BUILDING TEXT AND SPEECH DATASETS FOR LOW RESOURCED LANGUAGES: A CASE OF LANGUAGES IN EAST AFRICA

Anonymous authors
Paper under double-blind review

ABSTRACT

Africa has over 2000 languages; however, those languages are not well represented in the existing Natural Language Processing ecosystem. As language technologies become more advanced, African languages still lack essential digital resources to be engaged effectively. This growing gap has attracted researchers to empower and build resources for African languages to transfer the various Natural Language Processing methods to African languages. This paper discusses creating, curating, and annotating language text and speech datasets for low-resourced languages in East Africa. This paper focuses on five significant languages: Luganda, Runyankore-Rukiga, Acholi, Lumasaaba, majorly spoken in Uganda, and Kiswahili which is majorly spoken across East Africa.

Speech, Text, Luganda, Common Voice, ASR, Swahili

1 INTRODUCTION

From smart speakers to smartphones, speech recognition has revolutionized how we interact with machines. While technology continues to mature and become more ubiquitous, we still see significant barriers to innovation for most of the world's languages. Existing speech recognition services are only available in major languages such as English, French, and Germany. Neither Amazon's Alexa, Apple's Siri, nor Google Home, the leading players in the global voice assistants market, supports a single native African language. The speech data held by these tech giants are still proprietary, which stifles innovation.

Automatic Speech Recognition (ASR) systems automatically convert human speech into transcribed text. These systems can transcribe audio and video data, allowing for multimedia data analysis. Typical ASRs work with models trained by machine learning algorithms. However, these Artificial Neural Network-based approaches require large amounts of data to produce accurate results (Alsharhan & Ramsay, 2020). Therefore, to develop a speaker-independent and high-quality performance ASRs, an extensive collection of speech samples must be gathered from various speakers along with the transcriptions of those speeches. Therefore, collecting extensive and balanced amounts of text and speech data is necessary to build functional ASR systems (Alsharhan & Ramsay, 2020).

Uganda is a culturally diverse and multilingual country with about 41 indigenous languages (Ssentanda & Nakayiza, 2017). However, the steady advance of speech recognition technology has not benefited any of these languages classified as low-resource for ASR applications. The adoption of speech technology is due to the limited amount of available transcribed speech required for training ASR models, which is a fundamental problem because technology adoption is gaining momentum in Uganda. Many services are going digital, notably government services, which are now only accessible online. While this is a positive development, it poses significant challenges to the providers and recipients of these services. There is a slow increase in platform adoption because of digital illiteracy. There is thus a need to build ASR applications for these low-resourced languages.

The existing limited text data resources are hard to discover, primarily since these are published in closed journals or not digitized (Wilhelmina et al., 2020). Furthermore, the lack of sufficient language resources for these low-resourced languages has affected the advancement of research in fields of Natural Language processing such as text classification, text summarization, information

retrieval, and machine translation. This paper describes a practical process to create speech and text data for four low-resourced languages in East Africa.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 describes the motivation taken in creating the text and voice datasets. Section 4 and 5 describe the collection process and characteristics of the text and speech corpus respectively. Section 6 describes the tools used to collect the data and also supports the creation of the parallel text corpus. Section 7 describes the ethics and legal considerations adhered to while creating the data. Section 8 discusses the key challenges faced and Section 9 concludes the paper.

2 Related Work

Previously, efforts have been made by different researchers to create and curate high-quality text and speech datasets for low-resourced languages. The work in Mukiibi et al. (2020) presents an agricultural-specific speech keyword dataset. The dataset contains 193 keywords, of which 129 are Luganda while 64 are in English. The keyword list was compiled from agricultural radio talk show recordings and data from online Luganda newspapers. This 193 keyword dataset contains different speech instances categorized into crops, diseases, fertilizers, herbicides, and general agriculture-related keywords.

Furthermore, efforts have been made to create and curate a news headline dataset on two low-resourced languages in South Africa; Setswana and Sepedi Vukosi et al. (2020). The dataset included 219 news headlines in Setswana and 491 news headlines in Sepedi. These were categorized into Legal, General News, Sports, Others, Politics, Traffic News, Community Activities, Crime, Business, and Foreign Affairs.

The work in David et al. (2016) developed a mobile device application for collecting parallel speech data for under-resourced languages such as Mboshi, Myene, and Basaa. Similar work has been done for Kinyarwanda which was the first African language launched on Common voice platform[1]. The Kinyarwanda team built their speech and text data, which currently has about 2,300 hours of recorded voice and 1,900 hours of validated voice.
Kinyarwanda, a widely spoken language in Rwanda with over 12 million speakers[2]. In 2019, Mozilla and Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) co-hosted an ideation hackathon in Kigali to create a data corpus for Kinyarwanda and to lay the foundation for local voice-recognition applications, which gave way to Digital Umuganda. Digital Umuganda brought the concept of communities to the digital edge to help in building digital infrastructure such as voice data.

Digital Umuganda created a team of "Commoners" volunteers who help mobilize contributors and manage offline events. Communities were formed in schools, although these were closed due to Covid-19. They then proposed venues to provide free internet for contributors living in the same neighborhood and send internet packages to those who could not make it to this venue. Digital Umuganda continued to use its core contributors as commoners to keep up the momentum by adding more contributions and mobilizing other contributors. They have now set up a hybrid model, on-site and off-site recording, by leveraging experience in organizing and managing events and involving commoners to mobilize online events contribution [3].

3 Motivation

Uganda has a diverse culture with a population made up of different ethnic groups with unique customs and norms[4]. The country is multilingual with forty-one native languages used by the inhabitants across the country (Ssentanda & Nakayiza, 2017). All native languages in Uganda, including the languages of our research focus; Luganda, Runyakore-Rukiga, Acholi, and Lumasaaba mainly used in the central, western, northern, and eastern regions, respectively, are low resourced languages.

[1]https://foundation.mozilla.org/en/blog/how-rwanda-making-voice-tech-more-open/
[2]https://en.wikipedia.org/wiki/Kinyarwanda
[3]https://commonvoice.mozilla.org/rw/
[4]https://en.wikipedia.org/wiki/Languages_of_Uganda

These languages lack large monolingual and/or parallel corpora to build NLP applications. For advances to be made towards applying AI and/or NLP-aided tools for (1) preservation of African languages in general, (2) building of educational tools for communities with lower literacy levels, (3) monitoring of radio broadcasts for topics of public interest, emergency response, for example, in the current COVID-19 pandemic where access to vulnerable people is reduced, and access is restricted [5]and building voice recognition technology[6], the first step is the creation and curation of high-quality datasets for these languages.

Despite the vast array of native languages in Uganda, the reach of these languages at a regional level in East Africa is quite limited. Therefore there is a need to collect data on native languages with a much broader reach. Although Kiswahili is natively spoken in Kenya and Tanzania, it has widely been adopted as a regional language by The East African community [7]. It is increasingly being used to facilitate cross-country communication. As an extension to creating corpora for native Ugandan languages, this research seeks to create a Swahili text and audio corpus to support the development of NLP-powered solutions with the Kiswahili language.

4 TEXT DATA

4.1 DATA SOURCES

The first step in creating voice datasets was the curation of large corpora of text datasets. The text corpora in this project was sourced from the different platforms used by the local population to diffuse information, promote awareness and facilitate communication. These included social media platforms (Facebook and Twitter), online Uganda newspapers (New Vision, Independent Magazine, Bukedde), and available online articles like Wikipedia. The other rare data source considered was the storybooks. The choice of these sources was based on: the diversity of the data around the local context. However, some of the data sources also had open-source data that could be openly published. All of these sentences were in English and needed to be translated to the local languages. To increase the Luganda dataset we also had to collect Luganda text data from other sources. These included: content creators, for example Luganda teachers, Luganda online newspapers Gambuuze[8], and the Buganda Land Board which is a Luganda newspaper specifically for Buganda land issues.

The Kiswahili text data was collected in Tanzania and Kenya from two primary sources: pre-existing curated text data and contributions from participants in the community. Pre-existing sources include news websites, published reports, published storybooks, and published open-source Swahili text datasets. The community contributions are original creations of the Kiswahili text made by individuals participating in the project.

4.2 DATA PRE-PROCESSING AND PREPARATION

The text dataset generated was unique because it was obtained from both structured (mass media, online outlets) and unstructured sources (social media outlets). Whereas the structured dataset was less erroneous, the unstructured dataset had a lot of noise and redundant information since people communicate informally. However, this whole set had to be preprocessed so as to be able to extract useful information from topics of interest (local events) to statements that carry peoples' sentiments. After preprocessing, the data was prepared through a tweaking exercise to generate a text corpus that we will publish under the Creative Common license (CC- 0).

4.2.1 DATA PRE-PROCESSING

The data preprocessing process was unique for the two kinds of data sources we had in this project. The data preprocessing process for the social media text involved using a Python script to remove any unwanted characters such as URLs, email addresses, mentions, and hashtags. After which, the

[5]https://medium.com/sciforce/nlp-for-low-resource-settings-52e199779a79

[6]https://reliefweb.int/report/world/twb-and-kobo-inc-develop-speech-recognition-technology-capture-voices-speakers

[7]https://en.wikipedia.org/wiki/Swahili$_language$

[8](https://gambuuze.ug/)

team carried out a sentence generation process where the original sentence was used as a basis for creating a new sentence while maintaining the context of the original sentence.

The data preprocessing steps specific to the Kiswahili text included removing texts with English words, abbreviations, and digits such as dates and removing sentences with more than 14 words. This process ensured that the Swahili text data was not interpreted ambiguously. One such example is when the text containing an abbreviation is being read, the abbreviations could be represented differently, which can negatively impact the quality of voice datasets collected with this data.

4.2.2 DATA PREPARATION

The data preparation step majorly involved creating a text corpus that can be publicly made available. Part of the text data could not be used in its original form since it is copyrighted, and thus this was only used as a baseline to create new sentences while preserving the context therein. The sentence creation activity was carried out by a group of university students and researchers proficient in the English language. Each person was given a set of original sentences (the sets varied in size, minimum being 100 sentences) and was required to create new data instances on the same topic of discussion. The sentences were then reviewed at two levels to ensure de-identification from the original sentence; to ensure that the sentences were well punctuated and meaningful. Figure 1 shows an example of a sample sentence created based on the original sentence.

Original Sentence	**New Sentence**
The Bill seeks to reduce the withdraw tax on mobile money from 1 to 0.5%.	The proposal focuses on reducing the tax on mobile money withdrawals.

Figure 1: A sample pair of a generated English Sentence. The senntence on the left hand side is the original sentence while the sentence on the right hand is the created sentence.

4.3 TRANSLATION

The source English text data was translated into the four Ugandan languages via a translation management system. The aim was to create a parallel text corpus that can be used to build NLP applications such as machine translation and support the design of other NLP tools for these specific low-resourced languages.

4.4 QUALITY ASSURANCE

Guidelines were developed to support the English data collection process's sentence creation and review processes. The guidelines captured the following concerns: (a) syntax and semantics of the data, (b) creation of unbiased data (race, religion, gender, political, culture), (c) de-identification of the new sentences from the source sentences (d) preservation of the context or topic of discussion in the parallel source sentences. Furthermore, the guidelines were developed and emphasized to ensure that the generated text data could be efficiently used when creating the speech dataset. These included: (1) The generated sentences had to be speakable – this was supported by creating sentences with a given length and replacing complex words with simpler words. This was important because the longer the sentence, the harder it is to memorize and read it out. The required length for each sentence was four to fourteen words. (2) The sentence had to be natural, a representation of the local context. (3) The sentence had to contain no abbreviations; to avoid creating ambiguous sentences. (4) Lastly, English sentences had to be both syntactically and semantically correct.

The quality assurance guidelines for the Kiswahili text were adopted from the text requirements on the Common Voice platform. These are designed to ensure that the text data is valid for voice contributions. The process involves a selection of a sample to be validated, a manual check by native language readers to identify different issues within the sample, and the calculation of the error rate for the validated sample. For a given batch of sentences, the parameters for selecting the sample size are a margin of error of 2% and a confidence level of 99%. The calculated sample determines the size of randomly selected sentences to be validated. For the Kiswahili text, at least 30% of the sample was randomly selected from each of three pre-defined groups of sentences; sentences with less than five words, sentences with 5-8 words, and sentences with 9-14 words. This was done to

ensure a fair representation of the different sentence lengths. Three validators score each sentence in the validation sample against four types of issues; incorrect spellings, incorrect grammar, difficulty to speak and non-native words, and any other type of issue they can identify. The error rate for a sample is calculated as the percentage of sentences in the sample that have any of the issues identified by the validator. The error rate for the entire validation sample is the average error rate reported by validators, and for a batch to be considered valid, this value should not exceed 5%. We conducted four validation exercises for the Swahili text with an average error rate of 3.14%.

5 SPEECH DATA

In this research, the speech data was collected for one native language in Uganda, Luganda, and the Kiswahili Language, a widely spoken language in the East African community. The speech data was collected using Mozilla's Commonvoice[9] crowdsourcing platform. The Commonvoice platform is an initiative to make voice data freely and publicly available and make sure the data represents the diversity of people. The Commonvoice platform allows people to donate their voices by reading, recording, and uploading sentences available on the platform. It then allows people to validate the accuracy of other people's recorded clips.

5.1 COMMUNITY CROWDSOURCING APPROACH

The Makerere University Artificial Intelligence Lab led the Luganda speech data on Common Voice mobilization. The mobilization team created communities that included: University student groups and association of Luganda teachers of the Luganda Language. These communities would have regular physical meetups to read and validate sentences on the Commonvoice platform.

Due to the Covid-19 pandemic, it was no longer possible for the communities to meet as the country went into a lockdown. After consulting with the existing community leads, we agreed to give weekly data facilitation to the contributors as reimbursement for the mobile data they would have spent while reading and validating sentences. We agreed to give an equivalent of USD 1.43 in terms of airtime to each contributor per week, and in turn, they would be required to make a minimum of 100 voice contributions per week. Cash prizes were also given to the monthly top three voice contributors, the top three voice validators, and the top community mobilizers to further incentivize the contributors.

Speech data for Kiswahili was collected in both Kenya and Tanzania by leveraging on student communities at universities. These were Maseno University and Kabarak University in Kenya and the University of Dodoma, Nelson Mandela African Institution of Science and Technology, and Dar es Salaam School of Journalism in Tanzania. These communities were established to congregate different enthusiastic individuals about NLP and create and promote the Kiswahili language. The members in these communities held workshops during which they were trained on how to use the Common voice platform to contribute text and voice data. During these meetups, the communities curated over 200,000 Kiswahili sentences and 100 hours of voice contributions from 80 participants in Kenya and 90 participants in Tanzania. Creating these communities also helped to ensure that members could still actively participate even outside the meetups as they have groups and maintain communication online through platforms like WhatsApp.

5.2 CONTRIBUTIONS

Speech data is measured in terms of the number of hours that have been recorded and validated. Since the launch of Luganda on the Commonvoice platform in September 2020, we have recorded a total of 406 Luganda validated hours from over 486 unique contributors. The demographics of the dataset contribution include: 41% (19 - 29), 22% (30 - 39), 5% (40 - 49), 3% (50 - 59), and 1% (60 - 69) and 1% (<19). The dataset has 40% female contributions and 34% male contributors.

Kiswahili has recorded a total number of 146 validated hours from over 288 uniqgue contributors[10]. The demographics of the dataset contribution include: 45% (19 - 29), 15% (30 - 39), 6% (50 - 59), 5% (40 - 49), and 1% (60 - 69). The dataset has 41% male contributions and 34% female contributors.

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

[10] https://commonvoice.mozilla.org/en/datasets

6 SOFTWARE TOOLS USED IN THE PROJECT

6.1 SENTENCE CROWDSOURCING APPLICATION

A web application was developed to help in collecting more monolingual sentences in Luganda using the crowdsourcing approach. With this approach, a group of linguists was sourced to carry out the sentence creation task using the application. The application has an administrator interface, which facilitates the generation of reports on daily, weekly, and overall submissions.

Application features

1. User account creation feature. One can only access the application features once they are logged in.
2. Sentence categorization feature in the different topics of interest such include Education, Sports, Politics, Entertainment, COVID-19, Agriculture, Culture, Gender-Based Violence, Business, Health, Land, Legal, Transport, Security, Technology, and Others (in scenarios were the sentence creator cannot determine the subject of the sentence). The linguist does this categorization while they are creating the sentences.
3. Sentence validation feature: The application also facilitates sentence validation to ensure the created or uploaded sentences meet the guidelines defined in the protocol. Some of the defined rules include: (1) The sentence should be more than three words and less than fifteen words. (2)The sentence should not contain digits. (3) The sentence should be a new or unique entry. Each new instance is compared with all the other existing sentences in the database to ensure no duplicates.
4. Bulk upload feature - The application also facilitates Bulk uploading sentences via an external file that must be a word document. These sentences are validated using the defined rules in the protocol as mentioned above.

Specifically for Tanzania, a web-based tool was created to enable users to make and submit original text creations and to upload files (text and CSV) containing text data and this ensured that users could participate actively at various times and different places.

6.2 SENTENCE TRANSLATION TOOL

The sentence translation was done using Pontoon, a translation management system developed by the Mozilla localization community. The system has a web interface that facilitates the translation of the text into other languages. These languages are added to the system by creating new locale instances in the Pontoon database. The system facilitates translation and review of the translated sentences via the contribute and approve/reject features, respectively. The system has different user accounts: (1) administrator - who has complete control over the system (2) manager - who can upload text data to the system, assign batches to a contributor and also validate the contributions made by the translators, and the (3) contributor - who can translate text data.

The team leaders of the four Ugandan languages were given managerial rights to assign sentence batches to the translators and validate the translations (contributions).

6.2.1 PIPELINE FOR TEXTDATA UPLOAD ON PONTOON SYSTEM

The text data is organized under 'Projects' on the system. We describe the steps taken on how projects are created on the Pontoon System.

1. Add new project.
2. Provide project name.
3. Select locales and move them to this project. Locales are the different languages. Once the administrator has added the new languages as locales they will be depicted on this page. All locales in which the source text is to be translated into are selected at this step. This feature enables the project to be mirrored in all the locales and thus helping us generate a parallel text corpus as depicted in Figure 2, once the translated strings are downloaded.

4. Select Datasource: either from a repository (Github) or the database.
5. Make project visible by selecting 'Public' from the visibility dropdown.
6. Save project.

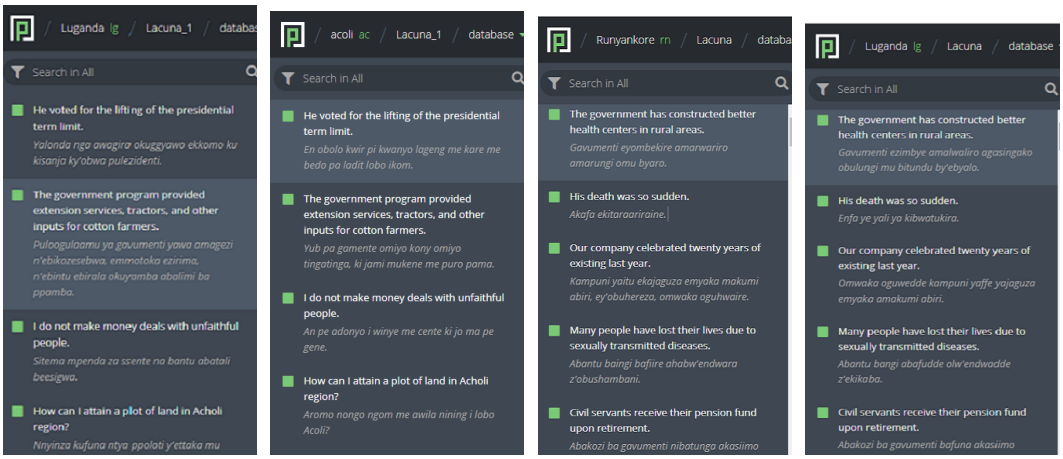

Figure 2: A figure depicting translation of the text data across the different languages

6.3 Quality Assurance Tools

The quality assurance tools included a set of protocols used at the various phases of collecting and creating a curated dataset.The protocols included sentence creation guidelines for the English source sentences, translation guidelines per language for the parallel text corpus, and sentence collection guidelines for the monolingual corpus.

7 Ethical Considerations

7.1 Copyright for Text sources

Copyright is a type of intellectual property that gives its owner the exclusive right to make copies of creative work, usually for a limited time. To make the acceptable text contributions to the Common-Voice dataset, we allowed source text available under a Creative Commons (CC-0) license[11]. Using the CC0 standard means it is more difficult to find and collect source text but allows anyone to use the resulting voice data without usage restrictions or authorization. Ultimately, we want to make the multi-language dataset as helpful as possible to everyone, including researchers, universities, startups, governments, and social purpose organizations.

7.2 Contributions

The Commonvoice database is available under the Creative Commons CC-0 public domain dedication. That means it is public, and all copyrights have been waived to the extent possible under the law. Participants in Commonvoice are required to do the same. They must agree that Mozilla may offer all of their Contributions (including text, recordings, validations, and feedback) to the public under the CC0 public domain dedication. Voice contributions are a purely voluntary activity, and thus there is no remuneration for the contributions. The contributors need to agree to this before participating.

8 Challenges

This section discusses some challenges we encountered in the text and voice creation efforts.

[11] https://creativecommons.org/share-your-work/public-domain/cc0/

- Text Data Licensing - It was challenging to get compatible licenses with CC0 for existing data sources. This implies that such data cannot be used on the Commonvoice platform unless the authors provide a CC0 waiver. Getting CC waivers was quite challenging as the license is considered too liberal by some content creators like news websites. Whereas it was easy for Swahili which has a large text corpus, it was not the case for a low resourced language like Luganda.
- Busy university schedules - The crowdsourcing communities were built around universities in East Africa. Student communities could not fully participate in the voice drives due to other ongoing university obligations. Additionally, there were delays in the translations because the linguists also had busy schedules.
- COVID-19 restrictions - The text and voice data were collected during the COVID-19 pandemic. This also came with challenges due to the travel and gathering restrictions.
- Inadequate Sentences - We experienced a situation known as over-contribution. This is a state in which the number of hours recorded on Commonvoice exceeds the recommended number of sentences. At 90,000 sentences, Luganda's recorded hours were above 300 hours. So we were forced to halt the voice contribution exercise until more sentences were uploaded to the Commonvoice platform.
- Anonymous Contributors - We also had contributors who preferred not to sign up on the Commonvoice platform. This made it difficult for us to keep track of the contributor progress in comparison to other contributors but also against personal and project goals. With such contributors we were also not able to collect demographic information such as gender and age which were important parameters in ensuring we create a rich diverse speech dataset.
- Platform related Issues The Common Voice platform came with its own challenges;
  - There was an email confirmation loop error where new users were kept in a loop of confirming their emails. This prevented many new contributors from signing up on the platform.
  - Some contributors are not tech-savvy and thus struggled to navigate the Commonvoice site. They were not able to fully utilize all the available options, such as repeating a recording listen to their recordings before uploading.
- Cost - We saw the number of contributions spike after facilitating the contributors with airtime. However, these airtime and cash prizes are not sustainable, so there is a need for continued motivation of the project participants to contribute to the project irrespective of airtime facilitation.

9 Conclusion and Future Work

Our work has focused on some of the low resourced languages in Uganda and East Africa. Working across 5 different languages unveils much about how low resourced each language is. And ideally, this is not the same across languages. We have presented different approaches used to collect both text and voice datasets for these languages. We introduced a number of data collection tools that were used to collect the data. These were used to collect and curate a parallel corpus for 4 languages. However, future work may focus on the biases that may result in using these tools across the different languages, challenges in adoption to the tools, strategies for creating representative datasets in terms of dialects and how this may affect the downstream tasks that may result in building models with this corpus.

Acknowledgments

We want to thank the Department of African Languages, Makerere University for the translation work on the parallel text corpus. We would like to thank Buganda Kingdom for partnering with us and also for the support on the monolingual text collection for Luganda through its agencies. We would also like to acknowledge the various organisations and individuals from which we sourced this data; the Independent News, Jerum Agency. We want to thank all the community contributors in East Africa who have enabled us to achieve the voice contributions. This work was carried out with

support from Lacuna Fund, an initiative cofounded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre; Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) and Mozilla.

REFERENCES

Eiman Alsharhan and Allan Ramsay. Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition. *Language Resources and Evaluation*, 54(4): 975–998, 2020.

Blachon David, Gauthier Elodie, Besacier Laurent, Kouarata Guy-Noël, Adda-Decker Martine, and Rialland Annie. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:1527–1554, 2016.

Jonathan Mukiibi, Claire Babirye, Andrew Katumba, and Joyce. Nakatumba. Agriculture keywords dataset (version one) [data set]. zenodo. *https://doi.org/10.5281/zenodo.4347308*, 2020.

Medadi E Ssentanda and Judith Nakayiza. "without english there is no future": The case of language attitudes and ideologies in uganda. In *Sociolinguistics in African contexts*, pp. 107–126. Springer, 2017.

Marivate Vukosi, Sefara Tshephisho, Chabalala Vongani, Makhaya Keamogetswe, Mokgonyane Tumisho, Mokoena Rethabile, and Modupe. Abiodun. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. 2020.

Nekoto Wilhelmina, Marivate Vukosi, Matsila Tshinondiwa, Fasubaa Timi, Kolawole Tajudeen, Fagbohungbe Taiwo, Oluwole Akinola Solomon, Muhammad SH, Kabongo S, Osei S, and Freshia S. Participatory research for low-resourced machine translation: A case study in african languages. 2020.