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

Claire Babirye, Joyce Nakatumba-Nabende, Jeremy Franics Tusubira & Jonathan Mukiibi
Department of Computer Science
Makerere University
Uganda, Kampala
clarybits68@gmail.com, joyce.nabende@mak.ac.ug
tusubirafrancisjeremy@gmail.com, jonmuk7@gmail.com

Andrew Katumba & Ronnie Ogwang
Department of Electrical and Computer Engineering
Makerere University
Uganda, Kampala
andrew.katumba@mak.ac.ug,rogwang123@gmail.com

Medadi Sentanda
Department of African Languages
Makerere University
Uganda, Kampala
medadies@gmail.com

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

Davis David
TYD Innovation Incubator
Dar es Salaam, Tanzania
davisdavid@tyd.or.tz

ABSTRACT

Africa has over 2000 languages; however, those languages are not well represented in the existing Natural Language Processing ecosystem. African languages lack essential digital resources to be engaged effectively in the advancing language technologies. 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 the process we took to create, curate and annotate language text and speech datasets for low-resourced languages in East Africa. This paper focuses on five languages. Four of the languages: Luganda, Runyankore-Rukiga, Acholi, and Lumasaaba, are majorly spoken in Uganda, and Kiswahili which is a majorly spoken language across East Africa. We have run baseline: machine translation models on the English - Luganda dataset in the parallel text corpora and Automatic Speech Recognition (ASR) models on the Luganda speech dataset. We recorded a BiLingual Evaluation Understudy (BLEU) score of 37 for the English-Luganda model and a BLEU score of 36.8 for the Luganda-English model. For the ASR experiments, we obtained a Word Error Rate (WER) of 33%.

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 signifi-

cant 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 global voice assistants market players, supports a single native African language. The speech data held by these technology 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 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). It is necessary to have an extensive collection of speech samples and transcriptions to develop a speaker-independent and high-quality performance of ASRs.

Uganda is a culturally diverse and multilingual country with about 41 indigenous languages (Ssentanda & Nakayiza, 2017). All native languages in Uganda, including the languages of our research focus; Luganda, Runyankore-Rukiga, Acholi, and Lumasaaba mainly used in the central, western, northern, and eastern regions, respectively, are low resourced languages. These languages lack large monolingual and/or parallel corpora to build NLP applications.

The steady advance in 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, text-to-speech, speech recognition, information retrieval, and machine translation. For advances to be made towards applying Artificial Intelligence (AI) and/or NLP-aided tools for example for the: (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 and building voice recognition technology[1], and (4) provision of agricultural specific services to smallholder farmers, the first step is the creation and curation of high-quality datasets for these languages. To cover this gap, this paper describes a practical process to create speech and text data for four low-resourced languages in East Africa. The paper also describes baseline models that we have developed for machine translation and speech recognition based on these datasets.

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 larger text and speech datasets on Ugandan languages with a much broader reach. Although Kiswahili is spoken in Kenya and Tanzania, it has widely been adopted as a regional language by The East African community [2]. 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 speech corpus to support the development of NLP-powered solutions with the Kiswahili language.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 and 4 describe the collection process and characteristics of the text and speech corpus respectively. Section 5 describes the tools used to collect the data and also supports the creation of the parallel text corpus. Section 6 describes the ethics and legal considerations adhered to while creating the data. Section7 discusses the benchmark results from Machine Translation (MT) and Automatic Speech Recognition (ASR) experiments based on these datasets. Section 8 discusses the key challenges faced and Section 9 concludes the paper.

[1] Kobo Speech recognition Technology

[2] Swahili Language

2 Related Work

There have been previous efforts in the field of Natural Language Processing to enable data collection, building NLP research communities, and advance NLP research efforts on the African continent. Masakhane is one of the communities that has enhanced NLP research in Africa through a participatory approach to community building Orife et al. (2020). Recently, the Lacuna fund was created to support researchers on the African continent to create openly accessible text and speech resources for NLP research. [3]. An AI4D - African language program was launched with focus on the collection and curation of language datasets Siminyu et al. (2021). One of datasets created through the AI4D program was the English-Luganda parallel text corpus Mukiibi et al. (2021), that was a starting point for the data collection efforts reported in this paper. The AI4D fellowship also enabled the creation of an agricultural-specific speech keyword dataset Mukiibi et al. (2020) that contains 5290 keyword utterances of specific keywords in English and Luganda. The Luganda keyword utterances dataset was used in a Machine Learning challenges that was hosted on the Zindi platform [4].

Efforts have been made to create and curate a news headline dataset for 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. The news headlines were categorized into several topics in the dataset. The authors in David et al. (2016) developed a mobile application for collecting parallel speech data for under-resourced languages for three languages: Mboshi, Myene, and Basaa. For the creation of speech datasets, work has been done to create openly available speech datasets using Mozilla's Commonvoice platform. Kinyarwanda was the first African language to be launched on the Common voice platform [5]. The Comonvoice Kinyarwanda dataset currently has about 2,300 hours of recorded voice and 1,900 hours of validated voice. There have been several addition of African languages on the Commonvoice platform including Luganda, Swahili and Igbo. The approach taken with voice contributions has been community mobilizations of voice contributors across the several countries where the languages are spoken.

3 Creation of Text datasets

In this section, we describe the approach taken to create the text dataset for the five languages of interest.

3.1 Data Sources

The first step in creating voice datasets is the building and curation of large corpora of text data. The text corpora in this project was sourced from the different local text sources for example, data from social media platforms, online Uganda newspapers, and available online articles like Wikipedia. Data was also obtained from Ugandan storybooks in collaboration with the linguists in the Department of African Languages. We were interested in diversity of the text sources. The initial data collection process generated text sentences in English which had to be translated to the four local Ugandan languages. We also collected more Luganda text data from other sources to increase the diversity and build a larger text corpora that was necessary for building voice resources on the Common Voice platform. The Luganda text sources included: text content created by Luganda teachers, content from Luganda online newspapers and content from the Buganda Land Board which is a Buganda agency incharge of the kingdom land.

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 included news websites, published reports, storybooks, and open-source Kiswahili text datasets. The community contributions are original creations of the Kiswahili text made by individuals participating in the project.

[3] Lacuna Language Datasets

[4] Agricultural Keyword Spotter for Luganda

[5] Kinyarwanda dataset on Commonvoice platform

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

ORIGINAL SENTENCE	NEW SENTENCE
The Bill seeks to reduce the withdraw tax on mobile money from 1 to 0.5%	The proposal is to reduce the tax on mobile money withdrawals.

3.2 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).

3.2.1 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 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 preprcessing 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.

3.2.2 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. Table 1 shows an example of a sample sentence created based on the original sentence prompt.

3.3 Sentence Translation

The source English text data was translated by a team of experts into four Ugandan languages. Efforts to translate this data into Kiswahili are ongoing. Translation was done using both online and offline tools depending on what was convenient for the team. Online tools included Pontoon, a translation management system developed by Mozilla to facilitate translation of text into other languages. Offline tools included: Microsoft excel. The translation teams were divided into two: contributors who provide the translations and the validators who reviewed and validated the translation given by the translatos.

3.4 TEXT QUALITY ASSURANCE

Guidelines were developed to support the process of the creation and review of the English dataset. 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. Such guidelines 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. 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 (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.

3.5 TEXT CONTRIBUTIONS

As a first step, we created a total of forty thousand $(40,000)$ sentences as source text data in English. The English sentences were translated into the four Ugandan languages, i.e., Acholi, Runyankore, Luganda and Lumasaaba to create a parallel corpus highlighted in Table 2. Efforts are underway to translate the English dataset into Kiswahili. We have also created monoligual corpora for the five languages of interest in this project. Table 3 shows the size of the monolingual corpora across the five languages.

Table 2: Parallel text corpora.

PARALLEL CORPORA	
Language	**Dataset size**
English - Luganda	$40,000$
English - Acholi	$40,000$
English - Lumasaaba	$40,000$
English - Runyankore	$40,000$

Table 3: Monolingual Text corpora.

MONOLINGUAL CORPORA	
Language	**Dataset size**
Luganda	$400,000$
Acholi	$40,047$
Runyankore	$40,211$
Lumasaaba	$40,184$
Kiswahili	$200,000$

4 CREATION OF SPEECH DATASET

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 crowdsourcing platform[6]. The Commonvoice platform is an initiative to make voice data freely and publicly available and make sure the data represents the

[6]Mozilla Commonvoice platform

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.

4.1 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 internet facilitation to the contributors as reimbursement for the mobile data 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.

4.2 VOICE CONTRIBUTIONS

For the Commonvoice project, the amount of speech data that was created was measured by the number of hours that had been recorded and validated. Since its launch on the Commonvoice platform in September 2020, Luganda has recorded a total of 406 Luganda validated hours from over 486 unique contributors. The Luganda Common Voice dataset was contributed by 39.2% women and 33.5% men while the remaining 27.3% were anonymous contributors. Table 4 shows a detailed breakdown of the Luganda Common Voice dataset based on the age of the contributors. Kiswahili has a total number of 146 validated hours from over 288 unique contributors platform[7]. The demographics of the dataset contribution are also shown in Table 4. The dataset has 41% male contributions and 34% female contributors while the remaining 25% were anonymous contributors.

Table 4: Demographics of Commonvoice validated corpus for Luganda and Kiswahili languages.

AGE GROUP	**PERCENTAGE (%)**	
Language	**Luganda**	**Kiswahili**
< 19	1	-
$19 - 29$	41	45
$30 - 39$	22	15
$40 - 49$	5	5
$50 - 59$	3	6
> 60	1	1
Unspecified	27	28

[7]Mozilla Commonvoice datasets

5 SOFTWARE TOOLS USED IN THE PROJECT

5.1 SENTENCE CROWDSOURCING APPLICATION

A web application was developed for collection of Luganda monolingual sentences using the crowdsourcing approach. With this approach, a group of linguists was sourced to carry out the sentence creation task. The crowdsourcing application also has an administrator interface that facilitates the generation of reports on the submitted sentences.
The application specific features include:

1. User account creation feature where one can access the application features once logged in.
2. Sentence categorization feature - where contributors are allowed to create sentences under different topics of interest: Education, Sports, Politics, Entertainment, COVID-19, Agriculture, Culture, Gender-Based Violence, Business, Health, Land, Legal, Transport, Security, Technology, and Others.
3. Sentence validation feature - The application facilitates sentence validation to ensure the created or uploaded sentences meet the guidelines defined in the sentence creation protocol. Each new instance is compared with all the other existing sentences in the database to ensure no duplicates.
4. Bulk upload feature - The application facilitates bulk sentence upload. The sentences are validated using the defined rules in the sentence creation guidelines.

For the Kiswahili sentence contribution, a web-based tool was created to enable users submit original sentence creations in either text or CSV format. This ensured that users could participate actively at various times across different places.

5.2 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 or contributions that were made.

5.2.1 PIPELINE FOR TEXT DATA UPLOAD ON PONTOON SYSTEM

As shown in Figure 1, the text data on the Pontoon platform is organized under 'Projects' on the system. The intial step is to create a new project on the platform which is assigned a project name. The different locales that correspond to the different languages to which the translations have to be made are selected. Once the administrator has added new languages as locales they are depicted on this page. All locales in which the source text is to be translated 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 1, once the translated strings are downloaded. The next step is to select the sentence datasource either from a Github repository or the database. Finally, the project is made visible by selecting 'Public' from the visibility drop down menu.

5.3 QUALITY ASSURANCE TOOLS

The quality assurance tools included a set of protocols used at the various phases of collecting and creating of the text 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.

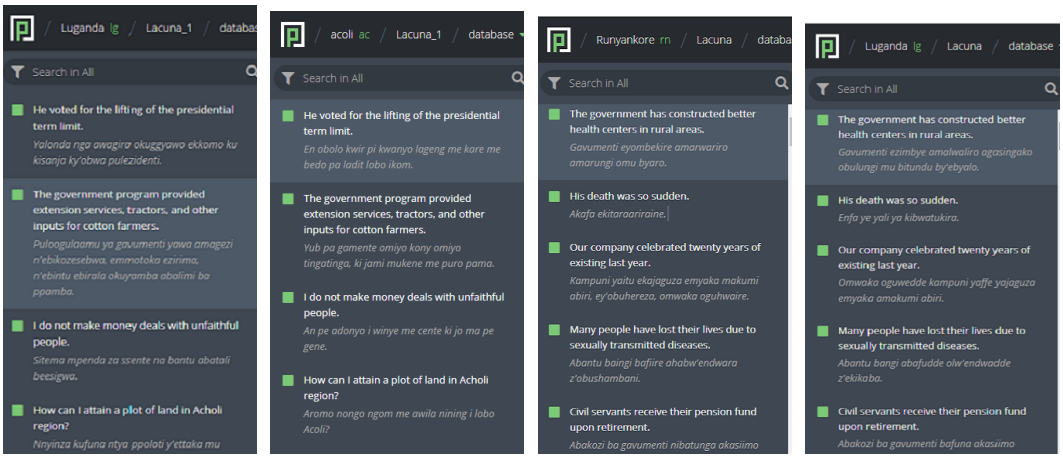

Figure 1: A figure depicting translation of the text data across the different languages

6 Ethical Considerations

There were a number of ethical issues that we had to take into consideration as we created the text and voice datasets.

6.1 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 Commonvoice dataset, we allowed source text available under a Creative Commons (CC-0) license[8]. 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.

6.2 Voice Ethical considerations

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. Participants had to agree that Mozilla may offer all of their contributions for example: text, voice recordings, validation and feedback to the public under the CC0 public domain dedication. The contributors were made aware that voice contributions were voluntary activity and there was no remuneration for the contributions.

7 Experiments and Results

In this section, we describe the baseline machine models we created for Machine Translation between English and Luganda and for the Luganda Automatic Speech Recognition model.

7.1 Machine Translation Experiments

Currently, experiments have been conducted on two languages in the parallel text corpus: the source language (English) and one local Ugandan Language (Luganda). The parallel corpus was split into a training set (80%); a test set (10%) and a validation set (10%). The Machine Translation experiments were carried out on the Hugging face platform. We trained a transformer learning model Vaswani et al. (2017) specifically the MarianMTmodel Junczys-Dowmunt et al. (2018) which is a variant

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

of the encoder-decoder model. We leveraged the use of transfer learning by training the *Helsinki-NLP/opus-mt-lg-en* and *Helsinki-NLP/opus-mt-lg-en* pre-trained models deployed on the Hugging Face platform[9]. The BiLingual Evaluation Understudy (BLEU) scores for the English-Luganda and the Luganda-English models are shown in Table 5.

Table 5: Machine Translation results.

Model	**BLEU Scores**
English - Luganda	37
Luganda - English	36.8

7.2 ASR EXPERIMENTS

An efficient transfer learning approach: a two tier pre-training approach was utilized to get a good performing model. Transfer learning was done from a pre-trained English DeepSpeech model to Kinyarwanda (Josh, 2019). The Kinyarwanda model was fine tuned to Luganda for 200 epochs. This was done because Kinyarwanda and Luganda are linguistically related. Model training was done on 300 hours of Luganda Common voice data (Rosana et al., 2019). The dataset was split with 80% training set, 10% validation set and 10% test set. On evaluating the model, we obtained baseline models of Word Error Rate (WER) of 33%.

8 CHALLENGES

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

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 and the COVID-19 pandemic that hindered physical meetings. Additionally, we also faced challenges in the language translations by the linguists we engaged on the project.

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 Luganda sentences on the Commonvoice platform - The collection of both the text and speech datasets was done in parallel and yet the speech data collection was dependent on the availability of text data. 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 Common Voice platform.

Anonymous Contributors - To build a more diverse voice dataset, it is important that we keep track of the metrics around the voice contributors for example age group and gender. However, some of the voice contributors preferred not to sign up on the Commonvoice platform and remain anonymous. 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.

[9]Hugging Face Models

Commonvoice platform challenges - The Commonvoice platform also had some challenges which included:

- 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.

Incentiving of contributors - We observed that the number of contributions increased after providing an incentive to contributors, for example data reimbursement. However, the incentives are not sustainable and there is need for continued motivation of the community to continue with voice contributions irrespective of the incentives.

9 Conclusion and Future Work

In this paper, we have provided a description of the process we undertook to create text and speech datasets for low resourced languages in East Africa. Working across five different languages avails the technicalities that are involved in building resources for low resourced languages. The data creation process is not the same across the 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 the five languages. We have also provided results from baseline models for Machine Translation and Automatic Speech Recognition models. Future work will focus on understanding 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.

Meyer Josh. *Multi-task and transfer learning in low-resource speech recognition (Doctoral dissertation, The University of Arizona).* The University of Arizona, 2019.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*, 2018.

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.

Jonathan Mukiibi, Claire Babirye, and Joyce Nakatumba-Nabende. An english-luganda parallel corpus [data set]. zenodo. https://doi.org/10.5281/zenodo.5089560, 2021.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020.

Ardila Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Lindsay Morais, Saunders, Francis M. Tyers, and Gregor Weber. "common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David I Adelani, Amelia Taylor, et al. Ai4d–african language program. *arXiv preprint arXiv:2104.02516*, 2021.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 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.