Building Corpora for Low-Resource Kenyan Languages

Audrey Mbogho
USIU-Africa
Nairobi, Kenya
ambogho@usiu.ac.ke

Quin Awuor
USIU-Africa
Nairobi, Kenya
qawuor@usiu.ac.ke

Andrew Kipkebut
Kabarak University
Nakuru, Kenya
akipkebut@kabarak.ac.ke

Lilian Wanzare
Maseno University
Maseno, Kenya
ldwanzare@maseno.ac.ke

Vivian Oloo
Maseno University
Maseno, Kenya
voloo@maseno.ac.ke

Abstract

Natural Language Processing is a crucial frontier in artificial intelligence, with broad application across public health, agriculture, education, and commerce. However, due to the lack of substantial linguistic resources, many African languages remain underrepresented in this digital transformation. This article presents a case study on the development of linguistic corpora for three under-resourced Kenyan languages, Kidaw'ida, Kalenjin, and Dholuo, with the aim of advancing natural language processing and linguistic research in African communities. Our project, which lasted one year, employed a selective crowd-sourcing methodology to collect text and speech data from native speakers of these languages. Data collection involved (1) recording and transcribing conversations and translating the resulting text into Kiswahili, creating parallel corpora, and (2) reading and recording written texts to generate speech corpora. We made these resources freely accessible on open-research platforms, namely Zenodo for the parallel text corpora and Mozilla Common Voice for the speech datasets, thereby facilitating ongoing contributions and developer access to train models and develop Natural Language Processing applications.

1 Introduction

Natural Language Processing (NLP) has become a vital component of modern artificial intelligence (AI), shaping various sectors from healthcare to commerce. In recent years, advances in hardware, sophisticated algorithms, and the availability of large text data sets have enabled NLP systems to deliver unprecedented capabilities, transforming many aspects of human activity (Chollet, 2021). Generative AI tools, such as ChatGPT, exemplify the reach and impact of NLP innovations. However, these advances have exacerbated the global digital divide, particularly in linguistic terms. Many African languages, including those indigenous to Kenya, remain excluded from these developments, limiting access to critical information and technological benefits for their speakers (Nekoto et al., 2020).

The linguistic divide is rooted in historical inequalities that stem from colonialism, which imposed foreign languages such as English, French, and Portuguese as the official means of communication in Africa. These languages dominate education, governance, and technology, often at the expense of indigenous languages (Bamgbose, 2011). Language policies in education, while sometimes supportive of mother-tongue instruction on paper, often lack practical implementation mechanisms (Awuor, 2019). This legacy has led to the marginalisation of a large part of the African population, who do not speak these colonial languages fluently, restricting their access to essential information and the technological tools developed for those languages (Nduati, 2016). During the COVID-19 pandemic, for example, critical public health information in Kenya was predominantly disseminated in English and Kiswahili, excluding speakers of indigenous languages, especially in rural areas.

This linguistic exclusion underscores the urgent need to integrate African languages into AI and NLP technologies. Building high-quality linguistic corpora for underrepresented languages is crucial to enable language technologies such as machine translation, speech recognition, and speech synthesis for African languages. In this context, our project focused on developing language corpora for the three Kenyan languages Kidaw'ida, Kalenjin, and Dholuo, by collecting text and speech data and translating the text data into Kiswahili to generate parallel corpora. These resources are intended to facilitate the development of NLP applications that meet the needs of African communities, thereby promoting linguistic diversity in AI development.

In the last five years, there has been a surge in NLP activity for African languages. There is a

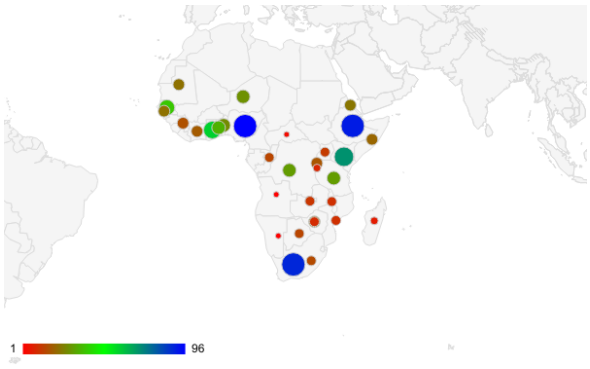

Figure 1: Distribution of NLP activity in Africa based on the number of NLP terms appearing in relation to each country in public sources

need to get an overall picture of what is being done and where. A search for "African language NLP" reveals a concentration of activity in only a handful of countries, among them South Africa, Nigeria, Ethiopia, Kenya, and Ghana. This state of affairs is depicted in Figure 1 based on the number of mentions of specific African languages in publicly available Internet sources. Although our project focused on only three Kenyan languages, one objective of this article is to inspire others to initiate similar projects across the African continent and to ensure that no language is left behind.

This research aimed to answer the following questions:

1. What methods are effective in acquiring data for building language corpora for low-resource languages?

2. What can motivate members of target language communities to contribute language data?

3. What methods can be applied to ensure the quality of the language data collected?

2 Background on Kidaw'ida, Kalenjin and Dholuo

In this section, we provide background information on each language and describe any language data available for it. We also describe the status of the language with respect to NLP resources.

2.1 Kidaw'ida (ISO 639-3 Code: dav)

Kidaw'ida is a Bantu language spoken Kenya by approximately half a million people in Taita Taveta County in Southeastern. Apart from the Bible, and the Anglican Hymn Book, few publications in Kidaw'ida can be found. Frank Mcharo and Peter Bostock published, in Kidaw'ida, a small 71-page book on the customs and traditions of the W'adaw'ida (Mcharo, 1995). This is a copyrighted resource and can only be used with permission. Furthermore, it is only available in print. Although online articles, social media posts and self-published story books written in Kidaw'ida are known to exist, they are hard to discover and consolidate. Indeed, social media platforms like Facebook, X and WhatsApp could be a rich source of Kidaw'ida text, but to extract it from the mixture of languages used on such platforms presents a significant challenge.

Kidaw'ida faces the threat of glottophagy from the more dominant Kiswahili due to their close geographical proximity and the intermingling of speakers of different Kenyan languages. Glottophagy is a term coined by the French linguist Louis-Jean Calvet (Calvet, 1974) to describe the "eating" of a language by a more dominant one. It perfectly captures what is arguably the greates current threat to Kidaw'ida. This threat is witnessed in real time as Kiswahili words are rapidly replacing Kidaw'ida words in day to day conversations. Under normal circumstances, borrowing of words is a boon to a language as it gives the language words that it lacks, thus strengthening the expressive power of the language. But the replacement of words when not necessary is very concerning as it could mean the erosion of the culture and the loss of crucial indigenous knowledge. This is not to deny the inevitable evolution of all languages, but rather to raise an alarm about the rapidity of the change affecting languages like Kidaw'ida that are spoken by a small population and that are threatened by a dominant language.

2.2 Kalenjin (ISO 639-3 Code: kln)

Kalenjin is a Nilotic language, belonging to the Eastern Nilotic branch of the Nilo-Saharan language family. It is primarily spoken by the Kalenjin people, who predominantly reside in the Rift Valley region of Kenya, particularly in counties such as Uasin Gishu, Elgeyo-Marakwet, Bomet, Baringo, Nandi, and Kericho. The language is also spoken by smaller populations in parts of Uganda and Tanzania (Naibei and Lwangale, 2018). The Kalenjin people are part of the larger Nilotic group that migrated from the Nile Valley to the East African Great Lakes region thousands of years ago (Chelimo and Chelelgo, 2016). The migration process,

which began around the fifteenth century, saw the Kalenjin spread across the Rift Valley and into the surrounding areas. As a member of the Eastern Nilotic group, Kalenjin shares linguistic features with other languages such as Turkana, Pokot, and Maasai, all of which are spoken by ethnic groups in the same region (Naibei and Lwangale, 2018). Kalenjin holds significant importance in the cultural and social life of its speakers. It is used for daily communication, in traditional ceremonies, and in storytelling, with oral literature being a vital part of Kalenjin cultural identity. However, similar to many African languages, Kalenjin use in formal domains like education, governance, and media is limited. In Kenya, English and Kiswahili dominate in these sectors, relegating Kalenjin primarily to informal communication (Ogot, 2002). Kalenjin, like other indigenous languages in Kenya, faces the challenge of language shift, where younger generations tend to prefer English and Kiswahili for social mobility and economic opportunities. Despite this, the language remains a central part of Kalenjin identity and continues to be passed down through generations, particularly in rural areas.

2.3 Dholuo (ISO 639-3 Code: luo)

Dholuo is a Nilotic language belonging to the Western Nilotic branch of the Nilo-Saharan language family. It is primarily spoken by the Luo ethnic group, which resides predominantly in the western parts of Kenya, along the shores of Lake Victoria, and parts of Tanzania and Uganda (Omulo and Williams, 2018). The language traces its origins to the migration of Nilotic peoples from southern Sudan, who gradually settled in the Great Lakes region of East Africa around the 15th century (Heine and Nurse, 2000). As a Western Nilotic language, Dholuo shares linguistic similarities with other languages in that family, such as Acholi and Lango spoken in Uganda and South Sudan.

Dholuo plays an important role in the cultural and social life of the Luo people, serving as a medium of communication in daily life, traditional practices, and artistic expressions such as music and oral literature. However, its usage is restricted to informal settings, as English and Kiswahili dominate formal domains such as education, governance, and commerce in Kenya (Nduati, 2016).

2.4 Challenges Facing Low-Resource African Languages

The challenges faced by the three languages in this study exemplify common barriers that prevent low-resource African languages from becoming digitally viable for NLP applications. Drawing from the experiences with Kidaw'ida, Kalenjin, and Dholuo, we identify the following critical gaps (Mbogho et al., 2025):

1. Insufficient Digital Resources: Comprehensive text corpora, speech recordings, and annotated datasets are scarce or non-existent. Without substantial linguistic data, it is challenging to develop effective NLP applications such as speech recognition, machine translation, and language generation.

2. Limited Research Capacity: Low-resource language research typically requires speakers of the language with relevant technical training. Smaller language communities face challenges in finding researchers with both linguistic knowledge and NLP expertise to spearhead such initiatives.

3. Lack of Institutional Support: Formal domains such as education, governance, and media predominantly use dominant languages (English and Kiswahili in Kenya), relegating indigenous languages to informal communication. Language policies, while sometimes supportive on paper, often lack practical implementation mechanisms (Awuor, 2019).

4. Community Engagement Challenges: While there is initial excitement, sustaining long-term community participation in language data collection projects requires consistent engagement and appropriate remuneration. Volunteers need motivation beyond linguistic preservation alone.

5. Digital Marginalisation Risk: Without substantial efforts to build linguistic resources and NLP tools, low-resource languages face digital extinction, where speakers are excluded from AI-powered communication technologies and digital services.

6. Resource Scalability: Current African language projects generate modest datasets (tens of thousands of tokens or hundreds of hours of

speech), while state-of-the-art language models require massive datasets in the order of trillions of tokens (Liu et al., 2024). Methods must be scalable and sustainable.

3 Related Work

Many African languages are primarily oral, with limited written materials and even less available electronically. This means that standard approaches to the harvesting of online data, such as web crawling, are excluded in many cases. Therefore, for many African languages, researchers must rely on other methods to collect data, such as fieldwork, community involvement, and collaborations with native speakers. By actively seeking and collecting data through these means, researchers can ensure that their work is inclusive and representative of the diverse linguistic landscape of Africa. This section reviews a few examples of other researchers' approaches to building African-language corpora from which we drew inspiration for our own project.

Kencorpus (Wanjawa et al., 2023) is a corpus of Kiswahili, Dholuo, and Luhya, three of the most widely spoken languages in Kenya. The corpus includes text documents, voice files, and a question-answering dataset. Data collection was conducted by researchers who were deployed to communities, schools and media houses. Kencorpus collected 4,442 text sentences and 177 hours of recorded speech. From the raw data, the project annotated Dholuo and Luhya texts with part-of-speech tags and created question-answer pairs for the Kiswahili corpus. In addition, a parallel corpus was created by translating the Dholuo and Luhya datasets into Kiswahili.

Adelani et al. (2021) used online news sites to develop named entity recognition (NER) datasets for ten African languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian-Pidgin, Swahili, Wolof, and Yorùbá. The data consisted of about 30,000 sentences that were annotated for the NER task. The population sizes of the speakers of the ten languages range from 4 million for Dholuo to 98 million for Kiswahili. Volunteers annotated the raw data with entity labels for person, location, organisation and date. The annotators were part of the Masakhane community and were not paid but were included as co-authors on the paper. Multiple annotators of the same entities provided reliability and quality assurance. This approach offers a viable alternative to the more common paid or unpaid crowd-sourcing approach. Volunteers who are part of a community such as Masakhane that has been organised for language work are already motivated and knowledgeable and may be better prepared to participate.

Nakatumba-Nabende et al. (2024) developed text and speech resources for four Ugandan languages and Kiswahili. The four Ugandan languages are Luganda, Runyankore-Rukinga, Lumasaba, and Acholi. Some of the text data were created by translating English text into the five African languages. The English text was obtained from various published sources, including English Wikipedia, social media, online local newspapers, story books, novels and human rights charters. In addition to generating data for the African languages, this process also generated parallel corpora for English and the African languages. Additional text was obtained by recruiting native speakers through crowd-sourcing to write sentences and review them for quality. Due to a lack of access to computers, some contributors wrote their sentences by hand, which had to be typed later. Text sentences were later read by native speakers, again through crowd-sourcing, and recorded to generate speech corpora. Five parallel text corpora were created consisting of 40,000 sentence pairs each. The project also recorded 582 hours of Luganda speech and 1100 hours of Swahili speech.

Ogayo et al. (2022) built a speech synthesis dataset for 11 African languages: Dhouo, Lingala, Kikuyu, Yoruba, Hausa, Luganda, Ibibio, Kiswahili, Wolof, Fongbe and Suba. The data consisted of a total of 65,537 utterances distributed unevenly across languages, ranging from 125 to 11,971 utterances. These utterances corresponded to a total of 113.84 hours of recordings, ranging from 0.33 to 24.82 hours. Religious texts and other online and print sources were used as data sources. The data were supplemented with contributions from recruited participants. To help encourage community participation, comprehensive guidance on data collection methods and data licensing was provided. Due to the quality of speech required for speech synthesis, voice talents were carefully vetted and remunerated in cash and kind.

The examples discussed in this section give an idea of the kinds of activities that are going on in corpus building for African-language NLP, and suggest approaches that have been shown to work well and that others can emulate. It is essential to note that these examples focus on specific re-

gions and languages and only partially represent Africa's linguistic diversity. They also offer only a glimpse into the significant amount of ongoing NLP work (including corpus building) in Africa. As momentum in African language technologies builds, researchers need to be intentional in ensuring that no language is left behind and that global inequalities are not replicated locally.

As these examples show, current African language projects, regardless of the methods used, only generate modest amounts of data, typically tens of thousands of tokens or hundreds of hours of recorded speech. On the other hand, the latest advances in NLP are in large language models (LLMs), which require massive datasets. Llama 2 from Meta was trained on 2 trillion tokens (Touvron et al., 2023); OpenAI's latest LLM was trained on even more data than this, although the specifics have not been officially disclosed. Therefore, it is important that the methods employed for data collection are scalable and sustainable to support more expansive research efforts across the continent. This is a significant challenge that requires innovative solutions and we hope that our work makes a positive contribution to the available approaches.

4 Our Methodology

Our main objective was to collect text and speech data for three indigenous Kenyan languages, namely Kidaw'ida, Kalenjin, and Dholuo. These languages were chosen because they are the home languages of the research team. The project proposed to collect data from members of the language communities through crowd-sourcing. The data collection was realized through a grant from the Lacuna Fund, which made it possible to pay small stipends to compensate language data contributors for their time, effort and the knowledge shared.

4.1 Text Data Collection

Two approaches were used to generate text data:

1. The contributors wrote down sentences in the three languages covered by this project. These could be made-up sentences or sentences from public-domain materials. Some contributors wanted to use the bible, but most local-language bibles in Kenya are copyrighted by the Bible Society of Kenya and the British and Foreign Bible Society. However, we advised that they could use such copyrighted materials for inspiration. For example, a Bible story could serve as a prompt and remind them of something else they could write about. The issue of cultural appropriateness of data is often cited as militating against using texts of foreign origin, but we propose that such texts can catalyse ideas.

2. Contributors recorded conversations in the three languages, and the conversations were later transcribed. Participants in conversations were informed of the recording in advance and asked to sign a consent form. During transcription, words that were in a language other than the target language were replaced, as code-switched speech was outside the scope of the project.

The sentences collected were translated into Kiswahili to generate three parallel corpora. Microsoft Excel and Google Sheets were used to compose and translate sentences. The resulting spreadsheets were stored on Google Drive and GitHub throughout the duration of the project.

Quality assurance was achieved by identifying contributors with high levels of language proficiency who were designated as Data Collection Leads (DCLs) and paid a monthly salary. The DCLs identified language contributors they knew were qualified and recommended them for recruitment. DCLs also contributed data and checked the contributor data for correct spelling, grammar, fluency, and proper translation into Kiswahili. The DCLs corrected any errors they found in the data.

To increase inclusion in NLP work, 7 of the DCLs selected were female and 5 were male, and 5 of the researchers were female and 1 was male. Women were also well represented among the contributors, as can be seen in Table 1. It is important to have gender balance in language projects to ensure that applications work for everyone and also to avoid bias, for example, in generative AI.

4.2 Voice Data Collection

To enable our geographically distributed contributors to easily make voice data contributions easily, we determined that an open-access online platform would be ideal. Quite surprisingly, we were only able to find two options: Living Dictionaries (`https://livingdictionaries.app`) and Mozilla Common Voice (`https://commonvoice.mozilla.org`).

Initially, it seemed that Living Dictionaries would be easier to use. Mozilla Common Voice involves a rather steep on-ramp, which we explain below. Living Dictionaries, although easy to get started with, turned out not to be fit for purpose as the main aim of that platform is language documentation and it is tailored for linguistic work rather than NLP projects. For example, when recording voice for NLP, the same phrase is required to be read by multiple people. This is not well supported in Living Dictionaries as the interest there is in capturing how a phrase is spoken and having one person speak it suffices. On the other hand, in NLP and, more specifically, in automatic speech recognition (ASR), the different characteristics of people's voices while speaking the same phrase must be captured, so that applications built on those data can work for everyone.

Thus, we ultimately settled on Mozilla Common Voice. The first step was to make a language request. Once this request was approved for each of our three languages, we had to localize the pages associated with our languages using a platform provided by Mozilla known as Pontoon (https://pontoon.mozilla.org). This involved translating more than one thousand technical strings into each language, a task requiring both technical knowledge and language skills, including the ability to create suitable terms for new technical concepts such as "database", "download" or "speech recognition". This process was quite time-consuming and was a necessary step before launching the language on Mozilla Common Voice. After launch, we uploaded the text sentences collected earlier and once enough text was available on the platform, we could start reading and recording. Our DCLs recruited more volunteers to read and record, ensuring the diversity of voices needed for speech data.

The DCLs provided continuous quality assurance by reviewing and correcting, as necessary, both the text and the audio data simultaneously with the data contribution.

5 Results

5.1 Text Data

We collected 30,000 text sentences for each of the three languages and had contributors proficient in both their mother tongue and Kiswahili provide translations. This resulted in the creation of three parallel corpora: Kidaw'ida-Kiswahili, Kalenjin-Kiswahili and Dholuo-Kiswahili. These parallel corpora are freely available for download from https://zenodo.org/records/13355021 (Mbogho et al., 2025). The same repository remains on Github, where it can be updated and reuploaded to Zenodo. This is important not only for the continual expansion of the datasets, but also for making any needed corrections and quality improvements.

To analyze the translation consistency of the corpora, we used two measures. The first is average Sentence Length Ratio that computes the ratio of average sentence length of the target language to the average sentence length of the source language. This measure shows whether translations expand or contract compared to the originals. A number between 0.8 and 1.3 is generally considered indicative of good-quality translations. For the three corpora in this study, the average sentence length ratio falls in the range 1.0-1.2, indicating that target sentences are generally longer by up to 14% for Dholuo-Kiswahili and 13% for Kidawida-Kiswahili. For Kalenjin-Kiswahili, the expansion is minimal (the source and target sentences are generally of the same length) with a percentage less than 1% (see Table 2).

Table 2 also shows the sentence length correlation computed using Pearson's r measure. This shows how strongly sentence lengths correspond across the two languages in a parallel dataset. It measures translation consistency where values above 0.8 show a likelihood of good alignment and consistent translation length patterns while values lower than 0.5 show possible alignment problems or major structural differences between languages. The three corpora show consistent alignment patterns between Kiswahili and the corresponding local languages with values over 0.8, except Kalenjin, which has about 0.75.

The above insights will be useful when machine translation models are trained and the corresponding applications are developed.

5.2 Voice Data

The collection of voice data is ongoing, but currently the speech data on Mozilla Common Voice is as shown in Table 2.

Like gender, age is an important consideration for voice data collection as a person's voice normally changes with age. For this reason, age categories should span a wide range and should match the distribution in the general population as much

Measure	Dholuo-Kiswahili	Kalenjin-Kiswahili	Kidaw'ida-KIswahili
Average sentence length ratio	1.148	1.002	1.136
Sentence length correlation	0.893	0.747	0.842

Table 1: Translation consistency analysis

Language	Hours	Speakers	Female	Male
Kidaw'ida	56	24	60%	40%
Kalenjin	92	41	70%	30%
Dholuo	120	44	58%	42%

Table 2: Speech Data on Mozilla Common Voice

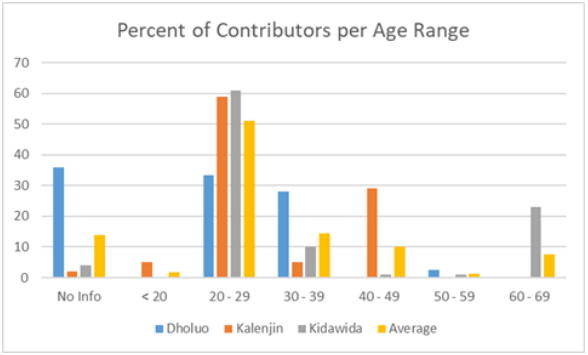

Figure 2: Age range of voice contributors

as possible. Figure 1 shows the age distribution of the voice contributors in our project. Most of the contributors were between 20 and 29 years of age. Thus, there is room for improvement in this regard as we plan for expansion of the work.

6 Discussion

This work was guided by three research questions. We aimed to find out what methods are effective in acquiring data for building language corpora for low-resource languages; how to motivate communities to contribute to language data; and how to ensure quality in the collected data.

We have referred to the type of crowd-sourcing we used as "selective" because the contributors were known to members of the project team. The researchers recruited DCLs they knew and, in turn, DCLs recruited contributors from among their close contacts. This was important because language proficiency was a key consideration, and using a "random" crowd could have compromised data quality. We found selective crowd-sourcing to be a useful approach for developing African language corpora, balancing rapid data generation with quality assurance. However, this approach may not scale well to larger projects. In that case, a project may have to accept data of varied quality and then apply additional quality control measures more aggressively in a subsequent phase.

With respect to motivating communities to participate in language projects like ours, the lesson from this project is that remuneration is critical. Some of our contributors were going through periods of unemployment, and this is what allowed them the freedom to take part. It is important to provide reasonable compensation for contributors' time, effort and linguistic expertise. We also found that many people with whom we spoke did not have a sense of the importance of their own language. Through our project, we have anecdotal evidence that there is increased awareness of the value of African languages and of the ways communities can participate in corpus building efforts.

Quality in the collected data remains a challenge. Although most of the data we have produced in this project is of good quality, we do sometimes find low quality contributions in the datasets. The work of reviewing and correcting is thus ongoing. We attribute this challenge to a lack of seriousness towards the task among a minority of the contributors. We speculate that the necessary compensation we have emphasised above can be a double-edged sword; a project may attract participants because of the pay, rather than any commitment to providing quality language contributions. However, we are encouraged by the fact that the majority of participants were serious about the task and felt they had a personal stake in promoting their own language. Educating African communities about the value of their languages and their potential to make important contributions should be a priority in language projects.

7 Conclusions and Future Work

The project ran for a year and generated approximately 90,000 sentence pairs for the parallel text corpora Kidaw'ida-Kiswahili, Kalenjin-Kiswahilli and Dholuo-Kiswahili. We also collected 268 hours of recorded speech for Kidaw'ida, Kalenjin and Dholuo by the time of the project's conclu-

sion. The researchers are actively seeking more funding to continue the work and extend it to additional African languages. We recommend that the datasets be used in their current state to train models to establish a baseline that can serve as an indication of the level of accuracy currently achievable. It is also important to document what is possible with small datasets as not all applications require massive amounts of data.

The datasets are available on open-access platforms, namely, Zenodo and Mozilla Common Voice. Anyone who wishes to download the data from either platform can do so at no cost and with minimal barriers, as the licences on both platforms are highly permissive. Developers are encouraged to take advantage of this unrestricted access to the data to train models and develop applications in these three languages. Language communities are encouraged to continue adding to the repositories to achieve greater accuracy in future models.

The project demonstrates how grassroots efforts in corpus building can support the inclusion of African languages in artificial intelligence innovations. In addition to filling resource gaps, these corpora are vital in promoting linguistic diversity and empowering local communities by enabling Natural Language Processing applications tailored to their needs. As African countries like Kenya increasingly embrace digital transformation, developing indigenous language resources becomes essential for inclusive growth. We encourage continued collaboration from native speakers and developers to expand and utilise these corpora.

Limitations

One limitation of this project is that we did not make sufficient effort to recruit participants in all age groups. Most of them fell in the 20-29 age range. This would limit the performance of automatic speech recognition models trained on the data. However, we expect that continued contribution from the wider community will give rise to greater representation with respect to age. Still, it is important to pay attention to the age aspect in future work.

Licensing issues are a major concern in corpus building for low-resource languages. A greater involvement of intellectual property specialists is essential to ensure that language communities are not disadvantaged but rather stand to benefit from the resulting data and associated applications. This consideration was precluded by the available funds.

As mentioned earlier, even though we made sure that each contributor was known to someone in the project team, we still had participants who did not take the work seriously and contributed low-quality data. In future projects, we will prioritise stronger vetting of contributors and recruitment of a larger quality assurance team.

Acknowledgments

This work was carried out with support from Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, Canada's International Development Research Centre, and GIZ on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ). The authors also wish to thank the Kidaw'ida, Kalenjin and Dholuo language communities for sharing their languages with the world.

Declarations

The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Quin Elizabeth Awuor. 2019. Language policy in education: The practicality of its implementation and way forward. *Journal of Language, Technology & Entrepreneurship in Africa*, 10(1):94–109.

Ayo Bamgbose. 2011. African languages today: The challenge of and prospects for empowerment under globalization. In *Selected proceedings of the 40th annual conference on African linguistics*, pages 1–14. Cascadilla Proceedings Project Somerville.

Louis-Jean Calvet. 1974. Linguistique et colonialisme. *Petit traité de glottophagie, Paris, Payot*.

Florence J. Chelimo and Kiplagat Chelelgo. 2016. Precolonial political organization of the kalenjin of kenya: An overview. *International journal of innovative research and development*, 5.

François Chollet. 2021. *Deep learning with Python*. Manning.

Bernd Heine and Derek Nurse. 2000. *African languages: An introduction*. Cambridge University Press, United Kingdom.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.

Audrey Mbogho, Quin Awuor, Andrew Kipkebut, Lilian Wanzare, and Vivian Oloo. 2025. Building low-resource african language corpora: A case study of kidawida, kalenjin and dholuo. *arXiv preprint arXiv:2501.11003*.

A.F. Mcharo. 1995. *Mizango na Maza ra Kidawida ra Kufuma Kokala*. Peter Bostock, Kenya.

Faith K Naibei and David Lwangale. 2018. A comparative study of the kalenjin dialects. *International Journal of Academic Research in Business and Social Sciences*, 8(8):476–503.

Joyce Nakatumba-Nabende, Claire Babirye, Peter Nabende, Jeremy Francis Tusubira, Jonathan Mukiibi, Eric Peter Wairagala, Chodrine Mutebi, Tobius Saul Bateesa, Alvin Nahabwe, Hewitt Tusiime, and 1 others. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: Experiences and lessons. *Applied AI Letters*, 5(2):e92.

Rosemary N Nduati. 2016. *The post-colonial language and identity experiences of transnational Kenyan teachers in US Universities*. Ph.D. thesis, Syracuse University.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building tts systems for low resource languages under resource constraints. In *Proc. S4SG 2022*.

Bethwell A Ogot. 2002. Historical portrait of western kenya up to 1895 bethwell a. ogot. *Historical Studies and Social Change in Western Kenya: Essays in Memory of Professor Gideon S. Were*, 13(4):99–120.

Albert Gordon Omulo and John James Williams. 2018. A survey of the influence of 'ethnicity', in african governance, with special reference to its impact in kenya vis-à-vis its luo community. *African Identities*, 16(1):87–102.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2023. Kencorpus: A kenyan language corpus of Swahili, dholuo and luhya for natural language processing tasks. In *Journal for Language Technology and Computational Linguistics, Vol. 36 No. 2*, pages 1–27, unknown. German Society for Computational Linguistics and Language Technology.