Kencorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Language Processing Tasks

Wanjawa, Barack*
University of Nairobi, Kenya
wanjawawb@gmail.com
Wanzare, Lilian
Maseno University, Kenya
Indede, Florence
Maseno University, Kenya
McOnyango, Owen
Maseno University, Kenya
Ombui, Edward
Africa Nazarene University, Kenya
Muchemi, Lawrence
University of Nairobi, Kenya

Abstract

Indigenous African languages are categorized as under-served in Artificial Intelligence and suffer poor digital inclusivity and information access. The challenge has been how to use machine learning and deep learning models without the requisite data. Kencorpus is a Kenyan Language corpus that intends to bridge the gap on how to collect, and store text and speech data that is good enough to enable data-driven solutions in applications such as machine translation, question answering and transcription in multilingual communities. Kencorpus is a corpus for three languages predominantly spoken in Kenya: Swahili, Dholuo and Luhya. This corpus intends to fill the gap of developing a dataset that can be used for Natural Language Processing and Machine Learning tasks for low-resource languages, with such languages usually being neglected due to few resources and research efforts. The Kencorpus is therefore a collection of text and speech data in the three languages. In the Kencorpus project, three Luhya dialects, namely Lumarachi, Lulogooli and Lubukusu, were sampled as Luhya has several dialects. Each of these languages and dialects therefore contributed text and speech data for the language corpus. Data collection was done by researchers who were deployed to the various data collection sources such as communities, schools and collaborating partners such as media and publishers. Kencorpus has a collection of 5,594 items, being 4,442 texts of 5.6 million words and 1,152 speech files worth 177 hours. Based on this data, other datasets were also developed as part of the project. These are a Part of Speech tagging sets for Dholuo and Luhya dialects, resulting in 50,000 and 93,000 words tagged respectively and Question-Answer pairs created from the Swahili text corpus that annotated 1,445 stories with 7,537 QA pairs. Translations of texts from Dholuo and Luhya into Swahili were done for 12,400 sentences. The datasets are useful for machine learning tasks such as text processing, annotation and translation. The project also undertook proof of concept systems in speech to text and machine learning for Question Answering task. These concepts provided results of a performance of 75% for the former, and 60% for the latter system. These are initial results that give great promise to the usability of the Kencorpus to the machine learning community. Kencorpus is the first such corpus of its kind for the low resource languages and forms a basis of learning and sharing experiences for similar works especially for low resource languages. Challenges in developing the corpus included deficiencies in the data sources, data cleaning challenges, relatively short project timelines and COVID19 pandemic that restricted movement hence the ability to get the data in a timely manner.

Keywords: Swahili, Dholuo, Luhya, POS tagging, Question Answer, Translation, Low resource languages, Corpus creation

1   Introduction

The intention to specifically focus research initiatives on low resource languages in Africa and other regions of the world is underlined by the quest to preserve these languages. Language serves as a tool for both cultural preservation and communication. Language can also be used to gauge a community's success in terms of its economic, emotional, and social development (Smith, 2019). The advancement of natural

*corresponding author

language processing in information technology, as used in machine learning and deep learning, has led to the creation of numerous useful applications such as text-to-speech and speech-to-text, machine translation, virtual assistants, text summarization, auto-correction, sentiment analysis, among others. However, these machine learning algorithms need training data, which is typically not available for languages with limited resources. Therefore, generating language datasets for languages with limited resources is an initial step to guarantee that machine learning tasks are feasible.

It is for this reason that this initiative was conducted with the aim of collecting text and speech data in low resource languages. Nonetheless, we face the challenge of dealing with the many low resource languages of the world. For example, Africa has many different languages spoken within and across borders to the tune of 2,000 different languages (Eberhard et al., 2021). A country such as Kenya alone has over 42 distinct language communities (National Museums of Kenya, n.d.). Therefore, as a start, the project has done a case study involving three Kenyan languages of Swahili, Dholuo and Luhya. The three chosen Kenyan languages of Swahili, Dholuo and Luhya in this case study are based on purposive selection on relative representativeness. Swahili is the national and official language of Kenya and Tanzania. The language is a cross border language in the Eastern part of Africa and is spoken by over 150 million speakers globally. Dholuo is a Nilotic language with an ethnic community of over 5 million speakers mainly around Lake Victoria in the three East African countries (National Museums of Kenya, n.d.; Omondi, 2020). On the other hand, Luhya is a Bantu language of about 7 million speakers also predominantly in the Western part of Kenya. It is a language with 17 sub-linguistic dialects i.e. Lulogoli, Luisukha, Luitakho, Lutiriki, Lubukusu, Lutachoni, Lunyore, Lumarachi, Lukhayo, Lusamia, Lunyala, Lumarama, Lushisa, Luwanga, Lutiriki, Lutsotso and Lukabras (Lubangah, 2018).

This research expects to develop and discover methodologies for collection, storage, and processing of corpora for under-resourced languages, while at the same time develop datasets that are based on the corpus such as Part of Speech (POS) annotation, translation across languages, Question Answering dataset and Speech to text modeling for these low-resource languages.

This data presented in this paper is available for the diverse machine learning data driven solutions such as question answering, machine translation and transcription. Through the project, we got insights into what it takes to prepare data and the accompanying part of speech (POS) annotations and translation pairs for such low resource languages. We have also developed a question answering (QA) dataset, speech and text parallel corpora and parallel translations across the low resource languages. Appreciating that such corpus and datasets open many potential opportunities in the machine learning communities, the paper reports on benchmark work done on two use cases, Question-Answering and Speech to text (STT) Transcription.

The machine learning community involved in aspects of research that uses corpora and datasets such as ours shall be immediate beneficiaries. Enterprises interested in human language technology (HLT) systems can also access datasets for developing and testing their language models as they develop practical information and communication technology (ICT) systems e.g. chatbots, searching tools, translation systems, teaching aids etc.

The rest of the paper is organized as follows – Section 2 provides the related work for this research while Section 3 provides the details of our methodology. Section 4 provides the results of the work, with Section 5 discusses these results. Finally, Section 6 provides the conclusion and points out areas of further research.

2   Related work

Though there are few corpora for low resource languages, several research efforts have gone into this initiative and more still need to be done. Some Dholuo texts have previously been collected but specifically for the task of machine translation (de Pauw et al., 2010). The Helsinki corpus of Swahili

(Hurskainen, 2004a) and Swahili language online part of speech tagging tool Swatag (aflat, 2020) are tools for Swahili language process, specifically Part of Speech (POS) tagging. The Kikuyu language of Kenya has a spell checker utility (Chege et al., 2010) and research on named entity recognition (NER) has led to the development of an NER system for ten African languages (Adelani et al., 2021).

Toolkits for neural machine translation already exist such as openNMT toolkit which uses neural networks to perform the translation (OpenNMT, n.d.). Work has been done on machine translation such as neural machine translation models used for machine translation across 5 different languages in South Africa with 50,000 sentences (Martinus et al., 2019). Practical applications of translations for low resource languages include work done in translating a glossary of COVID19 terms across 33 languages (Translators Without Borders, n.d.).

Several Question Answering (QA) datasets exist for high resource languages e.g. SQuAD (Rajpurkar et al., 2016), MCTest (Richardson et al., 2013), Common sense knowledge systems (Ostermann et al., 2018), WikiQA (Yang et al., 2015), TREC-QA (Voorhees et al., 2000) and TyDiQA (Clark et al., 2020). However, only a few datasets are available for low-resource languages, with TyDiQA being such a dataset since it has QA collection of 11 languages from Wikipedia corpus for languages including the low-resource language of Swahili. It is therefore desirable to deliberately develop more QA datasets, especially for low-resource languages.

Other machine learning methods that do not need training data, such as semantic networks (SN) can be tried on low resource language applications in the absence of training data. SNs are already used in domains such as Google Knowledge Graph (Singhal, 2012), LinkedIn (Wang et al., 2013) and Facebook (Sankar et al., 2013) amongst others. However, even such SNs would usually need some minimally processed data source such as a part of speech (POS) tagging. Once tagged, it is then possible to employ SNs to undertake tasks such as QA (Wanjawa et al., 2020, Wanjawa et al., 2021).

Part of speech (POS) tagging is therefore an important aspect of data curation that researchers of low-resource languages should consider. Some open-source toolkits exist for POS tagging high resource languages (Lamas, n.d.) but hardly any for low-resource languages. Developing such POS tagging datasets for low-resource languages is therefore also desirable. POS tagging is usually a requisite pipeline stage in many natural language processing (NLP) tasks, hence an essential dataset for language modelling. POS annotation of low-resource languages can be done using spreadsheets and online (Kituku et al., 2015; Pauw et al., 2006) after developing predefined tag sets (Tracey et al., 2019).

Data collection modalities need careful consideration and planning. Participatory methods of data collection is a workable method of sustainable data collection efforts (Nekoto et al., 2020). Some work has been done on using community initiatives to build corpora for low resource languages like Masakhane (Orife, et al 2020) and AI4D African language program (Siminyu et al, 2021), that focus on curation of language datasets. Such a participatory approach has been used in data collections for Digital Umuganda corpora in Rwanda (Digital Umuganda, n.d.) where communities gather at centralized locations for data collection activities. Open-source data can also be used for corpora compilations. Such sources include African story books (African Story Book, n.d.) and Tuvute Pamoja initiative (Tuvute Pamoja, n.d.) both of which have story collections in different African languages, while Edutab (Edutab, n.d.) provides Swahili story collections for end users. Speech data can be collected using methods such as GroupTalk where data collection is done through interviews in small groups (Cieri et al., 2002) of focused group discussions for formal settings. Conversion of written text to create speech data, such as the use of GroupMeet has been used previously to extend data corpora (Gelas et al., 2012). These tried and tested methods have been found useful and can be employed in projects such as Kencorpus.

3   Methodology

Kencorpus project collected primary data, both speech and text, in three Kenyan languages of Swahili, Dholuo and Luhya. The project then curated the data to create datasets of the raw data corpora and additional support datasets. These additional datasets were – Question Answering (QA) dataset for Swahili language, being the Kencorpus Swahili Question Answering (KenSwQuAD) dataset, set of translations of Dholuo and Luhya language texts into Swahili and part of speech (POS) tagging of Dholuo and Luhya texts. The project also developed proof of concept systems for speech to text modeling and question answering using machine learning. The datasets were geared towards machine learning for these three low resource languages. The details of each aspect of the project follows.

3.1   Choice of languages

Kenya has more than 42 indigenous languages and are grouped in three major families of Bantu, Nilotic and Cushitic (Iraki, 2009). Kenyan languages tend to have predominance in certain geographical locations within the country. The choice of Swahili as a language in the project was due to Swahili being the national language of Kenya. Despite it being spoken by many speakers in East Africa and interests globally, Swahili is still a low resource language. More deliberate research efforts are needed to provide corpora and machine processing tools for Swahili. There are many dialects of Swahili (Wald et al., 2018, Walsh, 2017). The Swahili data collected (text and speech) was mainly the Standard Swahili that is of general use in official and learning settings, though subtle differences were possible depending on the region of Kenya where the data came from. The Swahili data is therefore considered Standard Swahili in this research.

Dholuo is a Nilotic language spoken majorly in Western part of Kenya near Lake Victoria. It also has speakers in Tanzania and Uganda. It is the second most populous language in Kenya (Mazrui, 2012). Dholuo also has different dialects or sociolects, and the project collected language dialects that was available at the data collection field. Nonetheless, these dialects tend to have mutual intelligibility.

The Luhya language is a bantu language also spoken predominantly in the Western part of Kenya. This language however comprises approximately 17 different dialects within it (Lubangah, 2018), hence a language of interest due to its diversity amongst its speakers. Due to the constraints in resources, only three dialects within the Luhya language were selected, being Lumarachi, Lulogooli and Lubukusu. The purposive sampling was guided by Lubukusu and Lulogooli being the populous languages among the other dialects. Lumarachi dialect is of interest due to the geographical location of most of the speakers, bordering the Dholuo speakers in Western Kenya. The researchers also had ease of access to geographical reach and resources to interact with the Dholuo and Luhya languages.

3.2   Scope of project

The Kencorpus project aimed at collecting speech and text data for the three languages of Swahili, Dholuo and Luhya. The intention was to collect an equal number of speech and text data in these three languages.

Additionally, the collected data was to be annotated with part of speech (POS) tagging for Dholuo and Luhya texts. Swahili texts were not tagged since existing projects have already developed POS tagging for Swahili which are generally reliable (Hurskainen, 2004a, 2004b). The third aspect of the project was to undertake translation of text from Dholuo and Luhya languages into Swahili. This was to increase the Swahili corpus, while understanding the intricacies of such translations. The datasets from translation would also be useful for the machine learning task of translation. A fourth component of Kencorpus was to create a Question Answering (QA) dataset based on the Swahili texts corpus. Finally, the project was to develop proof of concept systems to confirm that the collected data and annotations were of practical use for the machine learning community. To this end, two proof of concept systems have been developed. These were a Question Answering model for

Swahili texts, and a speech to text (STT) model for Swahili speech files.

3.3 Data sources

The project identified data sources as mainly primary data, with a few secondary sources. The primary data was collected from institutions of learning (schools, colleges), local community settings and during social events. Secondary data was to come from partnering media houses and publishers. Research assistants, who are natives of the respective languages, would visit the data sources and assist in the collection of data. The project therefore employed three research assistants per language to lead data collection efforts. We had partnerships agreements with several institutions to enable us to access their existing datasets or provide links to data sources. This enabled us to collect texts from various genres such as articles, book sections, news texts, Africa short stories and other publications. The project also held story writing competitions in educational institutions with the aim of getting texts from different geographical regions and dialects. Our respondents were purposively drawn from the different genders, age groups and geographical locations.

3.4 Data collection

The project designed tools for collecting both text and voice data at their various sources. Text data was collected through story competitions mainly in schools. This meant providing the respondents with writing materials (pens, papers) and collecting the resultant creative writing. Publishers were to provide their data either as hard copies for photocopying or in some cases soft copies ready for direct processing by computers. Speech data was recorded using voice recorders (dedicated or apps on phones), while collaborating media houses were to provide speech data through computer files. All collected raw data was considered as Level 0 and was compiled by one of the researchers who would account for it and ensure it is secured and labeled.

In terms of Research Ethical consideration, we developed research consent forms for use in the project. The consent form spelt out the project objectives and how the data would be processed, accessed and used. Only respondents who were willing to provide data were allowed to participate in the project, subject to their informed consents. They also kept a copy of the consent forms. Consent forms for groups such as schools were executed by the school managers on behalf of the respondents. All members participating, whether individual or groups were listed by full names on the consent forms. Metadata collection forms, which were hard copy data forms ready for populating by pen, were also developed to capture the details of each data item collected. All these items were part of the research assistants' toolkit.

3.5 Data Cleaning

In this project we did data cleaning of the texts only. The voice clips were retained as originally recorded due to constraint in equipment to edit speech files. However, original recorded speech files passed through quality control checks to assure the best quality possible. Most of the speech files were also studio quality having been obtained from media houses.

Data cleaning of texts is an essential process of eliminating noisy signals that would otherwise degrade the quality of data sets intended for natural language processing. The noise in data can be due to the presence of corrupted characters, misspellings, inconsistent data, redundant data, missing characters, extra spaces, inconsistent punctuations, spelling variations, and codeswitching, among others. The data collected in this study was not devoid of these noisy elements. Therefore, a systematic approach was developed to ensure that the raw corpus comprising handwritten manuscripts, scanned pages, images, web-scrapped data, and text files were adequately processed to deliver quality and clean language datasets. The data cleaning approach adapted a four-stage process based on den Broeck et al. (2005) data cleaning framework, however, tweaked it by introducing a preceding stage that we named 'Digitize'. Therefore, the data cleaning five stages included the digitization, screening, diagnosis, treatment, and documentation stages. The data cleaning process flow is shown on Fig. 1.

The first step was to digitize the manuscripts by scanning and converting the scanned PDF images into an editable text format. Approximately three thousand manuscripts, largely composed of compositions written by primary students in grades five to eight, were scanned into PDF text images. Subsequently, several Optical Character Recognition (OCR) software were used to convert the PDF files to editable text formats, with mixed results: some very accurate digitization and sometimes inaccurate. This could largely be attributed to the handwriting quality and partly to the precision of the OCR software, as shown in Fig. 2. These included both mobile and web OCR software like the Pen-to-Print app, Expert PDF OCR, and Google doc internal file converter. The latter had the most plausible results.

Therefore, Google doc internal file converter was extensively employed to convert the rest of the scanned files into editable text documents bearing similar file names and stored under a designated Google drive, as depicted in Fig. 3. Besides, this eliminated further project costs of procuring proprietary OCR software and storage spaces for the cleaned files.

Some of the original manuscripts were improperly scanned, and a hazy image was obtained which turned incomprehensible when digitized. To resolve this, the respective raw manuscripts were availed and rescanned correctly. In some cases, the manuscripts contained crossed or strikethrough words that were incorrectly interpreted as ideographic characters by some of the OCR software. These were dropped to clean the final sentence. The tweets were cleaned automatically using Python programming and its associated libraries like the Natural Language Toolkit, regular expressions, and other libraries. The regular expression library was used to remove non-ASCII characters, duplicate tweets, punctuations, and other noisy characters. Moreover, all the tweets

were lowercase, and the URL section, images, and other sections of the tweet dropped to maintain only the message section.

The second step was the screening, which entailed analysis of the quantity and quality of the text output from step one and other raw digitized text like the scrapped tweets. Besides, screening involved utilizing Python's Natural language tool kit and other libraries to quantitatively and visually analyze and understand the datasets by observing patterns in data types, word occurrences, and other features. The third step was the diagnosis whereby incomplete data, inconsistent punctuations, data redundancies, spelling variations, and other errors were identified.

The fourth step involved treating the dataset by eliminating the errors identified in step three. This was the longest, most time-consuming, and most expensive step. Here, eighteen human data cleaners comprising indigenous language speakers from the three languages were sought, interviewed, hired, and trained on data cleaning and research ethics. Moreover, performance expectations were set including cleaning at least 100 file documents per week. To ensure data reliability, each file had a primary and secondary data cleaner. First, the primary data cleaner meticulously went through each sentence in the documents to clean it according to the training scheme that included how to identify and remove noisy elements, correct spelling and grammar errors, handling missing data, among others. Secondly, at the end of the week, the data cleaners were swapped within their respective language teams to look through the allocated files for any previously unseen errors and correct those to improve the data quality. Finally, a linguist, for the respective languages, randomly selected and perused sample files for quality assurance of the data cleaning.
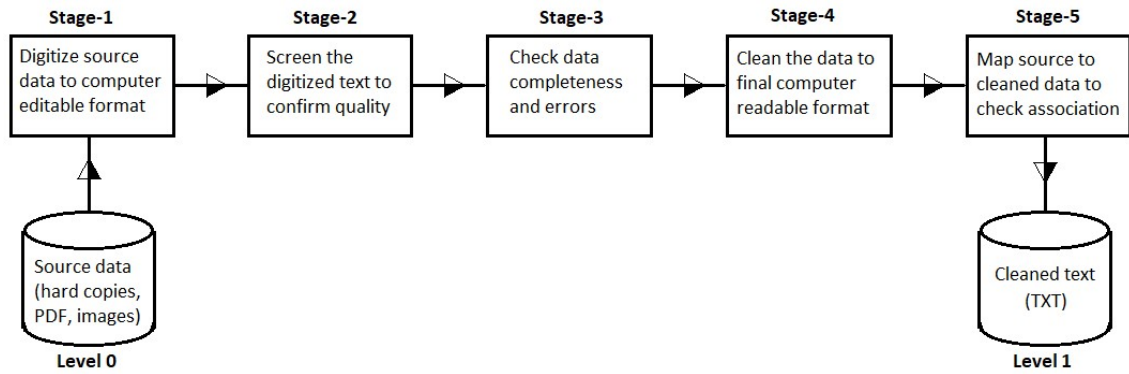
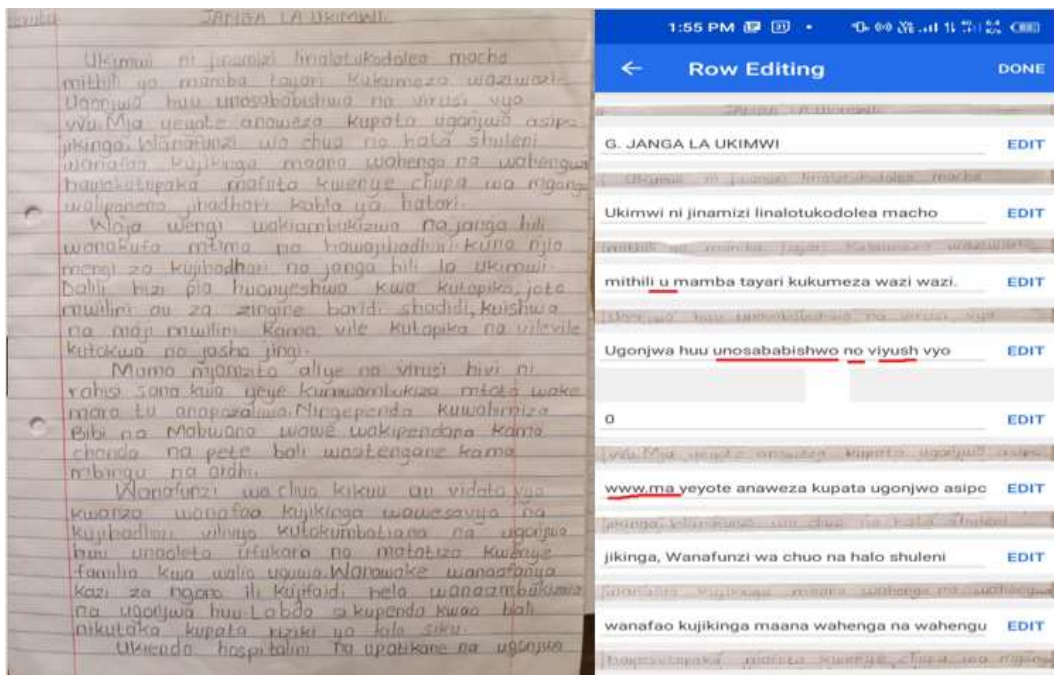*Figure 1: The 5-stage data cleaning process adopted by Kencorpus project (source: author)*

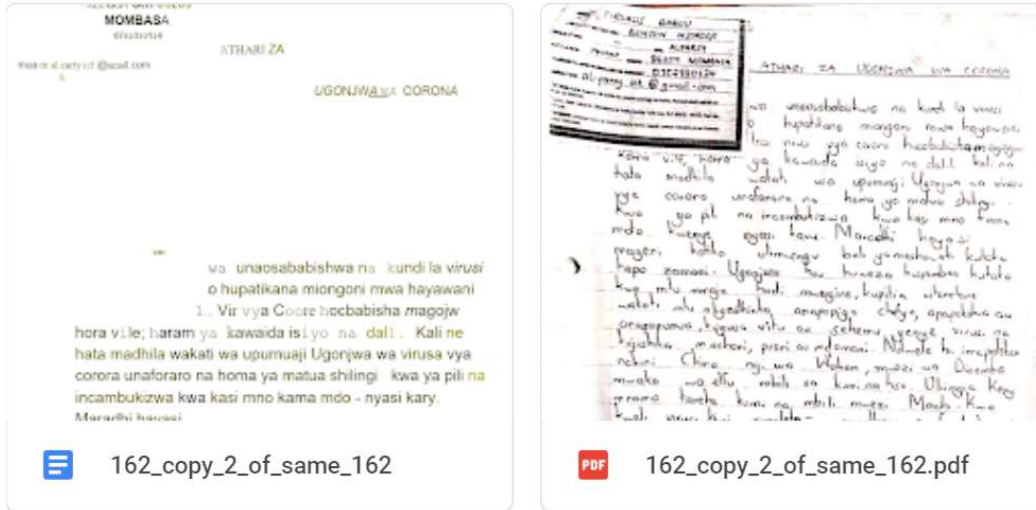*Figure 2: Digitization using Pen to print app from the play store*

Figure 3: Digitization of scanned PDF to an editable text document

The fifth step was documentation that involved indexing all the original scanned files and mapping the 'level 0' master folder with the equivalent cleaned files stored in a 'level 1' folder, under the respective language subfolders.

3.6   Staff Training and Piloting

The Kencorpus project undertook training of the identified research assistants, who were knowledgeable in the three research languages. This training was done at the beginning of the project. Both face to face and virtual training was conducted. Internal capacity building between researchers within the language categories was exercised. Data collection efforts were expected to be evenly distributed amongst the three researchers per language.

The researchers then started their field work while reporting back on field experiences. Such experiences enabled the project to tune their research tools to be better suited to the research environment and continually improve them over time.

Subsequent training was done for data cleaning and annotation tasks. Data cleaning was aimed at converting the raw collected data as Level 0, into edited computer formats e.g. text images or scans were to be converted to the standard text (TXT)

format. Any other test formatted documents were also to be reformatted back to TXT. Speech files were all to be converted to the standard Waveform Audio File Format (WAV) format.

3.7   Data annotation

We developed our own annotation guides for POS tagging, translation and QA tasks. These were made available to all research assistants for their reference. Part of speech (POS) tagging tasks were done by at least two research assistants per text document. The researcher in charge did random checking on the POS tags and confirmed agreement. The researchers, being the subject matter expert in the respective languages, had the final decision in case of disagreement. POS tagging was based on a predefined tag set developed by the subject matter experts and shared amongst the annotators. This tag set was made uniform for deployment across the two languages being annotated. The starting point of tag set development was the 12 universal tag set (Petrov et al., 2011). The tag set was only incremented upon consensus and agreement amongst the linguists in the four languages i.e. Dholuo and the three dialects of Luhya, to ensure uniform deployment of tag sets across the corpus languages.

Collection of Question-Answers pairs was done using an online form, where annotators filled in the Questions and Answers (QAs) for each story read. The methodology for QA was similar to what was employed for SQuAD (Rajpurkar et al., 2016) and TyDiQA (Clark et al., 2020). We set a number of QA pairs based on a story, with each annotator reading and setting the QA pairs as per a predefined criteria of number of questions, type of questions and that they be single answer and answerable.

The unit of translation was a sentence. The annotators were to annotate one sentence at a time. The translation followed literal translation of meaning in context. We adopted equivalence translation where cultural words, idioms, metaphors and sayings were translated into available equivalences in the target language. At the beginning, the research had group translations where we conducted comparative translation in order to agree on the guidelines. Native speakers and linguistic experts were involved in the translation. As supplementary material, dictionaries were used as references.

3.8 Quality control check

The project setup a system of checking each aspect of the project and ensuring that quality was assured. Quality control checking of our processes from collection to annotation was a continuous process and this assisted in ensuring that we setup datasets that are of high quality for use in machine learning tasks. Monitoring data items as they were being collected and transmitted enabled the project to check on the data as it came and made corrective actions in terms of quality of data itself or process of getting it to the centralized storage. Data collection was submitted and managed in a shared centralized storage under the accountability of one researcher. The research interrogated each received data item and communicated back to the researchers in cases where the data needed to be resubmitted for whatever reason.

Data collection was supervised by the linguists, while the researchers also visited the data collection sites to confirm and spot check the data collection efforts. Change of process included asking the research assistants to send data immediately upon collection instead of keeping it for long. Each member of the research team was assigned a task and worked with a team of research assistants to accomplish the task.

Data cleaning teams were guided on tools and methods of cleaning, with the supervising researcher checking on the outputs. Each annotation work was under a supervising researcher to ensure that the expected task was done, confirmed, and checked.

POS tagging was rechecked by the subject matter experts on a random sampling basis targeting 10% of all the annotated words. This was done by sampling the various genres of texts, but fully checking all the tags in the sampled text under consideration. For translation, the researcher in charge, being the subject matter expert, also sampled 10% of the translated work and checked the translations to confirm accuracy. QA tagging task also involved the annotators reviewing a sample of each other's work to confirm agreement. The researcher in charge was to have the final decision in case of arising issues on the set of QA pairs. The two proof of concept systems were also overseen by the project team to check and confirm that the project expectations were met.

4   Results

The results of the project is the Kenyan languages corpus (Kencorpus) of texts and speech, together with other datasets of translations to Swahili, part of speech tags of Dholuo and Luhya languages and a question answering dataset for Swahili language.

4.1   Kecorpus statistics

The details of the dataset that the Kencorpus collected is shown on Table 4.1 below. Note that the Total numbers (Texts, Speech and Total) refer to the unique filenames compiled for the collection. Words refers to the total number of words in the collection as directly counted for documents already in computer text format or estimated for raw data that was based on images or scans. The total time is the number of hours, minutes and seconds in files in that collection.

Table 4.1: Kencorpus project statistics

Language	Texts	Words	Speech	Time	Total
Swahili	2,585	1,829,727	104	19:10:57	2689
Dholuo	546	1,346,481	512	99:03:08	1058
Luhya	987	2,272,957	536	58:15:41	1523
Tweets (Swahili)	324	152,750	0	0:00:00	324
Total	4,442	5,601,915	1,152	176:29:46	5,594

The final outcome of the initiative is the Kenyan Languages corpus (Kencorpus) that has a dataset of 4,442 texts and 1,152 speech files. The Luhya language data was obtained from three different dialects, each with its own data distribution as shown broken down in Table. 4.2 below.

Table 4.2: Kencorpus project breakdown of Luhya language data into dialects

Dialect	Texts	Words	Speech	Time	Total
Luhya_ Marachi	483	67,812	138	15:37:46	621
Luhya_ Bukusu	135	876,257	354	30:11:00	489
Luhya_ Logooli	369	1,328,888	44	12:26:55	413
Total	987	2,272,957	536	58:15:41	1,523

4.2   Kecorpus annotations

The datasets created from the initial data collections above are described as shown in Table 4.3, 4.4 and 4.5.

Table 4.3: Kencorpus datasets created from the data collected - Translation

Task	Sentences
Dholuo-Swahili translation	1,500
Luhya-Swahili translation	11,900
Total	12,400

The final dataset has 12,400 translated Dholuo-Swahili and Luhya-Swahili sentence pairs (Table. 4.3), 143,000 POS tags in two languages (Dholuo and Luhya-Marachi, -Bukusu, -Logooli) as in Table 4.4, and 7,537 QA pairs from Swahili texts

(Table 4.5). More details about the Kencorpus Swahili Question Answering dataset (KenSwQuAD) including details of quality checks and machine learning systems developed as a result of that set is available as a separate project[1].

Table 4.4: Kencorpus datasets created from the data collected – POS tagging

Task	Words
Dholuo POS tagging	50,000
Marachi POS tagging	27,900
Bukusu POS tagging	30,900
Logooli POS tagging	34,300
Total	143,000

Table 4.5: Kencorpus datasets created from the data collected – QA annotation

Aspect	No.
Swahili text stories annotated	1,445
Total QA pairs	7,537

All the above-mentioned datasets are available for use by researchers and any other user under creative commons with attribution (CC BY 4.0) international license. Kencorpus data collections is available on the project website[2].

4.3   Kencorpus proof of concept

Kencorpus project developed two proof of concept systems to test the datasets for practical use in machine learning tasks. One proof of concept model was to test question answering (QA) based on the QA dataset from the project using deep learning. The deep learning system tried was transformers (BERT) using a dataset of 100 stories with 500 QA pairs. The test set was exposed to 80% of the dataset, while 20% was used for testing. The other machine learning method tried was that of using semantic networks as already tried in other datasets (Wanjawa et al., 2021). This method creates a network of nodes and edges based on the part of speech (POS) tagging and the inter-relatedness of the network can be queried.

---

[1] www.kencorpus.co.ke/kenswquad

[2] www.kencorpus.co.ke/corpora

The second proof of concept model was to develop a speech to text (STT) based on the Swahili speech files collected in the project. This Speech to text (STT) system was based on a collection of speech corpus that is 27hrs 31min 50 sec, with 7 male and 19 female speakers. This was based on a project that developed a Kiswahili phoneme dictionary of 31,759 words-phoneme pairs. The method used was Python programming language STT toolkits.

QA systems are applicable in systems such as internet search, dialog systems and chatbots, while STT are useful for processing text in instances such as educational materials for those with hearing challenges.

Two proof of concept systems developed for the project performed as shown on Table 4.6. Details of these systems are available on our project website (www.kencorpus.co.ke/transcriptions and www.kencorpus.co.ke/kenswquad) and other research papers (Wanjawa et al., 2022).

Table 4.6: Kencorpus proof of concept system performance

Test system	Accuracy
QA system using deep learning (BERT)	60%
QA system using semantic network	80%
STT using Python programming utilities	70%

5   Discussions

This project developed the Kenyan Languages corpus (Kencorpus) that has datasets in three languages of Swahili, Dholuo and Luhya. However, the Luhya language used in this research sampled three dialects, hence the final corpus has data on five distinct languages of Swahili, Dholuo, Luhya-Marachi, Luhya-Logooli and Luhya-Bukusu. These datasets were possible through data collection from both primary and secondary data sources by the project researchers.

Analysis of the data collected indicated that text files tend to be in the range of 100-300 words, while our projects were 2,000 words per text file. That turned out to be 10-times more that we projected in any typical text file. The analyzed data on speech files also confirmed that the file contents were much less than 5 minutes each, especially from storytelling scenarios. The longer speech files were those from media houses who could afford to provide audio clips that could run even upto an hour.

Getting data from primary sources such as education institutions and the community provided the corpus with rich data from such settings that reflected diversity in the age groups from lower primary to college levels. Gender and cultural diversity was also manifested at data collection points.

However, we also encountered some challenges in coming up with the Kencorpus datasets. These were:

Incentive to respondents - Challenges in getting primary data included cases where respondents expected or asked for compensation for data collection, which had not been planned, nor would such a modality fit into the budgetary considerations of the project.

COVID-19 restrictions - We had challenges of collecting and processing the data during the COVID-19 pandemic with curfews and movement restrictions.

Poor quality data - Data collected from lower grade schools were also a bit illegible due to their writing habits and were made worse by their use of pencils for writing. Scanning or retyping handwritten texts to create the computer formats needed in the corpus was difficult, with discovery of such challenges coming up much further in the project after the data collection phase has ended. A lesson learnt on this was the use of darker pens should be the norm when collecting handwritten stories. While it was fast and relatively easy to get secondary data from our collaborators, we still had challenges in processing some of the data that was still clear e.g. scans or photographed images of newspaper clippings. Sometimes the shortcomings on the quality of the scans were related to the equipment used in the scanning or the expertise of the operator. We had instances where the scanned images would cut off the edges of the raw documents. This made the reproduction of the original document difficult. Some secondary data was also quite voluminous

e.g. book scans or Bible sections. These were challenging to scan and eventually reprocess into a computer format. We continually gave feedback to our researchers on the quality of data being transmitted so that they could improve over time.

Speech files did not have many challenges apart from the quality of recording that depended largely on the equipment. However, most of the collections (over 60%) were provided by project collaborators from media houses which were already in high studio recording quality and in a suitable compression format such as MP3. Speech files tended to be big in size and sometimes the recording gadgets such as smartphones would run short of storage memory. Careful planning of work to anticipate the expected data volumes and preparing the recording hardware settings in advance assisted us in addressing such challenges. We also encouraged the research assistants to post their data to our central storage as soon as they could immediately after the data collection exercise.

Data cleaning - Data cleaning process was used to convert our raw documents into computer processable formats, which were TXT for all texts and WAV for all speech files. This process would lead to the creation of the corpus that is ready for machine learning tasks. Our project planned for the data cleaning to be done after data collection so that the data is fully available for cleaning when the exercise started. We however soon realized that the volume of data from the raw source that needed cleaning was much, despite the available cleaning time being limited. We also had other project tasks such as translations, POS tagging and QA annotation which needed the cleaned data and could therefore not start at their planned time as they needed the cleaned data. We overcame this by allowing parallel running of both the cleaning and the annotations. In cases such as QA annotation, we allowed the annotators to use the raw data (images) and set the QA pairs from them. Future projects dealing with corpus creation can benefit by starting the data cleaning immediately when data collection starts so that the two are parallel running and the workload in data cleaning is balanced through the project cycle.

The consideration as to whether cleaning of texts should also include correction of spelling and grammar remains an issue of concern. Text data from lower-level schools tended to have many of such mistakes. However, correcting them, especially grammar, would shift the source data from the original author's style to a new one that may not reflect the true status of the author. In our case we corrected the spelling only, and left the grammar as was provided by the original contributors. The collected data could also provide insights into the influence of aspects such as age, gender and geographical location on the topical issues of discussion or concern and is an interesting research area to explore. The original collected text data before cleaning is also a good data source for research on data cleaning, optical character recognition (OCR) and image processing. Though the corpus publishes the cleaned TXT data format, it is possible to obtain the original images on request for purposes of research.

Low volumes of data - Challenges experienced with the proof-of-concept systems were the low volume of data in the corpus for machine learning methods that needed lots of data. The performance of the proof system on QA was low due to inadequate data. Even the QA system confirmed that improvements were only possible when the data volume increased. The speech to text (STT) system did not have much data to train on since the Swahili speech files were relatively few, compared to the texts. More work is going on to revamp the data collection.

Lack of training and testing data remains the challenge with low resource languages when subjected to machine learning methods that need such datasets. Our Kencorpus project therefore contributes to resources of such languages as Swahili, Dholuo and Luhya to start placing them at the realm of machine learning systems so that users can access the many technological benefits of machine processing e.g. education materials for the physically challenged, internet search, chatbots, frequently asked questions (FAQs) that are developed from models such as STT and QA. These models need a corpus and the

accompanying annotations, both of which the Kencorpus project contributes to.

As this and other data collection and annotation efforts continue, we expect that machine learning researchers targeting Swahili and other low resource languages, many of which are in Africa shall start building resources for the benefit of the users.

6   Conclusion

In this paper, we have described the creation of Kencorpus which is the first corpus of 5,534 data items, both speech and text, for Swahili, Dholuo and Luhya. The creation of the datasets were achieved by collecting primary and secondary data from education institutions, community, media houses and publishers. The paper also reports on data cleaning efforts to ensure that the datasets are in the expected computer format that aids in further machine learning processes. The resultant collection in Kencorpus is the set of 4,442 text documents of about 5.6 million words and 1,152 voice files of about 177 hours across the three languages.

We reported POS tagging efforts for Dholuo, Luhya-Marachi, Luhya-Logooli and Luhya-Bukusu. These tags are useful for language processing systems such as spelling and grammar checkers, hence tools for low resource languages can now be developed in our word processing programs. The paper explained the development of parallel corpora for Dholuo-Swahili and Luhya-Swahili. This is useful in increasing parallel corpora for machine translation systems. A QA dataset of 7,537 QA pairs was also reported. This is useful in machine learning systems for machine comprehension tasks and enquiry systems such as chatbots, frequently asked questions (FAQs) and even internet search using low resource languages. The developed transcription corpus for Swahili will go a long way in developing speech technology.

By monitoring the data collection and dataset creation process at every stage, we ensured that the resulting corpus and datasets are of high quality and are also of practical use in typical machine learning systems as demonstrated by our proof-of-concept systems in speech to text (STT) and question answering using deep learning and semantic networks.

For future work, this research can be updated with new data and datasets over time to further enrich it and make it even more useful to machine learning models that require lots of training data.

References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., & others, 2021. MasakhaNER: Named Entity Recognition for African Languages. ArXiv Preprint ArXiv:2103.11811. https://arxiv.org/pdf/2103.11811

Aflat, 2020. Kiswahili Part-of-Speech Tagger - Demo AfLaT.org. https://www.aflat.org/swatag

African Story Book, n.d. Retrieved June 28, 2021, from https://www.africanstorybook.org/

Chege, K., Wagacha, P., de Pauw, G., Muchemi, L., Ng'ang'a, W., Ngure, K., & Mutiga, J., 2010. Developing an Open source spell checker for Gĩkũyũ. In Proceedings of the Second Workshop on African Language Technology - AfLaT 2010 (Issue Lrec).

Cieri, C., Miller, D., & Walker, K., 2002. Research methodologies, observations and outcomes in (conversational) speech data collection. Proc. HLT 2002.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J., 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. ArXiv Preprint ArXiv:2003.05002.

de Pauw, G., Maajabu, N., & Wagacha, P. W., 2010. A knowledge-light approach to Luo machine translation and part-of-speech tagging. Proceedings of the Second Workshop on African Language Technology (AfLaT 2010). Valletta, Malta: European Language Resources Association (ELRA), 15–20.

Den Broeck, J. V., Cunningham, S. A., Eeckels, R., & Herbst, K., 2005. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities, PLOS Medicine; journals.plos.org. https://journals.plos.org/plosmedicine/article?id =10.1371/journal.pmed.0020267

Digital Umuganda, n.d. Retrieved March 31, 2022, from https://digitalumuganda.com

Eberhard David M., G. F. S., & Fennig, C. D. (Eds.), 2021. Ethnologue: Languages of the World (Twenty-fourth). SIL International.

Edutab, n.d. Retrieved July 18, 2021, from https://edutab.africa/

Gelas, H., Besacier, L., & Pellegrino, F., 2012. Developments of Swahili resources for an automatic speech recognition system. Spoken Language Technologies for Under-Resourced Languages.

Hurskainen, A., 2004a. Helsinki corpus of Swahili. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hurskainen, A., 2004b. Tagset of SWATWOL A two-level morphological dictionary of Kiswahili Clause boundary tag Tags for marking noun class features. http://www.aakkl.helsinki.fi/cameel/corpus/swa tags.pdf

Iraki, F. K., 2009. Language and political economy: A historical perspective on Kenya. Journal of Language, Technology & Entrepreneurship in Africa, 1(2), 229–243.

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.

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.

Kituku, B., Musumba, G., & Wagacha, P., 2015. Kamba Part of Speech Tagger Using Memory-Based Approach. International Journal on Natural Language Computing, 4(2), 43–53.

Lamas, n.d. Retrieved March 2, 2021, from https://lamas.science.ru.nl/software/

Lubangah, L. J., 2018. Linguistic Versus Geographical Boundaries: A Lexical Semantic Assessment Of Luhyia Dialects. http://erepository.uonbi.ac.ke/bitstream/handle

/11295/104707/Lubangah%20_Linguistic%20V ersus%20Geographical%20Boundaries%20A%2 0Lexical%20Semantic%20Assessment%20Of%2 0Luhya%20Dialects..pdf?sequence=1

Martinus, L., & Abbott, J. Z., 2019. A focus on neural machine translation for african languages. ArXiv Preprint ArXiv:1906.05685.

Mazrui, A., 2012. Language and education in Kenya: Between the colonial legacy and the new constitutional order. In Language Policies in Education (pp. 151–167). Routledge.

National Museums of Kenya, n.d. The Languages of Kenya: The Nilotic, Bantu and Cushitic Language Groups. Retrieved March 31, 2022, from https://artsandculture.google.com/story/the-language-of-kenya-the-nilotic-bantu-and-cushitic-language-groups/2AJCwOhG6x7aIQ

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., & others, 2020. Participatory research for low-resourced machine translation: A case study in african languages. ArXiv Preprint ArXiv:2010.02353.

Omondi, D., 2020. Five tribes retain hold at apex of population as numbers increase. The Standard. https://www.standardmedia.co.ke/kenya/article/2001361377/2019-census-results-tyranny-of-big-tribes

OpenNMT, n.d. Retrieved March 31, 2022, from https://opennmt.net/

Ostermann, S., Roth, M., Modi, A., Thater, S., & Pinkal, M., 2018. Semeval-2018 task 11: Machine comprehension using commonsense knowledge. Proceedings of the 12th International Workshop on Semantic Evaluation, 747–757.

Pauw, G. de, Schryver, G.-M. de, & Wagacha, P. W., 2006. Data-driven part-of-speech tagging of Kiswahili. International Conference on Text, Speech and Dialogue, 197–204.

Petrov, S., Das, D., & McDonald, R., 2011. A universal part-of-speech tagset. ArXiv Preprint ArXiv:1104.2086.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P., 2016. Squad: 100,000+ questions for machine comprehension of text. ArXiv Preprint ArXiv:1606.05250.

Richardson, M., Burges, C. J. C., & Renshaw, E., 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 193–203.

Sankar, S., Lassen, S., & Curtiss, M., 2013. Under the Hood : Building out the infrastructure for Graph Search. http://www.facebook.com/notes/facebook-engineering/under-the-hood-building-out-the-infrastructure-for-graph-search/10151347573598920/

Singhal, A., 2012. Introducing the Knowledge Graph: things, not strings - Inside Search (Vol. 2013, Issue 7/22/2013). http://insidesearch.blogspot.com/2012/05/intro ducing-knowledge-graph-things-not.html

Smith, B., 2019. Preserving cultural heritage one language at a time. https://blogs.microsoft.com/latino/2019/12/12 /preserving-cultural-heritage-one-language-at-a-time/

Tracey, J., Strassel, S., Bies, A., Song, Z., Arrigo, M., Griffitt, K., Delgado, D., Graff, D., Kulick, S., Mott, J., & others, 2019. Corpus building for low resource languages in the DARPA LORELEI program. Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, 48–55.

Translators Without Borders, n.d. Retrieved March 31, 2022, from https://translatorswithoutborders.org/twb-creates-covid-19-glossary/

Tuvute Pamoja, n.d. Retrieved February 5, 2021, from https://digitalumuganda.com

Voorhees, E. M., & Tice, D. M., 2000. Implementing a question answering evaluation. Proceedings of LREC'2000 Workshop on Using Evaluation within HLT Programs: Results and Trends.

Wald, B., & Gibson, H., 2018. Swahili and the Bantu languages. In The World's Major Languages (pp. 903-924). Routledge.

Walsh, M., 2017. The Swahili language and its early history. In The Swahili World (pp. 121-130). Routledge.

Wang, R., Conrad, C., & Shah, S., 2013. Using Set Cover to Optimize a Large-Scale Low Latency Distributed Graph. Proceedings of the 5th USENIX Workshop on Hot Topics in Cloud Computing. https://www.usenix.org/conference/hotcloud13/workshop-program/presentations/Wang

Wanjawa, B., & Muchemi, L., 2020. Using Semantic Networks for Question Answering-Case of Low-Resource Languages Such as Swahili. International Conference on Applied Human Factors and Ergonomics, 278–285.

Wanjawa, B., & Muchemi, L., 2021. Model for Semantic Network Generation from Low Resource Languages as Applied to Question Answering–Case of Swahili. 2021 IST-Africa Conference (IST-Africa), 1–8.

Wanjawa, B., Wanzare, L., Indede, F., McOnyango, O., Muchemi, L., & Ombui, E., 2022. KenSwQuAD--A Question Answering Dataset for Swahili Low Resource Language. arXiv preprint arXiv:2205.02364.

Yang, Y., Yih, W. T., & Meek, C., 2015. WikiQA: A challenge dataset for open-domain question answering. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2013–2018.