Facebook platform—and sharing them with people or groups of people via WhatsApp. They also download content to their phones when they have access to free WiFi, sometimes supplied through government initiatives. Once content is taken offline (or if it is created directly on their phones), 25 % of participants said they recently sent or received media via Bluetooth, while 75 % of participants recently used SHAREit, a cross-platform file sharing app that leverages WiFi connections which is faster, but also less reliable than Bluetooth according to participants.

[3] Structural properties of language, such as arrangement of words and phrases to create well-formed sentences.
[4] Organisation and inventories of sounds in a language.

Table 2: Original transcription and translation of a WhatsApp voice message showcasing dynamic and creative code-switching between isiXhosa (upright text) and English (italic text).

isiXhosa/English Transcript	English Translation
Masithi *3 o'clock* ke e*Clocktower*. Mamela kyk hier ndiyamazi *I know him, I got him* [INAUDIBLE]. Ndizithi kuye masiye eWaterfront *I won't tell him that I'm meeting a friend, but* ndiyayazi *he won't mind* xasidibana nawe. *He will buy us drinks and some lunch then* sonwabe wethu.	Lets say *3 o'clock* then at the *Clocktower*. Listen, look here, *I know him, I know him, I got him* [INAUDIBLE]. I will say to him let's go to *Waterfront, I won't tell him that I'm meeting a friend, but* I know *he won't mind* when we meet up with you. *He will buy us drinks and some lunch then* we'll have fun man.

We collected less data during the media retrieval exercises in comparison to the earlier reminder exercise. This was in large part due to participants adhering to the initial task description of sharing and generating spoken queries for isiXhosa media items. Not only did finding content on their phones take time, but participants also concurrently uploaded that content to the paired WhatsApp groups using the WiFi connection at the workshop venue, which became overloaded. The screenshots, images, memes, music and videos that participants shared pointed at a rich isiXhosa mobile media ecology that exists largely outside of platforms such as Facebook, Twitter, YouTube, TikTok and so on that dominate the hyperdeveloped world [73].

On the topic of recording things in everyday life, participants immediately mentioned voice messaging. All participants reported that they send voice messages every day, without fail. One mentioned that it is convenient way to communicate, as they do not have to worry about misspelling words or whether the person they are sending the message to will understand it. According to the participants, who often return to more rural areas in the Eastern Cape during holidays and for ceremonies, voice messaging is even more widespread there. In this sense, voice messaging was seen as positive and equalising, broadening participation and facilitating communication within the participant's local social circles, but also with family in rural areas.

Participants also reported that it can be tricky keeping up with the voice messaging activity in larger WhatsApp groups including those local to Langa, where the majority of members send voice (rather than text) messages. Although they appreciated the key benefit of voice messaging, namely that it allows everyone to clearly understand the message, finding older voice messages proved challenging to participants. We also observed workshop participants often deleting voice messages immediately after sending them, which we later learned was because people wanted to re-record the messages. As there is no way of listening to a voice message before sending it, nor effective ways of finding older voice messages, this suggests that current voice messaging user interfaces such as WhatsApp do not adequately support the needs of this particular user group.

Given that our initial use-cases engaged, but did not inspire, participants, we reflected with them on the workshop and potential next steps. Overall, participants appreciated the exposure to Zoom and the focus on thinking about application areas of advanced technologies in their context. With regard to ASR application areas, the facilitator remarked after consulting the group of participants: "Voice Messaging is the way to go".

4.2 Materiality of Mobile Voice vs. Text Messaging

From this first workshop we could see how participants are already balancing and leveraging the different material qualities of voice messages and text messages: they appreciated how the modality of voice gave everybody equal chance to participate, but conceded that it can be difficult to keep on top of messages, especially in larger groups. Of course, the asynchronous nature of voice messages is itself a property that was the exclusive provenance of written forms [58] and demonstrates how voice messages are an emerging hybrid form of communication. However, in the mobile domain text still reigns supreme, for it can be edited, searched, copy-pasted; it is data-light, requiring little bandwidth to send or share, and is easier to back-up at low cost. A series of text messages, or a long message, can also be quickly skimmed, which is far harder with audio. Of course, text-input, especially on mobile devices, can be cumbersome. However augmenting text-input with AI (e.g., auto-correct, auto-suggest, or voice keyboards) can alleviate some of these issues, albeit with varying success [35] and not for all languages.

We could not find any research on voice messaging practices of users in the Global South, except for a Brazilian case-study of a publicly available dataset of a large WhatsApp group that was analysed to assess how (mis-)information spreads through the network [51]. Research on avid voice messaging users in the US has shown how "voice messages can be recorded faster than typing a text message, allow for greater expressiveness, and deeper emotional bonding", but also concedes that they are cumbersome to review and edit, tedious to scan, and inherently public during recording and playback (unless using headphones) [33]. In the US at least, voice messages carried a connotation as they "shift the effort necessary for communication towards the receiver, when compared to texting", and are a more niche practice reserved for communicating with a low number of intimate friends, family, or partners. In contrast, our participants revealed that voice messaging is not only pervasive, but that they are happy to take on that extra effort required, as it allows everyone to participate. Bidwell et al.'s research, though conducted before WhatsApp released voice messaging, shows how prototypes that support recording and sharing asynchronous voice recordings in a rural South African community, and on a bespoke application running on a communally owned tablet, were especially popular with women, despite men controlling access [7].

Creative appropriations of technology are familiar terrain in South Africa. Although not specific to South Africa, intentionally missed calls are often used to convey information (e.g., "I'll give you a missed call when I arrive"), but are also used to signal and

reinforce connection between intimate contacts [22]. A uniquely South African phenomena at the time, people leveraged USSD-based 'callbacks' to send highly constrained messages at no cost. Intended by cell phone networks to allow people without airtime to request a return call from a contact who did have airtime credit available, users appropriated the service to send very short messages, such as "ME.N.U.4EVER", which is delivered to the contact's phone as "Please call ME.N.U.4EVER"; a previously unreported and under-designed-for practice that Bidwell et al. reveal in their eponymous paper [6].

We believe that the pervasive, situationally aware and inclusive practices surrounding voice messaging that participants reported and we engage with here deserve similar attention from CHI. To further that aim and rather than pursuing our initial imagined use-cases, we conducted a follow-up workshop focused specifically on voice messaging and involving a bespoke probe that exemplify ASR capabilities in relation to WhatsApp voice messaging. At the same time, we transcribed the data we had collected in order to develop, test and refine an isiXhosa ASR model, reported later in this paper (see Section 6).

4.3 WS2: WhatsApp Voice Messaging & ASR Probe

The second Langa workshop centred around a bespoke ASR probe, consisting of an Android app that connects to a cloud service (see Fig. 3). After engaging in a series of discussions surrounding voice messaging practices on WhatsApp, we leveraged the probe to allow participants to explore ASR capabilities in relation to their voice messages, a topic we explored collaboratively.

4.3.1 The ASR Probe. The Android app that we developed—*Voice Notes*—can receive and process audio data from other apps, including WhatsApp (see Fig. 3). When users long-press on a voice message and click the share icon in a WhatsApp group or conversation, they can then select the Voice Notes app from a list of Share Targets that usually includes other common apps, such as Email, SHAREit, Bluetooth, etc.. The app creates a copy of the voice message's content and uploads it to our cloud service. The cloud service then creates an asynchronous transcription request using Google's Cloud Speech-to-Text service using the English (South African) language model, and the transcription results are sent back to the app. Audio recordings are subsequently deleted from the server, but transcripts are stored in a database that is kept synchronised with the app on a per user basis. However, the research team has committed to not look at or analyse the database in order to preserve user privacy. The duration of the entire transcription process is roughly proportional to the length of the submitted voice message. The voice notes with their corresponding transcripts are presented to the user in a scrollable and searchable list, and clicking on an item takes the user to a details screen where they can see the transcript or listen to the original recording, as well as delete and share the voice message and transcript.

Since the act of sharing a voice message is deliberate, and it is clear to the user at the point of interaction what content will be transcribed, no special permissions are required by the app. This contrasts with data collection apps other researchers have developed in order to analyse WhatsApp data, which require users to grant expansive permissions to the entire WhatsApp folder, causing issues with finding people who will agree to participate given the personal and sensitive nature of chat histories despite data protection assurances [33]. A drawback of our approach is that only the audio content stream is accessible, and not any metadata of the voice recording, such as the date it was created, if it was sent or received, and in the context of which conversation or group. This has the knock-on effect that, in order to be able to robustly handle transcription processing and support playback or re-sharing, the Voice Notes app has to create a copy of the recording, which takes up storage space. While we benefit from the authentic form of expression that WhatsApp represents to participants, the drawback of such an 'unplatformed' approach [46], is that it's primary use was never a research platform, and obtaining and exporting data is more cumbersome and prone to gaps such as missing metadata.

4.3.2 Participants & Setup. Twelve (six male; six female) isiXhosa-speaking residents of Langa participated in the second workshop, using the same recruitment process and criteria as the first workshop. Half of the participants had previously taken part in the first workshop. All participants were active users of WhatsApp with an Android phone and, after obtaining their consent, were asked to engage in discussions and experiment with the ASR Probe app. Participants received a data bundle and were again compensated an appropriate amount for their time, which was determined by the facilitator.

To adhere to stricter COVID-19 restrictions in South Africa, the workshop was again enabled through a Zoom teleconferencing link to connect co-located researchers in one location with participants and the facilitator in Langa. However, unlike the first workshop, participants could not be co-located and so each participant connected to the Zoom meeting on their own. Accordingly, before the workshop we organised phone-calls and practice sessions for the participants to familiarise themselves with the Zoom interface and turn-taking (e.g., mute/unmute, raising hand, etc.). At the beginning of the workshop we also obtained participant consent and permission to record the event. Participants were again encouraged to respond in either isiXhosa or English, translated by the facilitator and observed by the isiXhosa-speaking note-taker.

Participants used their phones—their primary computing and internet access device—to connect to Zoom, and while on the Zoom call could not, for instance, check a WhatsApp conversation to look up the answer to our questions (e.g., how many participants are in a typical group they are a member of, and in what language did they send their most recent voice message?). They also could not use the ASR Probe while on the Zoom call, so after the lunch break we sent video instructions and voice message samples to a WhatsApp group consisting of everyone in the Zoom meeting. Although the pragmatics of our remote workshop setup were more challenging than the first workshop, participants mentioned to us afterwards that they benefited from the experience and could practice and familiarise themselves with Zoom and its unnatural turn-taking mechanics.

4.4 Results & Reflections

Participants reported that voice messages are easier and quicker to send, and that they use voice, rather than text messages when they

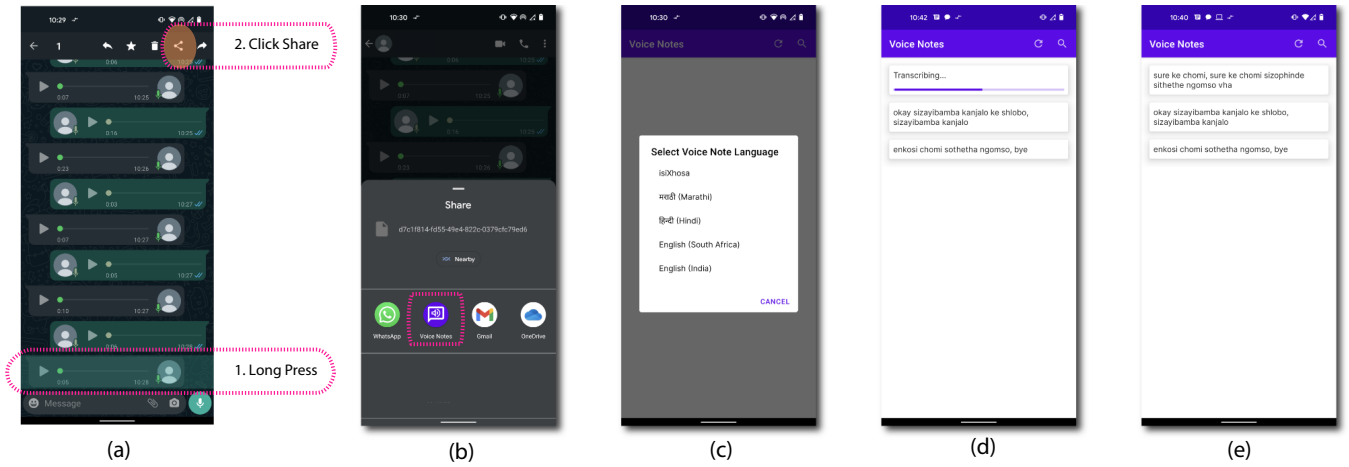

Figure 3: Interaction flow of the ASR Probe: users select a WhatsApp voice message (a1), tap the share icon (a2), and (b) select the Voice Notes app. (c) Users then select a transcription language: English (Langa Workshop), isiXhosa (Langa Study), or Marathi (Dharavi Study). The app then (d) uploads and (e) transcribes the voice messages.

are in a rush or conveying a message or sentiment that is harder to explain through typing. Instructions and directions were seen as particularly suitable to voice messages, which contrasts to the reported usage habits of US-based users [33]. Tone and emotion were also discussed – through voice messages participants could for example more clearly express anger, or whisper if they were gossiping.

While recipient design [29] had the strongest influence over language choice, participants mostly used isiXhosa in their voice messages and usually only reverted to English if the recipient could not understand isiXhosa, or they wanted to keep the conversation short; isiXhosa sentences are often three times longer than their English equivalents [7]. This was especially the case when communicating with family, although one participant reported mixing languages in a family group chat. Another participant also reported additionally sending SeSotho and SeTswana to friends.

Although many voice messages fade away into the chat history, never to be listened to again, all participants reported that they do revisit some voice messages, either 'in the moment' to check that the voice message they just sent contained all the right information, or to revisit older content. For example, participants reported recording and later relistening to meetings or church sermons through voice messages. They might also reference older voice messages if someone backtracks on what they said, or there is a need to check particular directions or instructions. One participant regularly keeps the voice messages of the person they are dating and re-listens to them again as a reminder of their voice.

To find an older voice message requires persistence: "I go to my WhatsApp chat and scroll up and keep listening to all the voice messages until I locate the one I'm looking for". Another participant lamented that "finding a VN is so tough because I have to scroll all the way up, past everyone's messages just to find that one VN". They agreed that remembering the approximate date and sender helps locate message. Another strategy employed is to search for a text message that was sent at around the same time as the voice message to at least be able to jump to a more promising starting point and continue their search from there. Those that are shared in a group with many voice messages are far harder to find; participants agreed that it is not worth the effort required to find voice messages older than about three months. Participants also need to routinely clear space on their phone to make space for new content, and often cannot afford the cost of cloud backups (or the required internet connectivity), so older voice messages may have already been deleted. One particularly tech-savvy participant reported a creative workaround, where he accesses voice messages through the 'Files' app on his phone. From here it is easier to sort all voice recordings by date and listen to them. Once the date has been identified, he returns to WhatsApp to select the correct message in order to, for instance, reply to it or hold someone to account.

Over lunch we asked participants to install the Voice Notes ASR Probe on their phones. We also sent video recordings introducing the app, how it works, and the data that it collects, emphasising that we do not listen to the voice messages, nor read message transcripts generated by the ASR system. We also sent sample voice messages to the WhatsApp group and encouraged participants to practice with transcribing these examples. We then asked participants to experiment transcribing some of their own English messages before reconvening the Zoom workshop which participants again accessed using their phones.

Participants felt that it would be useful if the app could translate between different languages of South Africa, and expressed some interest in automating transcriptions within WhatsApp. They also enquired about how much data the app requires, to which we responded that it was about the same amount as sending or receiving a voice message when initially transcribing, and negligible amounts to refresh the list of transcripts. Reflecting on how the app might help them, participants immediately recognised the value of being able to 'listen' to a voice message privately by looking at the transcript rather than listening when in a public setting. Others mentioned that by looking at the transcript of a voice message that

they had just sent, they could see if they made a mistake or error. A transcript could also be helpful for someone who does not understand a word—either because it is difficult to hear or the meaning is unknown—they could see the word clearly in the transcription and understand it better or look up its meaning elsewhere. Multiple participants suggested that the app could furthermore be helpful when people have unfamiliar accents that were hard to understand or follow.

Given the current state-of-the art of ASR systems, we might dismiss many of these remarks as unrealistic, and focus only on the use-case that is nearest at hand: to support looking at a voice message while in public. However, if we pause and look at these comments through a different lens, we can see that they are also visions of future AI capabilities that are similar to the rhetoric, hype, and promise that surround many contemporary AI products and services, which also extend far past their current capabilities [27]. We might instead, draw inspiration from allied efforts to diversify future-making [59] or reconstitute utopian visions in other domains of computing [49], and recognise the importance of creatively engaging with specific, local instances that taken together can create global forms.

To highlight the research that needs to be done at the intersections of HCI & ASR research, we deliberately juxtapose these ambitious proposals for the ASR-enabled systems that participants articulated with our comparatively modest efforts to develop and put ASR-enabled technologies in the hands of people.

5 DHARAVI STUDY & MARATHI ASR SYSTEM

While we concurrently developed an isiXhosa Language Model (see Section 6), we drew on a longstanding community partnership with Marathi-speaking residents in Dharavi to provide a comparative perspective. Scholars of everyday practices (e.g., [47]) have long recognised the value of comparative study and assessment, particularly for phenomena, such as language [18], chat [53] or digital infrastructures [24], that are largely taken-for-granted. We took the opportunity to deploy the best performing Marathi model [45] that was developed for the low-resource Indian language challenge at MUCS 2021 [21] in a real-world setting. Following the principles of the 'itinerative design' methodology [59], which demonstrates the benefits of pivoting between different communities in the context of a research project, we did not repeat all phases, but picked up the line of research from its current point in South Africa, albeit at a slightly smaller scale. So we also invited Marathi-speakers in Dharavi to reflect on their voice and text messaging practices and experiment with the ASR Probe, configured with a Marathi ASR system based on the MUCS competition-winning Marathi model. The differentiating factor of that particular language model is that it augmented training data supplied by the competition with crawled data from YouTube to improve performance [45]. This technique was successful for Marathi, not only because there are 100 million more people that speak Marathi than isiXhosa, but also because mobile data is an order of magnitude more affordable in India ($2.50 per month for 42GB of prepaid data) than in South Africa ($19 per month for 3GB of prepaid data). While the low-resource moniker refers to the dearth of readily available transcribed training data, everyday Marathi voices—in the form of YouTube video, articles, or user-generated text and speech content—are much better represented in online spaces. The deployed Marathi ASR model was not tweaked further, had a word-error-rate of 15.79, and produced Devanagari script as its output [45].

Table 3: Marathi participant demographics.

ID	Gender	Age	ID	Gender	Age
M1	Male	58	F1	Female	36
M2	Male	31	F2	Female	28
M3	Male	34	F3	Female	32
M4	Male	48	F4	Female	35
M5	Male	21	F5	Female	32

5.1 Participants & Method

Ten (five male, five female; see Table 3) Marathi-speaking residents of Dharavi participated in this study, recruited through a community liaison who also determined an appropriate participation compensation. As the study was conducted remotely, we individually gave participants a brief about the research and demonstrated the use of the Voice Notes app using a video. After watching the video, an 11th participant who had initially volunteered decided not to proceed with the study; the remaining ten participants consented. We conducted phone-based semi-structured interviews about voice note practices lasting approximately 20 minutes, then shared the Voice Notes app and assisted participants to install it.

We asked participants to experiment with the Voice Notes app for a period of one week and reminded them that their voice messages and transcripts are entirely private, unless they choose to share them with the research team during subsequent interviews or by forwarding individual voice messages and transcript screenshots via WhatsApp. A week later we followed up with participants to ask about usage impressions and reflections.

5.2 Initial Interviews

Initial interview questions revolved around mobile voice and text messaging practices and language preferences. Speed and ease of use were themes that participants immediately recognised: e.g., "in one minute we can talk a lot more than if we type text" [M4]. However, for one participant this also intertwined with being inclusive: "my uncle and sister are not educated – I don't get a reply fast if I send typed messages. I do get an immediate reply if I send a voice note" [F1]. Installing, configuring and typing on a Marathi-language keyboard and/or switching between different languages on Indic script keyboard requires more advanced digital skill-sets. Other participants reinforced the notion that voice messages remove the barrier of text input (cf. [4])—"my father has a mobile recharging shop – many of his customers are illiterate, so he uses voice notes"—but also recognised that even voice messages require some digital skills: "many people don't know how to use voice notes in WhatsApp – they touch [the record button] and leave it. [They] don't know the idea of long press" [F3]. Another participant remarked that people better understand voice messages, or they send voice notes "when I need to precisely communicate messages with colleagues" [F4]. Finally, participants also reported

issues finding salient information within voice messages compared to text messages – for instance, "recently during the [COVID-19] vaccination drive, I received a voice note with the gate number of the hospital where the vaccination centre was set up, so I had to search a lot to find the gate" [F3].

Participants also reflected on how they dynamically negotiate whether to converse in Marathi or Hindi. For instance, a nascent conversation might begin in Marathi, but then shift to Hindi if the Marathi-speaker detects that their partner is more comfortable speaking Hindi. The opposite phenomena also occurs, whereby a conversation begins in Hindi but then is continued in Marathi if the participants realise that both are native Marathi speakers. Further, while Dharavi is overall culturally and linguistically diverse, living areas are often grouped along linguistic lines, so Marathi-speaking residents will generally prefer to speak Marathi especially in their home environments and only revert to Hindi to accommodate others [64]. However, once a language is settled upon, this tends to remain.

5.3 ASR Probe Results

Participants generally saw the value of accessing transcripts. A participant who works as an educator said that "as the classes are now online, being a primary school coordinator, I am overwhelmed by voice notes from children" [F5]; along with another participant, they asked if we could keep the app running beyond the trial period, which we agreed to. The educator also reflected on how she currently records voice messages for her pupils that she then sends to their parent's mobile phone. However, she also creates a short text message summarising the content, which she hopes will ensure that busy parents won't miss the message, but still allows children to benefit from the increased accessibility of the voice message. Listening discreetly was also seen as valuable. For instance another participant remarked: "I was in a meeting and got a message from my supervisor. I use the app to check what is in the voice recording" [M3].

Multiple participants used the term 'dictionary,' when suggesting ideas about how to improve the app. They recognised that most commonly-used words are correctly transcribed, but certain words that are not quite mainstream, though still commonly used in Dharavi, are missed: "not-so-common words like 'Sahishnuta' [Marathi: tolerance] are not available on the app; only very frequently used words are available". Accents and pronunciations also affected transcription accuracy according to participants – for instance, one participant said that they tried re-creating a voice message: "if I try to speak slowly with the right enunciation, it works perfectly" [M2]. Another participant remarked that transcriptions were of better quality for short instructions in comparison to a longer voice message containing a reading from a poem that was not transcribed successfully.

Finally, one participant requested a way of recovering some of the metadata that is lost when the voice message is moved between WhatsApp and the Voice Notes app: "if I can add a keyword or a note to the transcribed file then it would be very helpful, as I won't be remembering the text content while searching for it weeks later, but I will remember the person or keyword I have added" [F2].

6 THE ISIXHOSA ASR MODEL & LANGA TRIAL

Concurrent to the Dharavi studies, our multidisciplinary team, led by ASR researchers and language experts, developed an isiXhosa ASR system. ASR development was guided by the findings from the Langa workshops and the high-level goal of supporting the transcription of WhatsApp voice messages, and the need to cope with everyday speech – that is, fast-paced, conversational and code-switching speech that reflects real 'language-in-use' [19]. It is important to acknowledge here however that unexpected challenges (discussed henceforth) meant that the pace of initial Xhosa model development was slower than expected: as such the model we deployed in Langa for the trial was not necessarily one that yet completely fulfilled all of these capabilities. We nevertheless expand upon the details of this deployed baseline model and, importantly, dissect the challenges we met surrounding data and transcriptions which led to the discrepancy between baseline and goal system. We also offer reflections on this challenge, which illuminate further areas of research at the intersection of HCI/ASR.

6.1 Baseline isiXhosa model

While precise technical ASR details are beyond this contribution's scope, for readers with an ASR background, the isiXhosa ASR model is a 'hybrid' one that uses a factorised time delay neural network (TDNN) [60] and is built using the Kaldi toolkit [61] with MFCCs and iVectors as input features and without grapheme-to-phoneme conversion. Hybrid models were chosen over end-to-end models because the latter are known to struggle with limited amounts of data, and are notably more black-box-like, so it is harder to adapt or tune individual parts to the use-case.

To train our isiXhosa model we utilised the NCHLT SA speech corpus [5] which, just like the data for the Marathi model [45], consisted of 50 hours of read speech. This dataset has previously been used in the isiXhosa ASR literature (cf. [9, 39, 75]). While the NCHLT data matches our use-case in terms of 'recording environment' (recorded on a phone), read speech is generally slower-paced and differs extensively from the more informal, fast-paced conversational speech that one would expect from voice messaging between community members, and also does not feature code-switching. Necessarily, then, this places a ceiling on what kind of quality of acoustic model can be achieved when training on read data-sets, which also lack 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. Furthermore, the dataset we collected during the Langa workshop was comparatively modest in size (see Table 4), so we decided to use it as testing rather than training data. Finally, although it might have been possible to tweak existing models of another language within the family of Bantu languages (e.g., isiZulu) to isiXhosa, we did not try this approach as we did not have access to a high-quality isiZulu model (those available, such as Google's, are not open); and, our attempts to incorporate data from the 'nearby' isiZulu and Sesotho languages had previously failed to improve model performance.

The mobile media samples we learned about during the first Langa workshop and that Walton refers to as content that is trafficked through the pavement internet (cf. [81]) are the exception rather than the rule in publicly-available isiXhosa content. Namely, what isiXhosa content that is available online is not reflective of actual language use [18], and the little publicly-available supervised data that has been previously collected for low-resource languages is often mismatched in domain to real-world use cases. Consequently, augmenting supervised data through Commoncrawl[5]—typically a useful source of language-specific text data—or by scraping YouTube for raw audio data revealed not only similarly mismatched data but was also particularly noisy for isiXhosa, as it features a lot of 'junk' and out-of-language content.

The challenges we encountered developing the isiXhosa Language Model (LM) are representative writ large of the issues ASR researchers encounter when working on low-resource languages. For instance Both isiXhosa and Marathi exhibit linguistic properties that are not found in the 'high-resource' languages for which ASR systems are typically developed. That is, models or approaches labelled 'state-of-the-art' when recognising English may not be directly applicable to low-resource languages without considerable re-working, re-tuning or even re-imagining of entire architectures. For instance, in isiXhosa the letters 'q', 'k' & 'c' are pronounced as different clicks and contain phonemes and sounds that are considerably different in articulatory features from any phoneme found in English. It is unclear how such features affect model performance and how they might be accounted for.

Both isiXhosa and Marathi also exhibit agglutinative morphology [32, 44]: what would often constitute separate words in, say, English are strung together to form longer words. Because of this, the isiXhosa model we created had to be modelled at the sub-word rather than word level. There is a fairly rich inventory of 'code-switching' ASR literature [69]. However, much of the earlier work focuses on languages for which large 'code-switch datasets' could be gathered (e.g., Mandarin-English [79]), or utilises existing linguistic tools (syntactic parsers; part-of-speech taggers) for both languages [2, 3]. Others have exploited 'translation'-based methodologies to some success; however, these methods rely on the existence of translation dictionaries [11] or large parallel corpora for the two languages [48, 84]. All these methods then are somewhat at odds with our overarching aim to develop approaches that could work across 'low-resource' contexts. And whilst many recent papers on code-switching ASR have advocated for 'end-to-end' approaches (e.g., for Mandarin-English [83, 86]), these too are known to be unstable when applied to smaller datasets that characterise 'low-resource' ASR. This 'linguistic-property' discrepancy extends to the level of language-in-use too; consider Section 3's discussion of "everyday multilingualism", and Section 4's first-hand evidence of such 'code-switching' naturally occurring in isiXhosa speech. If we are to produce an ASR model which recognises language as it is actually spoken in voice messages, then, this switching between languages at the sentence or even word level (cf. Table 2 'eClocktower') must be accounted for. But understanding how to model this code-switching in an ASR model is not trivial. Our isiXhosa model currently employs only a very basic code-switching solution: we trained separate English LMs and isiXhosa LMs on monolingual data and interpolated. This approach will not adequately model the nuances of when switches between languages are more likely to occur; for this, we really must consider more subtle linguistic and structural factors, and also the historical and socio-cultural aspects (cf. Section 4's discussion on English being more for the 'young'). Especially in post-apartheid South Africa, the unequal power dynamics between languages, who speaks them, and where they are spoken also play into what vocabulary is taken from which language. We can therefore already posit that our Marathi model will perform better on account of code-switching being less prevalent in Dharavi.

We had initially hoped that the Langa workshops would provide a rich 'in-context' data-set that we could utilise as a robust test data-set to evaluate against. However, we found that though audio from the workshops was collected successfully (see Table 4), obtaining accurate transcriptions for this data was a lot harder than expected (due to the more conversational nature of the discussion; participants' non-familiarity with the task of 'transcribing'; a lack of clarity regarding exactly what was being asked of the participants with regards to producing transcripts). Recall that we had paid the workshop note-taker to take on this task, but we were also expecting participants to produce simpler and shorter voice messages containing media queries and reminders. The conversational and code-switched voice messages participants generated instead are much more challenging to transcribe on a spreadsheet and using a phone-based media player designed for music playback. We thought that the benefit of familiarity of using existing tools would outweigh the challenges, but in retrospect, we wish we could have supported the person transcribing the data more by scaffolding the task and combining better tools. We have no doubt that the transcriptions and translations we received accurately reflected the meaning of what people said. However, in contexts where code-switching occurred, or when people repeated words or restarted a sentence, transcriptions were not verbatim enough.

Unfortunately, we only discovered these issues with participant-generated testing data quite late on in the development process: after having already begun model tuning based on the initial provided transcripts, a phenomena that Sambasivan et al. [65] refer to as a data cascade. Once we realised that these transcripts did not necessarily reflect ground-truth, this cascaded to word-error-rate (WER) calculations and the decisions that are made in response to these WERs. This made it harder to assess whether our sub-word modelling techniques were better than our word modelling techniques; whether the 'noise' from our Commoncrawl data-set was impeding overall speech recognition accuracy. We returned to the person transcribing the data to ask if they could review the output. We empathised that the task was more challenging than we envisioned and how it might be demotivating to revisit work. We also offered more tips and techniques and carefully explained the need for precisely verbatim transcriptions in more detail using examples from the dataset. Finally, we suggested that transcription accuracy is more important than quantity. So, we ended up with fewer transcripts to tune the isiXhosa model than we had initially expected. The deployed isiXhosa ASR system (tested using the limited re-transcribed data-set) had a WER of 87.52 and character-error-rate (CER) of 40.7.

[5]https://commoncrawl.org/

Table 4: Summary of testing dataset collected from the Langa workshop. (Note: due to the low accuracy of initial transcripts we do not have all details for row 1.)

	Minutes	Utterances per recording	Total tokens	Mixed/code-switched tokens
Langa dataset	25.02	118	–	–
Re-transcribed subset	4.75	36	577	17

6.2 Langa Reflections

We debated within the team if it was worth deploying the isiXhosa system, given the shortcomings we already identified, and the poorer performance compared to the Marathi model. However, we quickly recognised that we are indebted to participants to showcase the modest progress we had made and to invite formative feedback and reflections. Consequently, we deployed the isiXhosa model to the Voice Note app and server and asked participants from the second workshop to experiment with it.

Unsurprisingly, they found that the isiXhosa system was prone to errors, particularly for the "informal isiXhosa" that participants use from time to time. Participants also noticed "a few typos with spelling", an issue which relates to isiXhosa's agglutinative morphology that necessitated sub-word modelling. Typos therefore can emerge when the isiXhosa model incorrectly transcribes a sub-word prefix, but correctly transcribes the main word. Given the sociolinguistic norm of correctly-spelled and unabrreviated isiXhosa [19], at least in its written form, we were glad to have positioned the Voice Note app as a helpful tool for an individual rather than developing a bot that, say, transcribes voice messages in a WhatsApp group automatically and publicly as suggested in the Langa workshops. This choice resonates with Shneiderman's remarks that position Human-Centred AI systems as powerful tools to give users a high degree of control rather than only seeking automation [68].

Participants also identified false positives with English code-switching: "it somehow transcribed some words in English even though no English was used". Here isiXhosa phrases were being transcribed using similarly sounding, but in the context nonsensical, English words. When asked if it would be better to remove the ability to transcribe English and focus purely on isiXhosa, the participant emphasised that "people tend to mix English and isiXhosa" and that it's worth supporting both languages. Another participant agreed, "we tend to use both languages". However more work needed to be done "distinguishing when the switch happens".

Participants reported enjoying experimenting with advanced AI capabilities in their mother-tongue, which one participant thought was "pretty cool" and prompted him to identify other opportunities for NLP tools that could be useful for him – for instance to translate an isiXhosa message into SeSotho to facilitate communication with a Sotho person.

Participants also identified audio segments where the system performed well and saw value and purpose beyond the shortcomings of the ASR prototype's current implementation. They also encouraged us to improve the system. Such comments, are generally treated suspiciously within the field of HCI4D research as a form of response bias [17]. However, Langa community members did not hesitate to critique a prototype implementation in an unrelated project that embody ideas that could only be implemented rudimentarily, similar to our ASR prototype. Due to our longstanding engagements with the community and the facilitator's skill, we are therefore inclined to take these comments at face value. In contrast to the critiqued prototype, the area where the ASR prototype succeeded was integrating content that is close to participants' everyday lived experience; that of communicating through voice messages on WhatsApp.

7 SUMMARY, RECOMMENDATIONS AND OUTLOOK

Across the two deployments, participants in Langa and Dharavi demonstrated how switching between voice and text modalities can augment and support pervasive voice messaging practices. In Dharavi, where commercial ASR systems already support some Indian languages, participants mentioned using the voice typing featuring of the keyboard built into their Android phone. To be sure, readers might have similarly used ASR technologies to send hands-free messages (for instance while driving or otherwise occupied); or, perhaps, have experienced the frustration (or humour) when the ASR system does not get the transcription quite right. Our research shows the benefits of combining (and switching between) voice and text modalities when messaging. Given that all ASR systems are blighted by recognition errors to some extent, sending the audio recording alongside the 'voice-typed' message could alleviate user frustrations more broadly – that is, beyond the specific communities in Langa and Dharavi we partnered with. Or, vice-versa, as Langa participants mentioned, being able to access transcriptions of voice messages allows one to discretely 'peek' at a voice message's content, for instance while using public transport, but still be able to later appreciate the tone and emotion of the message, which is better conveyed in speech. Novel commercial video/podcast editing software such as Descript[6] already demonstrates the creative and collaborative potential enabled by pivoting between textual and spoken representations of content. However, low-resource languages are currently unsupported, and marginalised users—who also typically lack access to PCs—would be unable afford such a subscription service. We recommend further research on, and inclusive interface innovations targeting, the integration of voice and text modalities. At a more basic level, we recommend allowing users the option to preview a voice message before sending it, not only to conserve precious bandwidth, but also to avoid potentially embarrassing social situations if a voice message was already listened to before it could be deleted or edited by the sender.

Although ASR literature groups together isiXhosa and Marathi languages under the low-resource moniker, much like HCI4D research groups together users from Dharavi and Langa using the

[6] https://www.descript.com/

term 'emergent user' [20], our research further shows the importance of attending to and accounting for the nuances of the language-in-use and media ecologies of the specific communities we engage with. Not only are more language resources available for Marathi ASR development than isiXhosa, but the history and geography of Dharavi means that rather than code-switching between Marathi and Hindi or English, participants in our studies tended to speak Marathi with their friends and family and only fell back onto Hindi if they sensed that their conversational partner didn't understand them. Consequently, Dharavi participants not only benefited from a more robust Marathi ASR system, but that system did not have to cope with code-switched conversation.

We are now investigating other methods to improved code-switched modelling; in particular, to better model likely 'switch points' from isiXhosa to English and back again: see, for instance, the data augmentation method technique discussed in [75]. We have recently identified an additional source of more time natural, code-switched isiXhosa-English training data, sourced from South African soap opera corpus [56]. This provides a further 2.68 hours of training data with a mixture of English utterances, isiXhosa utterances and intra-sentential code-switched (i.e., switching between languages within a sentence) isiXhosa-English utterances. Particularly apt at illustrating the difference between this and the NCHLT data-set we used for the work reported here is the speaking-rate (phones-per-second), which is 8.77 in the read NCHLT corpus [5] compared to 19.98 for the Soap Opera corpus [56].

In 'low-resource' ASR contexts, such linguistic challenges are exacerbated by the paucity of representative and accurately transcribed data. The 'unplatformed' [46] approach we took in our research effectively engaged participants and generated a dataset of authentic, representative conversational speech from participants. Within HCI it is common practice to co-create user interfaces, scenarios and use-cases to ensure that new technologies address the needs of people. Our research demonstrates that collaborations across HCI and NLP that emphasise community engagement can, feasibly, also generate invaluable datasets: we recommend that future research in this space co-create use-cases, interfaces *and* datasets.

With our emphasis on ethical and transparent data-collection, we did not enlist the help of professional transcription services, which came at the expense of transcription accuracy. Increasingly, calls to decolonise speech and language technology not only draw attention to how labour-intensive, messy, incomplete and theory-laden the transcription process is, but just as importantly, "take seriously the sovereignty of local people over their data" [8]. Where our work fell short is providing better support for the transcription task when generating new datasets. So we furthermore recommend to assess transcription quality early to avoid data cascades [65] and to develop inclusive tools to enable communities to generate their own high-quality transcripts, retaining data sovereignty. HCI research in this space has shown that mobile-friendly transcription tools, such as *Respeak* [77], *Recall* [76] and *BSpeak* [78], simplify the transcription task to the benefit and empowerment of marginalised and excluded communities. However, these currently require users to iteratively either read a sentence or listen to a spoken audio segment and then subsequently clearly 're-speak' it, which a well-trained ASR system subsequently transcribes. Such tools not only present ASR development with a chicken-and-egg problem, but also have not been designed or evaluated with code-switched data in mind. We recommend that future ASR/HCI research develop mobile-friendly, inclusively designed, and code-switching compatible tools to generate the high-quality transcripts that are imperative for building robust ASR system.

It may be a stretch to think that such data-collection and transcription approaches are scaleable enough to, on their own, generate the supervised training data required for hybrid ASR methodologies (and those that similarly tune models to take linguistic knowledge and contextual insight into account more generally). However, 'low-resource' languages are not 'zero-resource' languages, so it is often possible to leverage existing datasets. The advantage of a co-created, authentic, and representative testing dataset in this context is invaluable during ASR development and can guide the myriad of small but together consequential decisions and trade-offs that are made along the way [27]. Here our research demonstrates that even modest amounts of testing data are already useful, but we also stress the importance of accurate transcription to avoid data-cascades [65].

ASR approaches such as ours have recently fallen out of favour, compared to 'unsupervised' end-to-end approaches and larger multilingual speech models, which rely on very high volumes of data. Of course, high volumes of data require equally high (and expensive) computational resources, so more popular 'state-of-the-art' approaches are increasingly becoming the exclusive provenance of large organisations able to afford the cost of computation and data, or are otherwise able to collect or extract it [13]. We believe, and are working towards, critical alternatives to this approach that could enable smaller entities—such as universities, smaller organisations embedded into communities with specific ASR use-cases, or someday perhaps communities themselves—to develop their own ASR systems.

8 CONCLUSION

Through our technical and creative methods to engage communities and develop and deploy ASR models of low-resource languages we have demonstrated opportunities and challenges of ASR system development. Most notably these revolve around the pervasive WhatsApp voice messaging practices of two communities of marginalised users in South Africa and India. Here, ASR-enabled systems have the potential to reduce information overload [34], broaden digital participation [4], and afford people enhanced opportunities to (re)discover and retrieve older voice message content, a form of digital possession [57].

Our research further reports on a series of creative voice messaging practices, such as combining both text and voice messages, and we call for further studies of such practices across the Global South.

Our research also aligns with rare but critically important scholarship on the challenges of developing high-stakes AI systems in the Global South [65], and shows that key insights, such as how low-resource contexts have "a pronounced lack of readily available, high-quality datasets" and the challenges of taking on such

'data work' also apply to developing ASR-enabled systems for low-resource languages and in postcolonial contexts [8] characterised by linguistic inequalities [18].

Reflecting on our research makes us now think of data, in both spoken and transcribed forms, as a boundary object [70] rather than a hand-over point [73] that can bind together the concerns of different communities of research—in our case HCI and ASR—but critically also extends into and engages with actual communities. Such collaborations have a role to play in future, for instance to better support collecting high-quality testing data of language-in-use annotated by accurate human transcriptions that also ensures that human efforts are leveraged to their fullest potential.

We are also mindful of critical AI commentators, such as Kate Crawford, whose "Atlas of AI" reconfigures AI as an industry that extracts and abstracts data away from the material conditions and the relationship it has with people and place [13]. Our attempts to augment data through commonly-used web-scraping techniques is an example of such practices which, as we reported, is also less effective for isiXhosa. However, a recurring theme identified by both Langa and Dharavi users of our ASR probe is a desire to 'listen' privately and discreetly by reading voice message transcripts. This demonstrates an intimate and sensitive relationship that users have to their voice messages. Furthermore, participants felt comfortable sharing voice message samples with us as part of our longstanding engagements with their communities. Here we can draw inspiration from the anthropologist Tim Ingold, who reminds us that the original meaning of 'collecting data' is to receive something that is given or offered, and not extracting what was not [37]. Of course, such acts implicate both giver and receiver in the norms, expectations and obligations of social life writ large as Marcel Mauss argues so beautifully in his seminal essay "The Gift" [52]. We experienced such obligations first hand, as we grappled with our decision to deploy the imperfect isiXhosa model. The clear recommendation of our work, then, is to make possible a future of ASR-enabled impacts through multidisciplinary collaboration and community partnership that relies more on data excellence and data ethics than data mining or scraping.

ACKNOWLEDGMENTS

We would like to thank Minah and Manik as well as the workshop participants in Langa & Dharavi for their contribution to this work. This work was supported by Engineering and Physical Sciences Research Council grants EP/T024976/1 & EP/M022722/1.

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. https://aclanthology.org/2020.lrec-1.343

[2] Heike Adel, Ngoc Thang Vu, Katrin Kirchhoff, Dominic Telaar, and Tanja Schultz. 2015. Syntactic and semantic features for code-switching factored language models. *IEEE/ACM transactions on audio, speech, and language Processing* 23, 3 (2015), 431–440.

[3] Heike Adel, Ngoc Thang Vu, Franziska Kraus, Tim Schlippe, Haizhou Li, and Tanja Schultz. 2013. Recurrent neural network language modeling for code switching conversational speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8411–8415.

[4] Devanuj Balkrishan, Anirudha Joshi, Chandni Rajendran, Nazreen Nizam, Chinmay Parab, and Sujit Devkar. 2016. Making and Breaking the User-Usage Model: WhatsApp Adoption Amongst Emergent Users in India. In *Proceedings of the 8th Indian Conference on Human Computer Interaction (IHCI '16)*. Association for Computing Machinery, New York, NY, USA, 52–63. https://doi.org/10.1145/3014362.3014367

[5] Etienne Barnard, Marelie H. Davel, Charl Johannes van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The NCHLT speech corpus of the South African languages. In *4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014*. ISCA, 194–200. http://www.isca-speech.org/archive/sltu_2014/sl14_194.html

[6] Nicola J. Bidwell, Mounia Lalmas, Gary Marsden, Bongiwe Dlutu, Senzo Ntlangano, Azola Manjingolo, William D. Tucker, Matt Jones, Simon Robinson, Elina Vartiainen, and Iraklis Klampanos. 2011. Please Call ME.N.U.4EVER: Designing for'Callback' in Rural Africa. In *Proceedings of the Tenth International Workshop on Internationalisation of Products and Systems*. Kutching, Malaysia.

[7] Nicola J. Bidwell, Simon Robinson, Elina Vartiainen, Matt Jones, Masbulele Jay Siya, Thomas Reitmaier, Gary Marsden, and Mounia Lalmas. 2014. Designing Social Media for Community Information Sharing in Rural South Africa. In *Proceedings of the Southern African Institute for Computer Scientists and Information Technologists Annual Conference (SAICSIT '14)*. ACM, New York, NY, USA, 104–114. https://doi.org/10.1145/2664591.2664615

[8] 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

[9] 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.

[10] Jeb Brugman. 2013. The Making of Dharavi's 'Citysystem'. In *Dharavi: The City Within*, Joseph Campana (Ed.). Harper Collins India, New Delhi.

[11] Houwei Cao, PC Ching, Tan Lee, and Yu Ting Yeung. 2010. Semantics-based language modeling for Cantonese-English code-mixing speech recognition. In *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 246–250.

[12] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (June 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

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

[14] Andiswa Mesatywa Dantile. 2015. *Language in Public Spaces: Language Choice in Two IsiXhosa Speaking Communities (Langa and Khayelitsha)*. Master's thesis. University of Stellenbosch, Stellenbosch, South Africa.

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

[16] Nicola Dell and Neha Kumar. 2016. The Ins and Outs of HCI for Development. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2220–2232. https://doi.org/10.1145/2858036.2858081

[17] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours Is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

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

[19] Ana Deumert and Sibabalwe Oscar Masinyana. 2008. Mobile Language Choices — The Use of English and isiXhosa in Text Messages (SMS) Evidence from a Bilingual South African Sample. *English World-Wide* 29, 2 (April 2008), 117–147. https://doi.org/10.1075/eww.29.2.02deu

[20] Devanuj and Anirudha Joshi. 2013. Technology Adoption by 'Emergent' Users: The User-Usage Model. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction (APCHI '13)*. ACM, New York, NY, USA, 28–38. https://doi.org/10.1145/2525194.2525209

[21] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Proc. Interspeech 2021*. 2446–2450. https://doi.org/10.21437/Interspeech.2021-1339

[22] Jonathan Donner. 2007. The Rules of Beeping: Exchanging Messages Via Intentional "Missed Calls" on Mobile Phones. *Journal of Computer-Mediated Communication* 13, 1 (Oct. 2007), 1–22. https://doi.org/10.1111/j.1083-6101.2007.00383.x

[23] Jonathan Donner. 2015. *After Access: Inclusion, Development, and a More Mobile Internet*. The MIT Press, Cambridge, Massachusetts.

[24] Paul Dourish. 2017. *The Stuff of Bits: An Essay on the Materialities of Information*. The MIT Press, Cambridge, Massachusetts.

[25] Paul Dourish and Scott D. Mainwaring. 2012. Ubicomp's Colonial Impulse. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/2370216.2370238

[26] Susan M. Dray, David A. Siegel, and Paula Kotzé. 2003. Indra's Net: HCI in the Developing World. *Interactions* 10, 2 (March 2003), 28–37. https://doi.org/10.1145/637848.637860

[27] M. C. Elish and danah boyd. 2017. Situating Methods in the Magic of Big Data and AI. *Communication Monographs* 85, 1 (2017), 57–80. https://doi.org/10.1080/03637751.2017.1375130

[28] Pedro Ferreira. 2015. Why Play? Examining the Roles of Play in ICTD. *Aarhus Series on Human Centered Computing* 1, 1 (Oct. 2015), 12. https://doi.org/10.7146/aahcc.v1i1.21264

[29] Harold Garfinkel. 1984. *Studies in Ethnomethodology*. Polity Press, Cambridge, UK.

[30] Sanjay Ghosh and Anirudha Joshi. 2014. Text Entry in Indian Languages on Mobile: User Perspectives. In *Proceedings of the India HCI 2014 Conference on Human Computer Interaction* (New Delhi, India) *(IndiaHCI '14)*. Association for Computing Machinery, New York, NY, USA, 55–63. https://doi.org/10.1145/2676702.2676710

[31] Google Cloud. 2021. Speech-to-Text: Automatic Speech Recognition. https://cloud.google.com/speech-to-text.

[32] Derek Gowlett. 2003. Zone S. In *The Bantu Languages*, Derek Nurse and Gérard Philippson (Eds.). Routledge, 609–638.

[33] Gabriel Haas, Jan Gugenheimer, Jan Ole Rixen, Florian Schaub, and Enrico Rukzio. 2020. They Like to Hear My Voice: Exploring Usage Behavior in Speech-Based Mobile Instant Messaging. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3379503.3403561

[34] Richard Harper. 2010. *Texture: Human Expression in the Age of Communications Overload*. MIT Press, Cambirdge, MA.

[35] Richard Harper. 2019. The Role of HCI in the Age of AI. *International Journal of Human–Computer Interaction* 35, 15 (Sept. 2019), 1331–1344. https://doi.org/10.1080/10447318.2019.1631527

[36] Richard Heeks. 2009. *The ICT4D 2.0 Manifesto: Where Next for ICTs and International Development?* SSRN Scholarly Paper ID 3477369. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3477369

[37] Tim Ingold. 2016. From science to art and back again: the pendulum of an anthropologist. *Anuac* V. 5 (Aug. 2016), 5–23 Paginazione. https://doi.org/10.7340/ANUAC2239-625X-2237

[38] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. Postcolonial Computing: A Lens on Design and Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1311–1320. https://doi.org/10.1145/1753326.1753522

[39] 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).

[40] Mark Jacobson. 2013. Mumbai's Shadow City. In *Dharavi: The City Within*, Joseph Campana (Ed.). Harper Collins India, New Delhi.

[41] Anirudha Joshi, Ashish Ganu, Aditya Chand, Vikram Parmar, and Gaurav Mathur. 2004. Keylekh: a keyboard for text entry in indic scripts. In *CHI'04 extended abstracts on Human factors in computing systems*. 928–942.

[42] Paul Kariuki and Lizzy Oluwatoyin Ofusori. 2017. WhatsApp-Operated Stokvels Promoting Youth Entrepreneurship in Durban, South Africa: Experiences of Young Entrepreneurs. In *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance (ICEGOV '17)*. Association for Computing Machinery, New York, NY, USA, 253–259. https://doi.org/10.1145/3047273.3047397

[43] Jasmeet Kaur, Asra Sakeen Wani, and Pushpendra Singh. 2019. Engagement of Pregnant Women and Mothers over WhatsApp: Challenges and Opportunities Involved. 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, 236–240. https://doi.org/10.1145/3311957.3359481

[44] Ashok Ramchandra Kelkar. 1958. *The Phonology and Morphology of Marathi*. Ph. D. Dissertation. Cornell University, Ithaca.

[45] Ondřej Klejch, Electra Wallington, and Peter Bell. 2021. The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Interspeech 2021*. ISCA, 2881–2885. https://doi.org/10.21437/Interspeech.2021-1035

[46] 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.

[47] Henri Lefebvre. 2004. *Rhythmanalysis: Space, Time and Everyday Life*. Bloomsbury.

[48] Ying Li and Pascale Fung. 2013. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7368–7372.

[49] Silvia Lindtner, Shaowen Bardzell, and Jeffrey Bardzell. 2016. Reconstituting the Utopian Vision of Making: HCI After Technosolutionism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1390–1402. https://doi.org/10.1145/2858036.2858506

[50] Nirav Malsattar, Nagraj Emmadi, and Manjiri Joshi. 2014. Testing the Efficacy of an Indic Script Virtual Keyboard: Swarachakra. In *Proceedings of the India HCI 2014 Conference on Human Computer Interaction* (New Delhi, India) *(IndiaHCI '14)*. Association for Computing Machinery, New York, NY, USA, 160–165. https://doi.org/10.1145/2676702.2677203

[51] Alexandre Maros, Jussara Almeida, Fabrício Benevenuto, and Marisa Vasconcelos. 2020. Analyzing the Use of Audio Messages in WhatsApp Groups. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 3005–3011. https://doi.org/10.1145/3366423.3380070

[52] Marcel Mauss. 2000. *The Gift: The Form and Reason for Exchange in Archaic Societies*. W.W. Norton, New York, NY.

[53] Moira McGregor, Nicola J. Bidwell, Vidya Sarangapani, Jonathan Appavoo, and Jacki O'Neill. 2019. Talking about Chat at Work in the Global South: An Ethnographic Study of Chat Use in India and Kenya. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[54] Andries Nel. 2021. South Africa - Languages. *Encyclopedia Britannica* (2021).

[55] Ngũgĩ wa Thiong'o. 1987. *Decolonising the Mind: The Politics of Language in African Literature*. Zimbabwe Pub. House, Harare, Zimbabwe.

[56] Thomas Niesler et al. 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)*.

[57] William Odom, John Zimmerman, and Jodi Forlizzi. 2014. Placelessness, Spacelessness, and Formlessness: Experiential Qualities of Virtual Possessions. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, New York, NY, USA, 985–994. https://doi.org/10.1145/2598510.2598577

[58] Walter J. Ong. 2009. *Orality and Literacy: The Technologizing of the Word* (reprinted ed.). Routledge, London.

[59] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, and Anirudha Joshi. 2019. Diversifying Future-Making Through Itinerative Design. *ACM Transactions on Computer-Human Interaction* 26, 5 (July 2019), 33:1–33:21. https://doi.org/10.1145/3341727

[60] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proc. Interspeech 2018*. 3743–3747. https://doi.org/10.21437/Interspeech.2018-1417

[61] 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.

[62] Shan M Randhawa, Tallal Ahmad, Jay Chen, and Agha Ali Raza. 2021. Karamad: A Voice-based Crowdsourcing Platform for Underserved Populations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 569. Association for Computing Machinery, New York, NY, USA, 1–15.

[63] 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

[64] Marie-Caroline Saglio-Yatzimirsky. 2016. *Dharavi: From Mega-Slum to Urban Paradigm*.

[65] 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

[66] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised Pre-Training for Speech Recognition. *arXiv:1904.05862 [cs]* (Sept. 2019). arXiv:1904.05862

[67] 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]

[68] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (Sept. 2020), 109–124. https://doi.org/10.17705/1thci.00131

[69] Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2020. A Survey of Code-switched Speech and Language Processing. arXiv:1904.00784 [cs.CL]

[70] Susan Leigh Star. 2010. This Is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology & Human Values* 35, 5 (Sept. 2010), 601–617. https://doi.org/10.1177/0162243910377624

[71] Statista Research. 2020. *Most Popular Mobile Apps Used in South Africa.* Technical Report.
[72] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]* (June 2019). arXiv:1906.02243
[73] Lucy Suchman. 2002. Practice-Based Design of Information Systems: Notes from the Hyperdeveloped World. *The Information Society* 18, 2 (2002), 139–144. https://doi.org/10.1080/01972240290075066
[74] Anna Lowenhaupt Tsing. 2005. *Friction: An Ethnography of Global Connection.* Princeton University Press, Princeton, N.J.
[75] 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.
[76] 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.
[77] 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.
[78] 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.
[79] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for Mandarin-English code-switch conversational speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4889–4892.
[80] Sarah Vuningoma, Maria Rosa Lorini, and Wallace Chigona. 2021. How Refugees in South Africa Use Mobile Phones for Social Connectedness. In *C&T '21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech (C&T '21)*. Association for Computing Machinery, New York, NY, USA, 128–137. https://doi.org/10.1145/3461564.3461569
[81] 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.
[82] Marion Walton, Vera Vukovic', and Gary Marsden. 2002. 'Visual Literacy' as Challenge to the Internationalisation of Interfaces: A Study of South African Student Web Users. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems - CHI '02*. ACM Press, Minneapolis, Minnesota, USA, 530. https://doi.org/10.1145/506443.506465
[83] Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Towards end-to-end automatic code-switching speech recognition. *arXiv preprint arXiv:1810.12620* (2018).
[84] Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. *arXiv preprint arXiv:1909.08582* (2019).
[85] World Bank. 2018. *Overcoming Poverty and Inequality in South Africa.* Technical Report.
[86] Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, and Haizhou Li. 2018. On the end-to-end solution to mandarin-english code-switching speech recognition. *arXiv preprint arXiv:1811.00241* (2018).
[87] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (first edition ed.). PublicAffairs, New York.