UnMute Toolkit: Speech Interactions Designed With Minoritised Language Speakers

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

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

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

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

Nina Markl
University of Essex
Colchester, UK
nina.markl@essex.ac.uk

Emily Nielsen
Swansea University
Swansea, UK
e.e.nielsen@Swansea.ac.uk

Gavin Bailey
Swansea University
Swansea, UK
g.bailey@swansea.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

ABSTRACT

In this paper and interactive exhibit we demonstrate a portfolio of systems and approaches that progressively vary on the theme of co-creating and situating spoken-language technologies to suit the needs, functions, and ways of speaking of diverse, resource-constrained, and under-heard language communities in South Africa and India. These systems demonstrate the benefits of human-centred machine learning methodologies and showcase how language technologies and conversational systems can broaden digital participation of minoritised language communities.

CCS CONCEPTS

• **Computing methodologies → Speech recognition**; • **Human-centered computing → Participatory design**.

KEYWORDS

Speech/language, information retrieval, co-creation, toolkit

ACM Reference Format:
Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Dani Kalarikalayil Raju, Nina Markl, Emily Nielsen, Gavin Bailey, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2024. UnMute Toolkit: Speech Interactions Designed With Minoritised Language Speakers. In *ACM Conversational User Interfaces 2024 (CUI '24), July 8–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640794.3665564

CUI '24, July 8–10, 2024, Luxembourg, Luxembourg
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665564

1 INTRODUCTION

An unfortunate reality is that most of the 7000+ languages spoken today have few – or no – existing digital language resources, which are needed to spur the development of spoken language technologies and conversational user interfaces. Such local, vernacular languages are generally spoken rather than written, lack standardised orthographies, and may even be entirely un-written [3]. Over the past three years, we have partnered with two such minoritised language communities, in South Africa and India, and uncovered salient opportunities for spoken language technologies to suit local needs, functions, and ways of speaking. A recurring theme in this line of research is the lack of representative datasets and access to even rudimentary, interactive demonstrator system to showcase and engage communities on what spoken language are and how they might be(come) meaningful or useful in context.

Through our research, and in response to these challenges, we have developed an open-source toolkit and adaptable blueprint: for partnering with and engaging minoritised language communities to co-create representative datasets that foreground ethics and consent; pipelines to develop interactive spoken language technologies; and, interactive demonstrator systems to deepen engagement, gather feedback, and stimulate design discussions.

We acknowledge that our demonstrator systems lack the finesse of more mainstream conversational systems for widely spoken languages that have standarised orthographies and widely available digital language resources. However, through this interactive demonstration we showcase the opportunities, innovations, and novel contexts at the margins of CUI. Our systems and methods also embody principles and practices that fall under the banner of responsible AI, and thus offer a critical alternative to the data

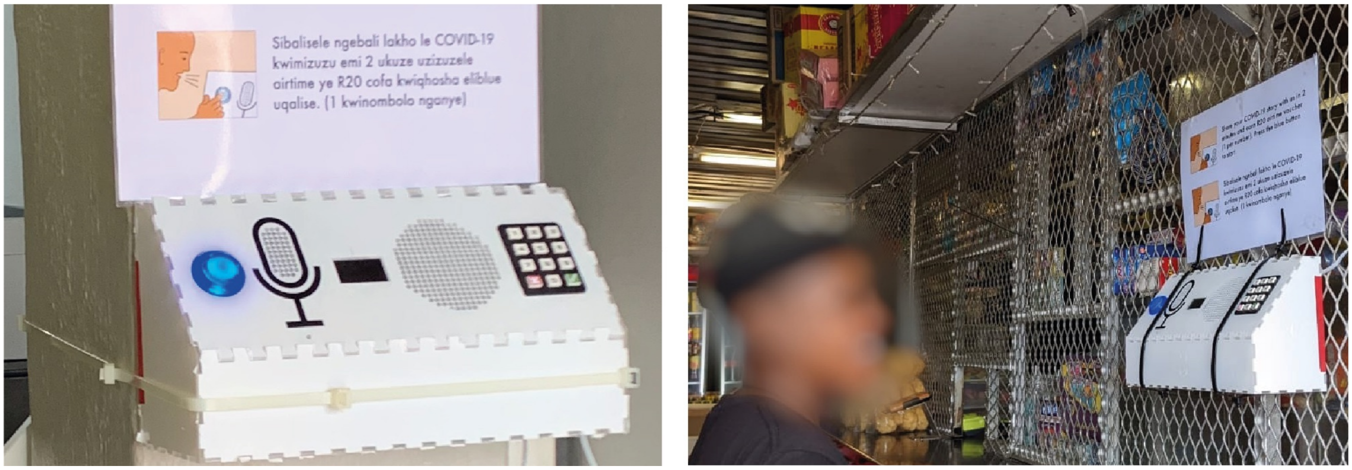

Sibalisele ngebali lakho le COVID-19 kwimizuzu emi 2 ukuze uzizuzele airtime ye R20 cofa kwiqhosha eliblue uqalise. (1 kwinombolo nganye)

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

mining practices, and questionable ethics, that undergird contemporary forms of spoken language systems [5]. Our demonstration is therefore also intended to provoke reflection and open up and advance such critical conversations.

2 CONTEXT

The demo consists of two journeys: one beginning in Langa, an isiXhosa-speaking community on the outskirts of Cape Town, South Africa; and the other in an agrarian community of Banjara tribespeople in Western India.

Langa is the oldest township in Cape Town, and was one of the original areas set aside for nonwhites during apartheid rule. The township is now classified as a previously disadvantaged area and has about 50,000 residents. While most residents speak isiXhosa at home, they draw on a linguistic repertoire of more than one language (e.g., English, Afrikaans, and/or isiZulu, amongst others) and often switch from one language to another in their daily interactions, a process that residents refer to as 'mixing' and linguists refer to as 'codeswitching' [9].

Access to and usage of digital technologies is strongly shaped by high data costs. Platforms like WhatsApp with subsidised data rates that compress digital media (e.g., pictures, audio and videos) before sending them, or peer-to-peer protocols with no associated data costs, are more popular than streaming services [6, 14]. Language speakers have limited practice with the written norms of the language, which have not kept abreast with the vibrancy and innovation of the language [7], particularly as it is spoken and mixed in public spaces in Langa. Consequently, residents of Langa—and the specific ways they speak and mix languages—are poorly represented in online spaces and existing speech datasets [2].

The Banjara tribe we partnered with in India were previously nomadic until forced to settle by British colonial rule. They speak a language called Gormati which does not have an indigenous script [4, p. 57], nor digital language resources. In this insular community, kith and kin speak Gormati to each other, however outside of their community use regional (Marathi) and national (Hindi) languages to converse [4, p. 53]. A major barrier to digital participation is text-input (see [8]), especially since older generations often cannot read or write, however data costs in India are some of the cheapest in the world and are locally affordable.

Our systems and approaches are tailored to the linguistic and digital ecology of each community and have been refined and stress-tested through deployments. The contrasting challenges of these communities tests the suitability of the toolkit to be applied to a number of contexts and languages.

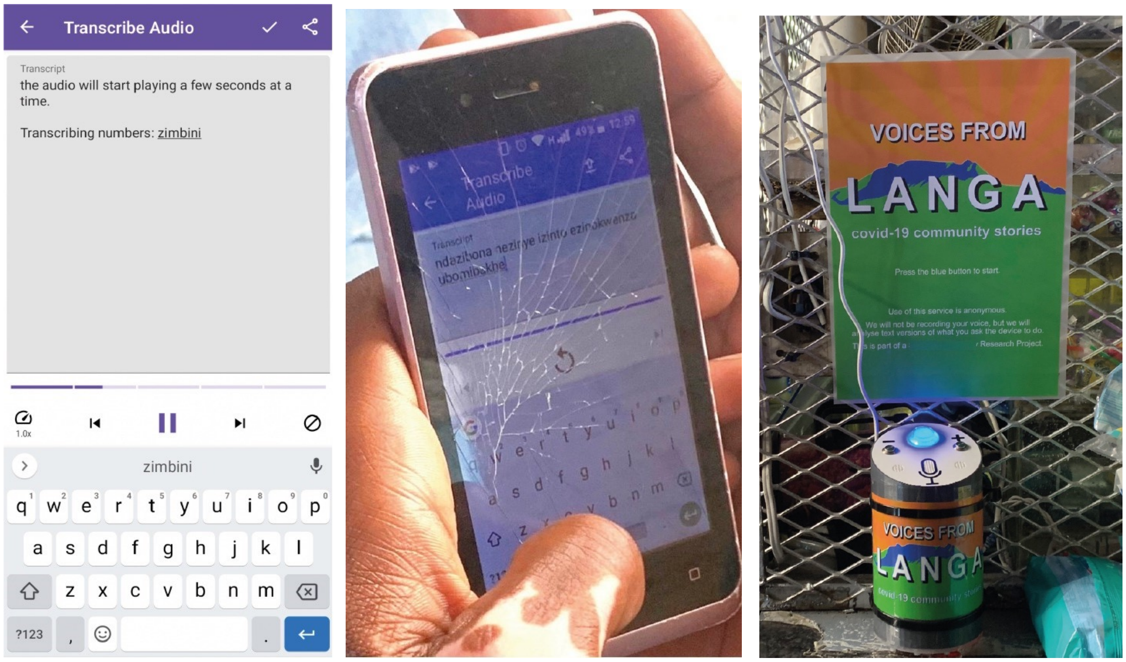
3 LANGA SYSTEMS

Early on in our research we uncovered widespread voice messaging practices of community members in Langa, who identified the use-cases of searching for old voice messages using automatically generated transcripts and reading transcripts of incoming voice messages in situations where it is not possible to listen privately (e.g., in a bus)[10]. However, such informal conversations, which featured pervasive code-switching between isiXhosa and English (see example in Table 1) was poorly supported by automatic speech recognition (ASR) systems trained on existing datasets [2] that lacked this vibrant and innovative language use.

3.1 Data collection using the SpeechBox

To gather more representative data that reflect the ways in which people in Langa speak in everyday life, we created, tailored, and deployed our SpeechBox system. It enables community members with little technological exposure, and budget for expensive data packages, to contribute their experiences about a topic of interest to a wider public audio collection. Deployed in public spaces, the device requires no prior training, and allows community members to share verbal narratives at the touch of a button.

The SpeechBox runs on bespoke client and sever software that in essence (1) prompts users to record their perspectives on a topic of community interest, (2) gives users an opportunity to re-listen (and if necessary re-record) the story, before (3) confirming that they are happy to add their narrative to the public collection, which



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

Transcribing numbers: zimbini

(a) (b) (c)

Figure 2: The main TranscriptTool interface to involve community members in the transcription of collected speech responses: (a) as a screenshot taken from the instruction video and (b) installed on a participant's phone with a cracked screen. Figure (c) shows the interactive demonstrator as it was deployed in one of five shops in Langa over a period of three weeks.

Table 1: Example community-sourced transcription and translation of a story recorded on the SpeechBox with code-switching highlighted in bold.

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

then uploads and stores their contribution on the server. Optionally, the user can input their mobile number to receive a small incentive payment[1]. We configured the SpeechBox to invite community members to record and reflect on their experiences of COVID-19 lockdowns, which at the time was a topic of great interest.

During a two-week deployment of two appliances in a local shop and internet cafe in Langa, we collected 318 audio recordings [11]. The contents of the recordings reflected the messiness and dynamic language of social life in Langa writ large that are missing from existing isiXhosa speech datasets (see example Table 1).

For our demo at CUI2024, we will invite attendees to record and reflect on their experiences at the conference. For online attendees we will place a webcam, microphone, and speaker in front of a second dedicated SpeechBox, and will press the physical buttons on their behalf.

4 COMMUNITY-SOURCED TRANSCRIPTION USING THE TRANSCRIPTTOOL

We developed a mobile-friendly transcription app to involve community members in the transcription of collected audio data (see Fig. 2a). The app supports and scaffolds the core audio transcription task, by breaking longer recordings into 5-second segments and overlapping consecutive segments slightly to ensure that if a word is cut-off at the end of one segment, it can be easily picked up in the start of the next segment. The tool also extends HCI/ASR scholarship on crowdsourced transcription because it does not depend on an existing ASR systems (e.g., [13]) or support only read speech (e.g., [1]).

Involving community members in the transcription of audio surfaced key insights on how isiXhosa is a weakly-standardised language and how written norms have not kept abreast with the vibrancy and innovation of the language [7], particularly as it is spoken and mixed in public spaces in Langa. Consequently, the

[1] Mobile numbers are deleted after payment is made and are not linked to recordings

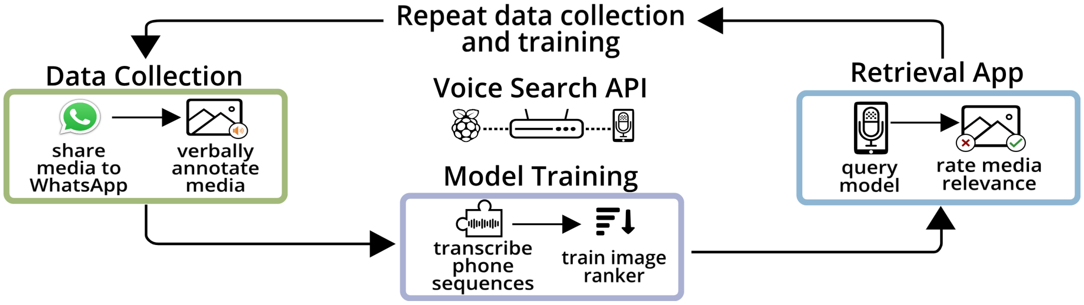


Figure 3: Diagram of our voice search toolkit to develop speech-based information retrieval systems for un-written languages.

community transcribed dataset challenges the orthodoxies and assumptions of ASR development pipelines.

For our demo at CUI2024, we will invite attendees to transcribe the recording they created previously using the TranscriptTool. If attendees did not consent to sharing their recording, they will be able to transcribe a different example recording instead. Online attendees will be able to use the TranscriptTool through screen-sharing.

4.1 Interactive Demonstrator

The final systems feeds back research results to community members and demonstrates a further use-cases for speech and language technologies in communities of minoritised language speakers. It gives community members the opportunity to query the corpus of collected stories on people's experiences of lockdown. It leverages an ASR system that was trained and tuned using a combination of existing isiXhosa language resources and with the help of data gathered in community collaboration.

Five instances of this system were deployed in Langa over a three week period in several shops (see Fig. 2c (right)). The system prompts users that it contains stories of lockdown, and that they can query that collection by stating a topic. The system transcribes the query and attempts to match it to stories in the corpus. Matched stories are then played back on the device. Around a third of queries generated a response on the device. Few participants chose to respond to the post-interaction rating question, but of those who did, 66 % rated the result as relevant [11].

For our demo at CUI2024, we will configure the interactive demonstrator to query and playback stories from the corpus that conference attendees recorded on the SpeechBox and then also explicitly consented to sharing. The system will be driven by an English ASR system. We will place a webcam, microphone, and speaker in front of a second device dedicated for online attendees and will press the physical buttons on their behalf.

5 BANJARA SYSTEMS

The systems we developed (Fig. 3) as part of our partnership with the Banjara community in India respond to the fact, that like the other ~3,000 endangered languages spoken across the world, their language does not have a written form, and have no digital language resources. Combining ethnographic and participatory methods reported elsewhere [11], we uncovered a first use-case for and demonstration of spoken language technologies in Gormati: an information retrieval system for multimedia content.

Without access to any existing Gormati language data, we split the core information retrieval system into two components: a multilingual phone recogniser and a ranker. We also asked community members to record multiple voice annotations describing each media item they wanted to store and retrieve. In total we collected 30 media items with ~4h of voice descriptions.

We used a *multi-lingual phone recogniser* trained on transcribed speech data from well-resourced languages to transcribe queries and annotations into phone sequences. We then trained the *ranker* to predict how these phone sequences correspond to each media item in the collection. Community evaluations showed how our system was able to return the relevant media item in response to a spoken query as a top-5 result 74% of the time.

Our demonstration consists of three main components: data collection; model training and a retrieval application. For the demonstration, this toolkit will be used to store and retrieve videos in response to spoken queries. The data collection, ranking, and voice retrieval systems all run entirely locally on a Raspberry Pi for communities to interact with, thus eliminating the need for data which can be costly and unreliable in digitally under-served populations.

5.1 Data Collection

Through our community-centred approach to data gathering, we ensure that language data is representative of everyday speech in the community and recruited community members to contribute and annotate media relevant to their context (i.e., farming images and video).

We experimented with digital storytelling software to collect both media items and voice annotations (see Fig. 4), and then pivoted to WhatsApp for adding new media and recording annotations in dedicated groups that we created for each media item. This pivot allowed us to leverage community member's familiarity with WhatsApp, but it became difficult for both researchers and participants to keep track of which media items are still in need of further annotation. We have since refined our approach to the web application, which can run on a Raspberry Pi micro-computer, seen in Figure 4.

For our demo at CUI2024, we will pretend that English is an unwritten language without any existing digital language resources. We will invite attendees to share photos and videos that are appropriate for, and of broad interest to, other attendees by posting these

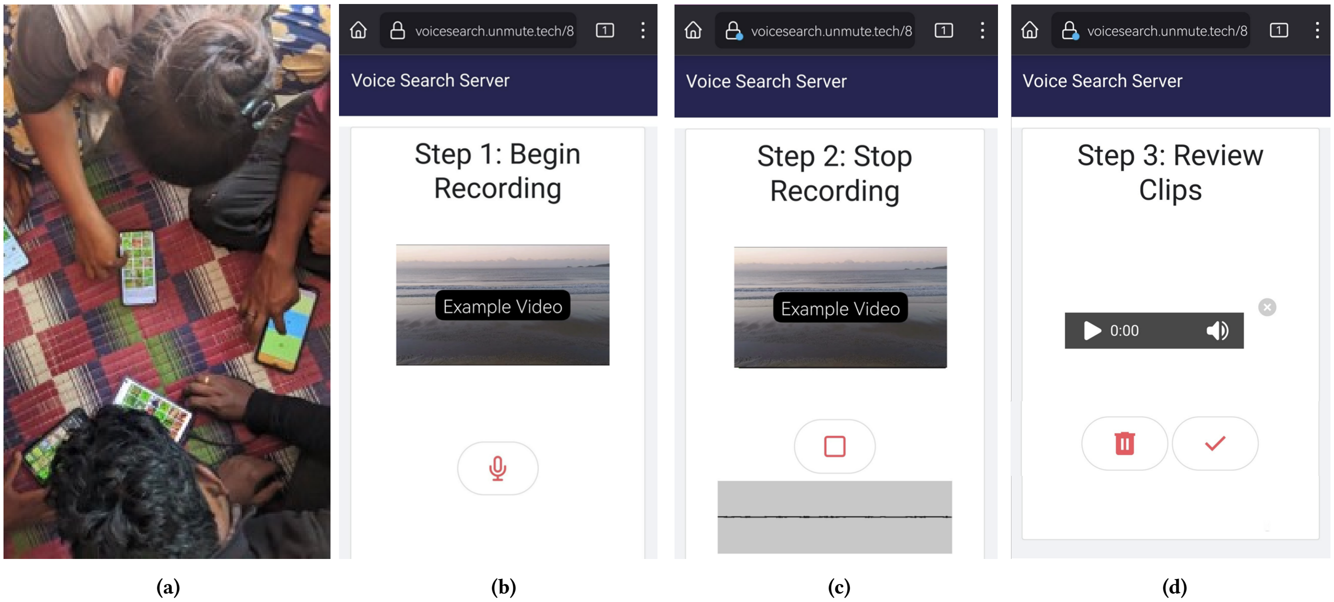


Figure 4: (a) Training community members to use the datacollection probe. (b–d) The data collection website pages for recording audio annotations for given media.

to a WhatsApp account. Furthermore, attendees will also have the opportunity to record voice annotations, describing videos that have previously been shared. Before any sharing takes place, attendees will need to consent to other attendees viewing or listening to recordings. Online attendees will also be able to contribute through screen and/or link sharing.

5.2 Model Training

The decoupled architecture of our information retrieval system, allows us to rapidly prototype and bootstrap the system with almost zero hours of spoken content in the target language. Furthermore, the system can be quickly retrained when new media items are added or further voice annotations are recorded.

During our demo at CUI2024, attendees will be able to explore the corpus of media items and voice annotations. They can also press the button to retrain the ranker and witness how quickly the ranker can be retrained (usually within a few seconds). Online attendees will also be able to participate through screen sharing.

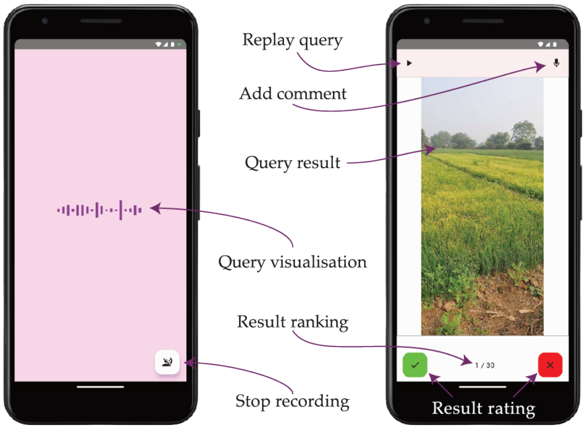


Figure 5: The media retrieval demonstrator app's screens for querying (left) and viewing/rating query results (right).

5.3 Retrieval Application

Following the training of the model, the voice search retrieval can be evaluated. We present a simple mobile application for this component: after speaking a query the user is shown the top-ranking media result as generated by our ranker. The user can then rate the result using either the green checkmark or red X-mark buttons, after which the next media in the ranked list is displayed.

During our demo at CUI2024, attendees will be able to query the collection of media items on a provided mobile phone. Through screen sharing, online attendees will also be able to interact with the mobile app, running on an emulator.

6 CONCLUSION

This demonstration presents a simple toolkit for developing a speech-based information retrieval system for low-resource languages. In this way the demonstration also respond to Seymour et al.'s provocative question of "*Who are CUIs Really For?*" and their finding that "participants recruited for user studies [published at CUI] were overwhelmingly from Europe and North America".[12]. Our research and toolkit thus also serve to stimulate discussions around approaches to include digitally and linguistically underserved populations within the CUI community and how this can be practically achieved through community partnerships.

ACKNOWLEDGMENTS

Across both research contexts, we thank community members, study participants, and data contributors for their contribution to this work. All source code for the systems reported in this paper is available at https://github.com/unmute-tech/. This work was supported by Engineering and Physical Sciences Research Council grant EP/T024976/1 and Responsible AI (EP/Y009800/1) grant IA012.

REFERENCES

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

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

[3] Steven Bird and Dean Yibarbuk. 2024. Centering the Speech Community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 826–839.

[4] J. J. Roy Burman. 2010. *Ethnography of a Denotified Tribe: The Laman Banjara*. Mittal Publications.

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

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

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

[8] 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

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

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

[11] Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Nina Markl, Lea-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2023. Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491102.3517639

[12] William Seymour, Xiao Zhan, Mark Cote, and Jose Such. 2023. Who Are CUIs Really For? Representation and Accessibility in the Conversational User Interface Literature. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3571884.3603760

[13] 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.

[14] 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.