Speech Technologies Datasets for African Under-Served Languages

Emmanuel Ngue Um
Institute of African DH
ngueum@gmail.com

Francis Tyers Morton
University of Indiana
ftyers@iu.edu

Eliette-Caroline E. Ngo Tjomb
University of Yaounde 1
eliette1404@gmail.com

Landry Dibengue
University of Yaounde 1
landrydibengue@gmail.com

Blaise-Mathieu Banoum Manguele
University of Yaounde 1
banoummanguelleblaisemathieu@gmail.com

Blaise Abbo Djoulde
Institute of African DH
tblaiso@gmail.com

Mathilde Nyambe A
Institute of African DDH
mathildenyambe@gmail.com

Brice M. Atangana Eloundou
Institute of African DH
martialatangana@gmail.com

Jeff S. Ngami Kamagoua
University of Yaounde 1
jeffsterlingngami@gmail.com

José Mpouda Avom
Institute of African DH
joseavom@gmail.com

Zacharie Nyobe
Institute of African DH
znyobe@gmail.com

Emmanuel Giovanni. Eloundou Eyenga
Institute of African DH
eloundoueyenga13@icloud.com

André Pascal Likwa
Institute of African DH
likwaiandre66@gmail.com

Abstract

The expansion of the speech technology sector has given rise to a novel economic model in language research, with the objective of developing speech datasets. This model is expanding to under-served African languages through collaborative efforts between industries, organisations, and the active participation of communities. This collaboration is yielding new datasets for machine learning, while also disclosing vulnerabilities and sociolinguistic discrepancies between industrialised and non-industrialised societies. A case study of a speech data collection camp that took place in September 2024 in Cameroon, involving representatives of 31 languages throughout the continent, illustrates both the prospects of the new economic model for research on under-served languages and the challenges of fair, effective, and responsible participation.

Introduction

There is a growing momentum in industry and academia to develop speech technologies on a massive scale. In the industrial domain, one of the most emblematic moves in this regard is the Massively Multilingual Speech (MMS) project initiated by Meta (Pratap et al., 2024), which aims to extend the coverage of speech technology across the global linguistic landscape. There are currently 336 African languages for which the MMS project has developed automatic speech recognition (ASR) and text-to-speech (TTS) models. MMS uses multilingual datasets to pre-train wav2vec 2.0 models, and the labelled dataset used for this pre-training consists of aligned New Testament recordings. This has enabled coverage of many of Africa's under-served languages, for which the Bible is often the only substantial textual resource. At an institutional level, academics and organisations are working together to build language datasets for machine learning in African languages. This is evidenced by initiatives such as The Lacuna fund[1], which has enabled the creation of a diverse range of language datasets, including speech datasets in more than 20 African languages over the past three to four years (Babirye et al., 2022).

Despite this progress, significant limitations remain, particularly in the dominant crowdsourced data collection model employed by platforms such as Mozilla Common Voice (MCV)[2] (Ardila et al., 2020). While MCV is widely recognised for enabling community participation in the creation of speech datasets, several critical flaws undermine its effectiveness for under-served languages. A significant challenge pertains to the dearth of publicly accessible text sources that can be collated for utilisation as reading prompts, compelling the reliance on religious texts such as the Bible, which are frequently the sole non-licensed text data sources.

[1] https://lacunafund.org/datasets/language/
[2] https://commonvoice.mozilla.org/en/about

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 82–90
March 4-5, 2025 ©2025 Association for Computational Linguistics

While the Bible may not be the predominant text source in most of the MCV's collecting interfaces for African languages, the absence of text diversity in under-resourced languages leads to a limited representation of language use, significantly differing from the fluid and varied nature of daily language usage. Additionally, the platform's framework tends to impose a single orthography model for each language, disregarding the linguistic diversity and orthography multiplicity found within many African communities. This rigid approach has the potential to marginalise certain dialects or writing traditions. Another challenge stems from the dependency on literacy participation, which excludes individuals who are fluent speakers but not proficient readers. Finally, the incentivisation of participation, while effective in the short term, raises questions about the sustainability of community engagement and the quality of collected data over time. The speech data collection camp organised by the Institute of African Digital Humanities (INHUNUM-A)[3]– in partnership with MCV, which constitutes a use case in this discussion – highlights these challenges. This experience has underscored the necessity for a more inclusive and adaptable approach to the development of speech technologies for African languages.

The initiative had two main goals. First, it sought to expand the reach of the MCV ecosystem in Africa by engaging community representatives to lead responsible, long-term crowdsourced speech data collection efforts. These efforts would be critical to the future development of speech technologies. Secondly, the initiative aimed to collect a 310 hour benchmark labelled speech dataset for 31 under-served African languages[4]. This paper reports on the key areas of the project and the challenges encountered during its implementation. These are grouped under (1) methodological, (2) technological, (3) sociolinguistic, (4) quality control, (5) incentivisation, (6) ethical aspects, and (7) discussion, and (8) recommendations.

1 Methodological aspects

In this section we discuss the approach to 1) the selection of languages and team members and 2) the collection and pre-processing of sentences.

1.1 Selection of languages and teams

The Institute of African Digital Humanities is a newly established organisation that aims to provide capacity building and networking in the use of digital methods and tools in the humanities and social sciences on the continent. Its outreach includes affiliated members, but more broadly any African-based institutional or individual stakeholder with an interest in digital humanities. In order to promote greater inclusivity across the regions and linguistic communities of the continent, an open call was launched to select teams, ideally consisting of two representatives of different genders and dialects within the same linguistic community. Candidates were also required to be fluent and literate in the language they were representing. In a sense, the selection was aimed at grassroots language enthusiasts who were not necessarily trained in linguistic research. In the same vein, the selection mechanism was designed to ensure, as far as possible, an equitable representation of linguistic diversity, to the extent that a given language was endowed with at least a standard orthography and a basic body of literature. Less emphasis was placed on criteria used in similar initiatives, such as regional representation, number of speakers or degree of standardisation (Butryna et al., 2020; Agirre et al., 2021). Languages with existing ASR or TTS models, including those developed in the MMS project, were excluded from the selection, even if they were more under-served. While this selection process was consistent with the principles of equity and representativeness that underpin the philosophy of our initiative, it did introduce some biases and inequalities. In terms of bias, the current ASR and TTS models developed within MMS, which are largely trained on biblical recordings, have not been sufficiently evaluated for performance, inclusivity and representativeness, raising concerns about the reliability of these technologies for the wider language community. In terms of inequality, the selection excluded de facto languages for which there was no existing orthography and/or a minimal body of literature.

Overall, The number of languages launched on MCV increased from 137 to 166, with the addition of 29 new languages[5], after the language data col-

[3]https://inhunumaf.hypotheses.org/

[4]https://github.com/Ngue-Um/INHUNUMA2024/blob/main/Inhunuma2024.md

[5]Setswana, one of the 31 languages involved, was already launched prior to the data collection event. Representatives of the Setswana languages attended the event with the objective of expanding the existing collection of sentence prompts to include the Kgatla dialect. At the time of this writing, Tunen,

Figure 1: MCV ecosystem in Africa before the data collection camp

Figure 2: MCV ecosystem in Africa after the data collection camp

lection camp held on September 9-14, 2024. This represents a growth of approximately 21.17%. The camp's contribution to expanding speech data collection for under-served African languages resulted in a significant increase in the platform's language offering, as represented on figures 1[6] and 2[7].

1.2 Sentence collection and preprocessing

There are two approaches to designing speech datasets using MCV. The first approach is Spontaneous Speech, whereby speakers are provided with prompts in their language, e.g. "What is the history of the origins of your community?", and are asked to respond in a few sentences, resulting in voice clip recordings. Subsequently, the recordings are listened to and transcribed, resulting in the alignment of voice and script labels. The second approach is called Read Speech, and consists of speakers reading sentence prompts. The resulting voice clips are then listened to by two different speakers who validate or invalidate the voice clip, assigning labels to the voice clip in the validation process. The second approach was used in our data collection camp. A prerequisite for the Read Speech approach is the provision of sentence prompts, which in the case of this project had to be provided by language teams. Each language teach was required to provide a minimum of 1000 sentences, the sources of which had to be licensed under Creative Commons (CCO). The majority of these sentences were either elicited by the team representatives or derived from their personal manuscripts, with some requiring digitisation and preliminary processing. Digitisation entailed the deployment of OCR (Optical Character Recognition) or manual typesetting by team members or project staff. In numerous instances, both processes resulted in inadequate rendering of characters, necessitating re-encoding or character conversion, and posing technological challenges. To address these challenges, language teams received support from language technologists and data scientists who are part of the MCV staff.

2 Technological aspects

In this section we discuss 1) the technological challenges of navigating competing writing norms and 2) the localisation of MCV interfaces.

2.1 The "ortho-graphy" challenge

The term 'orthography' has its roots in the Greek word *orthos*, meaning 'straight', 'correct' or 'right'. The emphasis on correctness in writing is based on the idea that languages are realities that can be reduced to coherent parts that reflect the range of possible uses within a linguistic community. The very notion of 'linguistic community' (Gumperz, 1968) is based on the assumption of the unity of the members of a given language group. While 'correctness' in orthography and 'unity' within the linguistic community are relatively easy to achieve in societies with a long history of political organisation and centralisation, with the exception of

a second language of the 31, is awaiting its launch.

[6] https://tinyurl.com/mcv-languages-before

[7] https://tinyurl.com/mcv-languages-after

societies such as Luxembourgish (Bellamy, 2021), many African societies in the post-colonial era have yet to achieve such ideals, if they have to at all. In the context of this study, there were regular instances where the materials submitted by the language teams revealed issues of competing orthographic norms. This was particularly pronounced in languages with a history of early missionary literacy before independence. Literature produced in the pre-independence missionary alphabet tended to contrast with post-independence orthographic standards. The latter were promoted by the second generation of missionaries, led by the Summer Institute of Linguistic (SIL) and Evangelical Missions, and operationalised by the first generations of linguists of African descent.

The coexistence of different, sometimes divergent, orthographic norms was difficult to resolve in the context of this initiative. In any case, the project leadership did not have the legitimacy and responsibility to make decisions regarding the choice of a particular orthographic norm. At the same time, the technological interface of linguistic infrastructures such as MCV is designed in accordance with the dominant, monolithic view that there should be one and only one orthographic norm for a given language. Final decisions about the choice of orthography were left to the team members. In such circumstances, an agreement was reached with the project leadership to give priority to the orthography standard that is widely used in the community.

2.2 Localisation of MCV Interfaces

Incidentally, decisions on the choice of spelling standard for the sentence collection did not always coincide with the choices made by the translators responsible for localising the interfaces in the various languages. For reasons related to the project schedule and the scarcity of competent human resources in the selected languages, the task of translating for localisation was sometimes entrusted to actors other than those involved in providing the sentence collections. The ideal situation would have been to reach a compromise between the translators and the sentence contributors. However, such arrangements were not always feasible, given the remote nature of the workflow between translators, sentence collectors, project management and MCV, and the critical impact of any delay on the project schedule. As a result, there are interfaces, such as that for Eton[8], where the localisation follows a different orthography standard from the sentence collection.

3 Sociolinguistic aspects

For want of a better option, the project managers had to force language representatives to pool their sentence samples. Initially, teams were asked to provide unified sentence collections for their languages. However, in cases such as Tupuri and Batanga, the two members of the team, each representing a particular dialect, provided a sample for their dialect. While in the case of Batanga the two samples used the same orthography, in the case of Tupuri the orthography used in the sentence sample from Tupuri Banwere, spoken on the border between Chad and Cameroon, differed slightly from the orthography used for Tupuri Bango, spoken in the area of Kaele in Cameroon. The two orthographies seemed to reflect the sociolinguistic configuration of the Tupuri linguistic community, and there did not seem to be any socio-political contestation of this reality. At the same time, MCV allows only one unique locale for each specific language, where the locale is represented by a two- or three-letter code, e.g. 'tui' (for Tupuri), 'bnm' (for Batanga), 'tn' (for Setswana). Technically, therefore, the MCV infrastructure does not appear to be configured to accommodate the sociolinguistic reality of Tupuri, which is manifested in the fluidity of usage in both spoken and written form. The example of Tupuri is not uncommon in accounts of applied language work in Africa. Roberts et al. (2021) refer to a similar situation among the Yambasa community in Cameroon, where groups of arguably distinct dialects have reclaimed orthographic autonomy and developed separate writing norms and practices.

4 Quality control

The quality control process was divided into seven stages and was subject to oversight from the MCV staff and a pool of local experts, as illustrated in Table 1.

5 Incentivisation

Incentivisation through cash and in-kind rewards is common practice in language work in general, for example in language documentation research involving community contributors (Ngue Um, 2019;

[8] https://commonvoice.mozilla.org/eto

Levels of control	Oversight
Localisation (sheets)	Local team
Sentences (Sheets)	Local team
Localised (Pontoon)	Local team
Approved (Pontoon)	MCV staff
Sentences (Checked)	MCV staff
Sentences (MCV)	MCV staff
Launched	MCV Staff

Table 1: Levels of quality control and oversight involved in the project

Akumbu, 2024). It has also been implemented in the creation of language datasets for machine learning as part of the Lacuna Fund initiative (Babirye et al., 2022). The benefits of paid labour can be measured in terms of the level of mobilisation of the actors involved and the extent to which they have contributed to the achievement of the project's objectives. In the specific case of the speech data collection camp organised by INHUNUM-A in September 2024, the impact of the incentives can be seen in the mobilisation of the participants before, during and after the data meeting, which enabled the recording and validation of more than 300 hours of voice data over a period of 30 days. In terms of diversity and linguistic representativeness, this represents a significant growth in the ecosystem of both MCV and speech datasets for machine learning.

However, there are a couple of side effects of incentivisation. One is the sustainability of community mobilisation beyond the scope of a particular project, such as the one undertaken. Withholding a portion of the monetary compensation for teams that did not meet the goal of 10 hours of voice recording and validation during the camp timeline, and paying it only after the goals were met, proved effective for continued mobilisation after the camp. However, for almost all the languages involved, once the incentives are fully paid, the tendency to contribute decreases significantly and sometimes stops altogether. This raises questions about the long-term sustainability of a crowd-sourced approach to speech data collection and, by extension, the voluntary, informed and qualitative participation of under-served communities in the development of speech technologies in their languages.

A notable dimension of this language data collection event is the under-representation of professional linguists, which contradicts the initial assumptions of the project leadership about a possible over-representation of linguists. In fact, of the 70 or so people who attended the meeting, only 3 professional linguists were listed. In comparison, there were three computer scientists. The majority of participants were grassroots language workers, either indigenous language teachers, translators, community literacy experts or language enthusiasts.

6 Ethical considerations and copyright

One of the major challenges in developing language datasets is the ethical considerations around data sources and community participation. For many under-served languages, existing text resources are sparse, and those that do exist are often limited to biblical texts. As a result, many existing ASR and TTS models in African under-served languages have been developed using these sources. This is the case with the MMS project, but also with the Building African Voices (Perez Ogayo, 2022) and Google Crowdsourced Speech Corpora for Low-Resource Languages and Dialects (Butryna et al., 2020) projects. This reliance on a religious text raises questions about the representativeness of the data, as it may not reflect everyday language use or cultural diversity within the community. In order to avoid expanding the inclusion of biblical texts in the language technologies of Africa's under-served languages, our project management reached an agreement with MCV to exclude such texts from the sentence collections. Although this provision was made explicit in the Call for Participation, a number of teams submitted sentence collections that were either entirely biblical or contained large swathes of religious texts taken from the Bible. In such cases, team representatives were asked to submit new collections. This has resulted in some of the initially selected teams dropping out of the project, or in long delays in the provision of the MCV interfaces for these languages.

In addition, the project had to deal with copyright issues, especially for languages such as Tunen, where the sentence sources were licensed under Creative Commons Attribution-ShareAlike (CC BY-SA), but needed to be licensed under Creative Commons (CCO) according to MCV standards. Community representatives were generally not well informed about copyright, and although the Call for Participation was explicit about these issues, the project leadership had not provided adequate guidance and resources to help community representatives navigate and resolve these issues as they arose.

7 Discussion

Crowdsourcing is a mode of participation that is becoming increasingly prevalent in social, behavioral, and educational research (Bagherzadeh et al., 2023; Kwek, 2020). Bagherzadeh et al. (2023) have identified two distinct approaches to the recruitment of participants in crowdsourced routines, which they have metaphorically designated as "fishing" and "hunting." The "fishing" routine targets a wide range of external knowledge on a specific domain, with the assumption that the diversity of the participants' input will enhance the robustness of the solution that is being engineered. In contrast, the "hunting" approach targets specific individuals with expert knowledge in the domain under investigation, seeking to elicit solutions from those with the greatest expertise.

In the domain of linguistic research, an analogy can be drawn with language documentation, a form of crowdsourced perspective of linguistic research in which data collection leverages the involvement of diverse contributions, profiles, and situations (Ajo et al., 2010; Grenoble, 2010; Maxwell, 2010; Himmelmann, 2006). While MCV's crowdsourcing perspective is generally of the "fishing" type, language documentation predominantly employs the "hunting" technique, with various accounts of success stories (Dwyer, 2010), as well as shortcomings (Akumbu, 2024; Ngue Um, 2019).

One aspect of crowdsourcing for speech data that appears to be overlooked in the "fishing" approach employed by MCV is the distinction between the literacy rate in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations and that in non-WEIRD ones (Brice et al., 2024). The implication of the literacy rate is that it indicates the degree of exposure of the average population to written text in the language for which speech datasets are collected. It is commonly assumed that a vast array of literacy expertise is readily available for crowdsourcing speech by reading sentence prompts, as well as for evaluating pre-recorded sentences. This is undoubtedly the case in literate societies and in WEIRD settings, but it is not the case in non-WEIRD, African under-served linguistic communities. Despite the fact that these communities have developed a considerable literacy rate through education, the reading and writing skills of individuals are still largely confined to the former colonial languages that serve as the medium of instruction in the majority of educational institutions across Africa. The implementation of the "fishing" approach in such circumstances thus renders crowdsourcing vulnerable.

Languages	Hours	Speakers	Validation
Duala	11	13	91%
Borgu Fulfulce	10	9	100%
Mbo	11	12	91%
Mokpwe	8	9	75%
Yoruba	7	123	72%
Hausa	13	50	39%
Ahmaric	3	34	67%

Table 2: Status of voice data contribution on MCV for 6 African languages (Language = “language name”; Hours = “total hours of speech recording, updated: 13th Oct. 2024 10:42am”); Speakers = “total number of contributors of recordings and validation”; Validation = “total number of labelled hours of speech data recording”.)

As previously noted in Section 5, in the context of the project described in this paper, 100% of the contributions for the 30 languages included in the collection have either ceased or decreased significantly after the final payment of incentives. This may be in alignment with the analysis presented by Bagherzadeh et al. (2023), which suggests that the "fishing" approach attracts a significant number of non-domain experts, primarily driven by financial incentives. This hypothesis can be further substantiated by examining the trends in speech data contributions for African languages that were launched on MCV but not included in our data camp, as illustrated in Table 2.

This analysis does not imply that participants who are primarily attracted by financial incentives lack domain expertise. In the context of this study, domain expertise is defined as literacy skills in the language in which speech data is crowdsourced. The argument, therefore, is that the motivation of those who are attracted primarily by financial motives is more likely to decrease drastically in the absence of incentivisation. Conversely, Bagherzadeh et al. (2023) suggest that elite experts, that is to say, the category of participants in crowdsourcing who are recruited using the "hunting" approach, do not engage out of the prospect of financial gain in the first place.

With respect to the number of contributing speakers and the total population of the linguistic community, the three languages indicated in the shaded section of Table 2 exhibit a comparatively larger population. This may justify why their contributing population is more significant than the number of the contributing population of the languages in the unshaded area. Thus, the "fishing" approach to crowdsourcing that represents MCV's standard contribution "doctrine" would result in a higher level of contribution from the languages in the shaded area compared to those in the unshaded area. As the data in Table 2 show, this is not the case. In particular, a greater number of contributors does not necessarily result in a proportional increase in hours of recorded speech and validation. The discrepancy in the contribution rate observed in this case can be attributed to at least two factors. First, the influence of incentives, which is reflected in the higher contribution rate of the languages in the upper part of Table 2. Second, in the context of under-served linguistic communities, the standard "fishing" approach of MCV does not attract elite experts, who are likely to spend more time recording and validating voices, even in the absence of financial reward. It is also noteworthy that the timing of the contribution rate in the languages at the top of Table 2 indicates that participation in the "fishing" approach is primarily driven by financial incentives.

8 Recommendations

The participation of individuals in crowdsourced linguistic datasets in exchange for financial compensation highlights the economic vulnerability of those engaged in such activities. In the specific context of African under-served linguistic communities, where literacy in indigenous languages is often low, this raises further questions about the quality of participation. In light of the above, there is an urgent need to develop robust protocols for crowdsourcing data for speech technologies such as ASR and TTS that aim for inclusivity and efficiency. This is especially true for crowdsourced participation aimed at collecting and labelling speech data. Similarly, the evaluation of the performance of ASR and TTS models trained on crowdsourced speech data in under-served linguistic communities should include an assessment of the crowdsourcing methods used, as well as an investigation of the potential influence of the socio-economic vulnerability of the contributors on the quality of the technological solutions developed. The success of the experience of the Speech Data Camp reported in this study, which we describe in terms of the achievement of the objectives initially stated, owes much to 3 main factors. The first is the incitement through cash payment of the contributors, which has attracted a critical mass of candidates to the speech contribution, and has enabled the management side to define selection criteria that could guarantee a reasonable level of literacy expertise of the selected participants, as well as the diversity of voices, in terms of representativeness of coexisting dialects and gender. Here it is important to emphasize that the design of the data camp model is an important step for the success of such an initiative. The second factor is the timing of data collection. In our model, most language teams achieved the best contribution scores in terms of number of hours and rate of progress during the camp. In other words, on-site mobilisation and emulation among peer groups is critical for the onboarding and self-motivation of contributors, even with the promise of financial reward. In comparison, the rate of contribution within one month after the data camp was significantly lower compared to the 6 days of contribution during the camp, despite the incentives. Reasons for this are related to the lack of focus when participants are in their normal social environment, as well as access to internet and electricity. The third factor is the quality of supervision and monitoring of the contributions. Once again, the examples of Yoruba, Hausa and Amharic in Table 2 show that in the absence of leadership to create a momentum of voice-data contributions, the growth of contributions may remain uncertain. The status of the Kinyarwanda[9] contribution illustrates this state of affairs. Namely, under the leadership of a speech data collection startup, Digital Umuganda[10], Kinyarwanda is currently the third most contributing language on MCV, just behind English and Catalan, and surpassing better endowed languages such as Spanish, French, and Chinese.

Conclusion

The initiative to enhance speech technologies for under-served African languages has highlighted both challenges and opportunities in language data collection. This paper details the methodological,

[9] https://commonvoice.mozilla.org/rw

[10] https://digitalumuganda.com/

technological, sociolinguistic, ethical, and incentive aspects of the project, while highlighting the significant progress made in collecting over 300 hours of speech data for 30 languages[11]. However, critical issues remain, such as uneven language representation, barriers to community engagement, and the biases introduced by reliance on pre-existing automatic speech recognition (ASR) and text-to-speech (TTS) models, many of which are rooted in religious texts.

The project also grappled with competing orthographic norms, issues of copyrights applicable to the sources of the sentence prompts, and the long-term sustainability of crowdsourced data collection efforts. Despite the tangible results achieved, ensuring continued community participation beyond financial incentives remains a challenge. Going forward, a deeper commitment to fostering authentic collaboration between language communities, linguists and industry is essential to ensuring the equity and efficiency of the new economy model brought by voice technologies.

In addition, expert linguists specialising in underserved African languages need to develop a critical awareness of the solution-oriented approaches driven by industry that are increasingly influencing applied linguistic work. Without a deep understanding of industrial and commercial practices in product and service design, linguists cannot critically and productively engage with industrial actors who own many of the technological solutions and financial resources. These industrial actors often lack key insights into which approaches are most appropriate for specific languages and contexts. Productive collaboration between linguists, communities and industry is essential to ensure that the technologies developed are not only linguistically sound, but also socially and culturally relevant to the communities they are intended to serve.

Acknowledgments

The study reported in this paper is the result of a data collection workshop (which we refer to here as the " Speech Data Camp") funded by Mozilla through its Common Voice initiative. The authors of this paper are grateful for the generous support that made it possible to achieve the goals of the project. More importantly, the organisation of this workshop has led to what one Mozilla staff member has called a "sea change" in the Mozilla Common Voice ecosystem. The authors of this paper are grateful to Mozilla for their trust and excellent oversight at all stages of the event.

[11]As of the date of submission of this paper, one language, Tunen, is awaiting clearance for copyright issues regarding the collection of sentence prompts submitted by representatives before its launch.

References

Eneko Agirre, Izaskun Aldezabal, Iñaki Alegria, Xabier Arregi, Jose Mari Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Kepa Sarasola, and Aitor Soroa. 2021. Developing language technology for a minority language: Progress and strategy. *ELSNews. 10.1*.

Frances Ajo, Valérie Guérin, Ryoko Hattori, and Laura C. Robinson. 2010. *Native speakers as documenters A student initiative at the University of Hawai'i at Manoa*, chapter 19. John Benjamins, Berlin.

Pius W. Akumbu. 2024. A community approach to language documentation in africa. In *ACAL in SoCAL: Selected papers from the 53rd Annual Conference on African Linguistics*, page 1–25.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 4211—4215.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira F., Jonathan Mukiibi, Medadi Ssentanda, Lilian D. Wanzare, and Davis David. 2022. Building text and speech datasets for low resourced languages: A case of languages in east africa. *AfricaNLP 2022*.

Mehdi Bagherzadeh, Andrei Gurca, and Rezvan Velayati. 2023. Crowdsourcing routines: the behavioral and motivational underpinnings of expert participation. *Industrial and Corporate Change*, 32(6):1393–1409.

John Bellamy. 2021. *Contemporary Perspectives on Language Standardization*, chapter 26. Cambridge University Press, Cambridge, UK.

Henry Brice, Benjamin Zinszer, Danielle Kablan, abrice Tanoh, Konan N. N Nana, and Kaja K Jasińska. 2024. Individual differences in leveraging regularity in emergent l2 readers in rural côte d'ivoire. *Scientific Studies of Reading*, 28(4):391–410.

Alena Butryna, Shan-Hui C. Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, Chenfang Li, Tatiana Merkulova, Yin M. Oo, Knot Pipatsrisawat, Clara Rivera, Supheakmungkol

Sarin, Pasindu de Silva, Keshan Sodimana, Richard Sproat, Theeraphol Wattanavekin, and Jaka A. Eko Wibawa. 2020. Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: An overview. *arXiv preprint*.

Arienne Dwyer. 2010. *Models of successful collaboration*, chapter 13. John Benjamins, Berlin.

Lenore A. Grenoble. 2010. *Language documentation and field linguistics: The state of the field*, chapter Conclusion. John Benjamins, Berlin.

John Gumperz. 1968. The speech community. *International Encyclopedia of the Social Sciences*, pages 381–386.

Nikolaus P Himmelmann. 2006. *Language documentation: What is it and what is it good for?*, page 1–30. Mouton de Gruyter, Berlin.

Adrian Kwek. 2020. Crowdsourced research: Vulnerability, autonomy, and exploitation. *Ethics & Human Research*, 42(1).

Judith M. Maxwell. 2010. *Training graduate students and community members for native language documentation*, chapter 18. John Benjamins, Berlin.

Emmanuel Ngue Um. 2019. *Achieving sustainable language preservation through economic empowerment in endangered language settings in West Africa*, chapter 20. Rüdiger Köpper Verlag, Cologne.

Alan W Black Perez Ogayo, Graham Neubig. 2022. Building african voices. *arXiv preprint*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ language. *Journal of Machine Learning Research*, pages 1–52.

David Roberts, Ginger Boyd, and JeDene Reeder. 2021. *Elip, Mmala and Yangben*, chapter 9. John Benjamins Publishing Company, Amsterdam.