KenSwQuAD – A Question Answering Dataset for Swahili Low Resource Language

Wanjawa, Barack[12]; Wanzare, Lilian[3]; Indede, Florence[3]; McOnyango, Owen[3]; Muchemi, Lawrence[2]; Ombui, Edward[4];

1 – corresponding author, 2 – affiliation University of Nairobi, Kenya, 3 – affiliation Maseno University, Kenya. 4 – affiliation Africa Nazere University, Kenya

Abstract

This research developed a Kencorpus Swahili Question Answering Dataset KenSwQuAD from raw data of Swahili language, which is a low resource language predominantly spoken in Eastern African and also has speakers in other parts of the world. Question Answering datasets are important for machine comprehension of natural language processing tasks such as internet search and dialog systems. However, before such machine learning systems can perform these tasks, they need training data such as the gold standard Question Answering (QA) set that is developed in this research. The research engaged annotators to formulate question answer pairs from Swahili texts that had been collected by the Kencorpus project, a Kenyan languages corpus that collected data from three Kenyan languages. The total Swahili data collection had 2,585 texts, out of which we annotated 1,445 story texts with at least 5 QA pairs each, resulting into a final dataset of 7,526 QA pairs. A quality assurance set of 12.5% of the annotated texts was subjected to re-evaluation by different annotators who confirmed that the QA pairs were all correctly annotated. A proof of concept on applying the set to machine learning on the question answering task confirmed that the dataset can be used for such practical tasks. The research therefore developed KenSwQuAD, a question-answer dataset for Swahili that is useful to the natural language processing community who need training and gold standard sets for their machine learning applications. The research also contributed to the resourcing of the Swahili language which is important for communication around the globe. Updating this set and providing similar sets for other low resource languages is an important research area that is worthy of further research.

Keywords

Swahili, Question Answer, low resource languages

1. Introduction

The quest for a question-answer (QA) dataset for natural language (NL) processing tasks continue to draw research interests globally. QA datasets are an important component in machine learning as it is one of the ways of data query that humans do. That means that machines can as well be programmed or given access to data to learn from and then undertake the same task for the benefit of users. QA is commonly used in many information querying tasks such as internet search, frequently asked questions and dialog systems.

Use of machines to process data and provide results has enabled fast information processing for the benefit of users, such as using computers for internet search. Machine processing of natural language is however not trivial. The NL has to be transformed into a format that computers understand, which could be by word embeddings e.g. one-hot encoding, term

frequency inverse document frequency (TF-ID), dense vectors such as Word2vec or GloVe (Pennington et al., 2014) or deep learning methods such as transformers e.g. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). Many of these NL transformations require data for training the models and perform quite well when the training data is available in abundance e.g. BERT performance in NL tasks is well documented (Libovický et al., 2019). The data source for such machine processing tasks can be provided through a corpus of the languages being processed or through gold standard datasets for tasks such as QA.

While high resource languages such as English, French, Chinese, Spanish etc. have vast amounts of data for training or even gold standard sets for QA, the low resource languages such as Swahili and many other African and world languages do not have such resources. This has been due to low research interests that would otherwise deliberately collect these datasets. The problem is made worse since those who need to do research on NL tasks are likely to have no option but to use the existing data resources that exist in high resource languages. This leads to further decline in low language resources while high language resources continue to get research interests.

There is therefore a need to deliberately target low resource languages by contributing NL datasets to benefit users and researchers. The task of QA remains important for information processing by computers. The only way of ensuring that a system is capable of QA tasks is by having an existing gold standard QA set, which can be used to verify the performance of the computer system. Deliberate research is therefore needed to get such gold standard QA sets, out of which computer processing systems that undertake QA related tasks can be tested upon.

Unfortunately, gold standard QA datasets are few or non-existent for low-resource languages. That means that research or testing of systems such as internet search and dialog systems for low resource languages is difficult. This is because QA datasets to test such models are not readily available. Contrast this to the high resource language of English where many gold standard QA datasets exits. These include SQuAD (Rajpurkar et al., 2016), MCTest (Richardson et al., 2013), WikiQA (Yang et al., 2015), TREC-QA (Voorhees & Tice, 2000) and TyDiQA (Clark et al., 2020). Very few public domain datasets have Swahili language texts, such as TyDiQA. TyDiQA is collection of QA sets in 11 languages from Wikipedia corpus of the different languages.

It is due to this problem of inadequacy of gold standard datasets for the low resource language of Swahili that this research developed Kencorpus Swahili Question Answering Dataset (KenSwQuAD), a Swahili language Question Answering Dataset. We developed KenSwQuAD by annotating primary data that was collected by the Kencorpus project. Kencorpus is a corpus of three Kenyan languages created through funding by the Lacuna fund of the Meridian Institute USA. The purpose of the Lacuna fund is to facilitate collection of data of low resource languages, both text and voice, for purposes of documenting and resourcing such languages to benefit the research community.

One of the languages in the Kencorpus dataset was Swahili, out of which KenSwQuAD was developed. Annotators formulated question and answer pairs based on over 65% of the

collected texts of Swahili language. A data quality check was then done on a sample of 12.5% of the annotated data for assurance that KenSwQuAD was a correct and reliable dataset for QA tasks of the Swahili language. A proof of concept on KenSwQuAD using semantic network modeling confirmed that NL processing was possible based on this QA dataset.

2. Background info of research topic

While datasets, both corpora and gold standard sets, remain important for natural language (NL) processing tasks as done by computers and on the internet, these datasets need to be developed through research for them to be realized and made available. The cycle of exploring data corpora then testing it on practical language models has led to many research efforts being concentrated on high resource languages such as English, French, Chinese, Spanish, which have readily available data corpora and language models. Data that is collected and not tested with some models may not be proved to be of practical use for computer processing tasks. This lack of resources can discourage research in low resource languages that may not be explored due to lack of tools to test them by their very nature of being low resourced. Deliberate research efforts for high resource languages have led to QA datasets such as MCTest (Richardson et al., 2013), SQuAD (Rajpurkar et al., 2016) and TyDiQA (Clark et al., 2020). Such efforts need to continue for all languages, including the low resource ones.

It is for this reason that the quest to collect a gold standard question answer dataset for the low resource languages of Africa become important. Africa has many different languages spoken within and across borders to the tune of 2,000 different languages (Eberhard David M. & Fennig, 2021). One of these languages is Swahili. First of all, Swahili, despite being a low resource language, is an important language of communication in Eastern Africa. It is the national language of both Kenya and Tanzania. It is spoken in many other countries of the world. Wikipedia estimates the number of Swahili users as between 100M and 150M worldwide (wikipedia, 2020), while Omniglot puts this number as 140M (omniglot, 2020). Swahili is therefore worthy of more research for the benefit of the users and the enthusiasts, both in terms of datasets and gold standard sets.

The objective of this research is therefore to develop a gold standard QA dataset for Swahili. This shall provide additional language resources to the low resource language of Swahili. The dataset is also useful for natural language processing tasks and language modeling to address issues such as internet search, dialogue systems and any machine learning system that requires question-answer type of data. The dataset shall be freely available in the public domain for research and exploration. Since the QA dataset is derived from full stories of Swahili language, the stories themselves shall also be available for users to read and enjoy, apart from doing any other NL processing task that relies on raw text.

The rest of the paper is arranged as follows – section 3 provides the related work for this research while section 4 provided the details of our methodology. Section 5 provides the results of the work, with section 6 discussing these results. Finally, section 7 provides the conclusion and points out to areas of further research.

3. Related work

There are many different question answering datasets available for exploration and research. These include MCTest (Richardson et al., 2013) for language comprehension questions and SQuAD (Rajpurkar et al., 2016) a dataset of 100,000 questions based on Wikipedia. Both these sets are intended for machine learning systems, hence rely on the availability of lots of data for the machine learning model to be successful. Lots of data resources are found for high resource languages where deliberate research has enabled these datasets to be collected over time.

High resource languages have benefitted from many data sources and hence machine learning models have been developed to exploit high resource language data. For example, Wikipedia (wikipedia, n.d.) has been used as the data source of machine learning based QA systems due to the vast data available in that site. Even SQuAD is a collection of QAs from Wikipedia, and models based on SQuAD use Wikipedia for their training when they are using that QA set.

Low resource languages however have tended to be neglected in research interests and application of machine learning models (Berment, 2004; Besacier et al., 2014). To this end, question answer datasets such as TyDiQA (Clark et al., 2020) have been developed to try and capture QA sets in more languages than just the high resource ones. TyDiQA is a collection of QA sets in 11 languages, both high resource and low resource languages. It is one of the few sets with the Swahili low resource language as part of the collection. It is also based on Wikipedia articles and the QA pairs are crowdsourced from web users. It exploits this vast data source and hence provides data that can be used for machine learning. TyDiQA dataset for Swahili has issues such as incorrect responses to some questions due to the crowdsourcing nature of the system as per analysis done by some researchers (Wanjawa & Muchemi, 2021). It however remains among the few sets that deliberately setups up QA pairs for some low resource languages and makes it available in the public domain.

Datasets that target low resource languages have been few. However, machine learning models that need data for training cannot work where data is little or unavailable. That means that models tend to be tested and continually get improved for high resource languages, such as English. This has progressively led to continued neglect of research efforts in resource low resource languages (Hirschberg & Manning, 2015). Some efforts have nonetheless been made to uplift the resources available for low resourced languages. These include the Helsinki corpus of Swahili (Hurskainen, 2004) and Swahili language online part of speech tagging tool Swatag (aflat, 2020). The Kenyan Kikuyu language has a spellchecker (Chege et al., 2010), while a named entity recognition set for ten languages of Africa has been developed (Adelani et al., 2021). This is a good start and much more research is still needed to resource such low resource languages.

The starting point of resourcing low resource languages still remains the provision of more datasets, tools and gold standard datasets such as QA sets. Testing of such data for low resource languages may however not be done using machine learning methods that need data for training, hence other methods such as language modeling using semantic networks can be explored (Novák et al., 2009; Wanjawa & Muchemi, 2020). Such modeling does not need training data but still presents the data in such a way that machines can read,

understand and then perform practical tasks such as question answering. SNs are already being used in domains such as Google Knowledge Graph (Singhal, 2012), LinkedIn (Wang et al., 2013)and Facebook (Sankar et al., 2013) amongst others.

Low resource languages therefore need to be given research focus by providing data sources and data sets (Selamat & Akosu, 2016). Additionally, we already have data modelling methods such as semantic network representation that do not need training data to represent such datasets in ways that make them of immediate benefit to users e.g. for internet search or even question answering. However, in time the low resource languages shall be resourced and even their processing may as well be done using the better performing statistical methods that need training data as proved with high resource languages. Before then, deliberate effort of developing the datasets should continue. This realization is what informs the need for a gold standard QA dataset for Swahili as done in this research.

4. Methodology

Kencorpus Swahili Question Answer Dataset, KenSwQuAD, was formulated using a method comparable to that used for SQuAD (Rajpurkar et al., 2016) and TyDiQA (Clark et al., 2020) but was tweaked to suit the available data source. The dataset used for generating KenSwQuAD was the Swahili portion of the data collected by the Kencorpus project (Wanjawa et al., 2022). Kencorpus project collected primary data, both text and voice, in three low resource languages of Swahili, Dholuo and Luhya. The first language listed, Swahili, is spoken throughout East Africa and in other parts of the world, while the last two languages are predominantly spoken in western part of Kenya and parts of Uganda and Tanzania that neighbour these populations where most speakers come from.

4.1 Data selection

The Swahili dataset from Kencorpus comprised of 2,233 unique texts and 104 unique voice files. KenSwQuAD shortlisted texts for annotation using purposive sampling. This was the chosen method because this was the first time such a dataset was being developed and the researchers needed to develop a predefined criteria for the choice of stories that meets the annotation objective. The project time and funding were also limited and the research had to get the balance right in the data selection.

The shortlisted stories were those that fitted the following criteria – at least 100 words in length but not more than 2,000 words. This was done to eliminate very short or very long stories that may be difficult to annotate or to follow. The other criteria was to target only prose and short stories. The research had realized that annotating texts such as plays, dialogue or poems would be difficult for the QA task since our research was based on the methodology on what was done for SQuAD (Rajpurkar et al., 2016) and TyDiQA (Clark et al., 2020). Our selection therefore meant that items such as Tweets, Facebook posts, long stories, religious texts, text on comics, texts of mixed languages and songs were excluded from the shortlist. The summary of the selection criteria is shown on Table 4.1 below.

Table 4.1 – Summary of texts sampled from the Kencorpus

Consideration	Total No.
Total Swahili texts in corpus	2,585
Texts over 2000 words	-42
Texts under 100 words	-325
Texts excluded for any other reasons	-50
Total Texts shortlisted for annotation	2,168
Total text provided to annotators	1,660
Proportion provided for annotation	76.6%

The final set of shortlisted texts based on the exclusion criteria, were therefore 2,168 texts, of which 1,660 (76.6%) were provided to the QA annotators. The method of data allocation to annotators was by equal number of stories in a given time duration (monthly, at the start of the month), then allowing individual annotators to access the next set of a fixed number of stories upon finalizing their targets. These subsequent sets were allocated weekly and replenished weekly upon confirmed completion.

4.2 Annotation guide

The research developed an annotation guide that spelt out the expectations of the annotation project, including issues such as the inclusion and exclusion criteria. This was to assist the researchers to pick the right stories from the vast Swahili dataset of Kencorpus project datasets. A final shortlist was availed to the annotators out of this dataset. After the selection of stories that met the annotation criteria, the research developed the criteria for the number and types of question and also of the type of answers to annotate. An analysis of the questions set on the TyDiQA (Clark et al., 2020) dataset informed the type of questions to formulate for KenSwQuAD.

The research decided to set a standard number of five questions per story. The questions were to be the object enquiry type (what, which, who, when). The research also decided to include at least one question that involved some reasoning (why, how). Most QA datasets usually revolve around such enquiry, including SQuAD (though it is in English language) and TyDiQA (the Swahili portion). These guidelines were also included in the annotation guide.

The research also provided guidance to the annotators on the number of questions to set for each question type, the desirable number of words in the question and also the answer, and on whether unanswerable questions should be allowed. These issues are summarized on Table 4.2 below, as was also included in the annotation guide. Some of the provisions on the guide were just suggestions e.g. number of words or questions to be set per type. The final decision depended on the text of the story, though the annotators were asked to try their best to follow the recommendations.

Table 4.2 – KenSwQuAD criteria for setting QAs

Aspect	Recommendation
Number of Type 1 questions (who, what, which)	3
Number of Type 2 questions (when)	1
Number of Type 3 questions (how, why)	1
Number of questions per story text	5
Number of words in the question	10 max
Number of answers per question	1
Number of words in the answer (Type 1 and 2)	3 max
Number of words in the answer (Type 3)	10 max
Multiple choice answers permitted	No
Unanswerable questions permitted	No

4.3 Annotator selection

The research recruited annotators then conducted both in person and online training for the annotators. The recruitment of annotators targeted speakers of Swahili language and additionally those who were currently engaged in teaching of the Swahili language in Kenyan educational institutions. This requirement was set to give the project the best personnel who could understand the intricacies of question and answering formulation based on their experiences with the Swahili language in their daily careers. QA sets also tend to mimic the real-world information retrieval needs, hence the annotators would leverage on their experience on what learners would usually query on such story corpus, apart from what they would also examine as teachers in testing language comprehension. Nonetheless, the research restricted the type and complexity of questions and excluded issues such as questions on language structure e.g. questions on parts of speech, functions of words etc.

The in-person training was done over a two-day workshop where the annotation guide was discussed and practical demonstrations done, followed by presentations, discussions and consensus building. The review and evaluation of the workshop confirmed that the annotators understood the expectations of the project. Online training was done for both those who had attended the physical workshop and also to the new members who had met the recruitment criteria and were joining the annotation team. The new members were assigned to respective mentors who had attended the in-person workshop. A practical training on annotation was once again done on the online meeting and discussions held on dos and don'ts.

4.4 Annotation tool

The researchers developed an online annotation tool on Google forms for collecting the QA pairs. The annotators were trained on the use of the online form and did a practical on it during the online sessions. The annotation team of six was then given twenty sample story texts each, and asked to test the annotation tool and give feedback over a period of one week. The tool was tested while the researchers monitored the data that was trickling in on the backend collation spreadsheet. The annotators confirmed that they were conversant and ready to use the tool based on their one-week test period. The submitted test annotations

were carefully reviewed by the researchers to confirm that they were fit for the project as per the annotation guide, training and discussions.

4.5 Data size and Distribution to annotators

We reset the data collection database after the test period and provided the annotation team with a new set of texts for actual annotation. Each annotator had different and unique stories to annotate still based on purposive sampling. The sampling dealt with the following considerations – first, the collection of 1,660 text stories was split into the format of their raw data e.g. those already typed texts in computer format (TXT, DOC, RDF), PDF, JPG and PNG. The stories not already in computer typed formats (such as images/PDF), were further categorized into handwritten versus those from typed sources.

The Kencorpus metadata had already provided information on the exact or approximate number of words on each story, hence each of the categories of stories was then sorted by the number of words. Each of the six annotators was then allocated the story texts in a category, one name at a time, then the allocation was repeated in batches of six until all the texts in the category were exhausted. The allocation for the six would then be done in the next category as per the list of stories sorted by number of words. This sampling ensured that each annotator got exposed to all types of texts (as per the different text types and formats) and also got comparable lengths of stories in an equitable manner.

We also spelt out the output expected per week, both in terms of minimums and maximums, just to ensure that the annotators gave the project the expected concentration that was needed. Too much work done over a short period of time had the danger of the annotators rushing through their work and hence not giving each story the time of reading and thoughtfulness that was needed to ensure that they formulated the questions and answers that were of the expected standards.

These measures were monitored weekly over the two-month annotation period. In anticipation for future machine learning purposes and to further verify the exact locations in the text where the answer is found, we annotated some stories with paragraph numbers. We therefore deliberately provided a 13.1% set of text stories (218 stories) to the annotators to include the paragraph number from where they got their answers from. The distribution of data to annotators is shown on the Table 4.3 below.

Table 4.3 – Description of Data provided to annotators of KenSwQuAD

Aspect	No.	Comment
Total no. of texts in Kencorpus Swahili	2,585	All texts in corpus
Total no. of texts Shortlisted for KenSwQuAD set	2,168	83.9% of text corpus
Total no. of texts Provided to the annotators	1,660	76.6% of shortlisted
Total no. of texts to annotate without indicating paragraph	1,442	86.9% of provided data
Total no. of texts to annotate and indicating paragraph no.	218	13.1% of provided data

4.6 Quality control check on annotated data

We finally did a quality control check on the annotation work. This was done over a one-week period at the end of the annotation project, where we sampled 12.5% of texts already annotated (180 texts, 900 questions) and switched them through to different annotators, ensuring that none of the annotators got their own work.

The switched over dataset consisted of the unique story identification numbers (story_ID) and only the questions that had been set by the initial annotator. We deleted the original answer and left that answer column blank. This annotation set was provided in the form of a spreadsheet. We then provided each annotator with the relevant stories and the spreadsheet containing only the questions. Their task was then to derive only the answers after reading the story text and the associated questions. The 180 texts sampled had both the QA types where paragraph number had been indicated (13%), while the balance 87% did not have paragraph numbers indicated. This sampling ensured that both types of QA annotation types were selected in the same proportion as their numbers in the original set.

4.7 Proof of concept testing of the QA dataset

As a proof of concept, the research used the semantic network (SN) method to create some networks of the final QA dataset stories in order to check if such an SN machine learning method was capable of undertaking QA tasks. The methodology used for setting up the model was that already done for Swahili QA in a previous research (Wanjawa & Muchemi, 2021). This method works in cases where there is no data for training a model. In this method, the Swahili text is first subjected to part of speech tagging (POST) which can be done using online tools (aflat, 2020), then an SN connecting subject-predicate-objects (SPO) of the text is created using tools and programming code. The created network is then queried using a query language - SPARQL Protocol and RDF Query Language (SPARQL). Deep learning methods such as transformer (BERT) were not tested since such would require a training corpus which is difficult to get for the low resource language of Swahili, despite BERT performing quite well when it is provided with vast training data as confirmed with English language experiments (Libovický et al., 2019).

5. Results

The results of the Swahili language QA annotation project is the Kencorpus Swahili Question Answer Dataset, KenSwQuAD. This started from the initial selection of 2,168 unique stories that were shortlisted for the project, with 1,660 of these stories being provided to the six annotators.

5.1 KenSwQuAD corpus statistics

The QA annotation responses from annotators was recorded on an online form. At the close of the project, a total of 1,547 annotated stories had been received. However, only 1,370 story texts had been uniquely numbered as expected, to correspond to their respective story_IDs as provided to QA annotators. The balance of 177 story texts listed in the data collection tool had repeated story_IDs. The analysis of the 177 annotations with repeated story_IDs is shown on Table 5.1 below. The column 'Set' shows the total number of QA

annotations in that category, while the column 'Final' shows the total number of unique story_IDs that were reconfirmed after reverification of the QA annotations.

Table 5.1 – Analysis of Repeated Stories collected during annotation

Aspect	Set	Final
Total no. of repeats of story_IDs (exact duplication) – pick one only	26	12
Total no. of repeats of story_IDs (different QAs set for the same story_ID) – combine the collected QAs to be more than 5 sets per story_ID	129	63
Total no. of annotations with repeats of story_IDs (different QA different QA contexts) – exclude from collection since story_IDs are erroneous	22	0
Total	**177**	**75**

From the results shown on Table 5.1, the reconciliation of annotations from repeated story_ID numberings resulted into 75 additional uniquely identified story_IDs with the correct story_IDs that matched the set of 5 QA pairs. The 75 unique stories where therefore added to the initial collection of 1,370 unique story_IDs to give a total of 1,445 as the final set of unique stories in the dataset, each annotated with 5 QA pairs. The collection of 1,445 stories included 201 story texts where annotators had indicated paragraph numbers where the answer was derived. The final set of 1,445 unique stories with QA pairs were therefore derived from both the set of those having paragraph numbers indicated (201 stories) and those that did not indicate the paragraph number (1,244). The derivation of the final KenSwQuAD dataset is shown on Table 5.2 below.

Table 5.2 – Final Results of annotating KenSwQuAD

Aspect	No.	Comment
Total no. of texts in Kencorpus Swahili	2,585	All texts in corpus
Total no. of texts Shortlisted for KenSwQuAD set	2,168	83.9% of text corpus
Total no. of texts Provided to the annotators	1,660	76.6% of shortlisted
Total no. of annotations collected on the online form	1,547	Not all were unique
Total no. of texts with unique story_IDs for QA set	**1,445**	**87.0% of provided stories**
Total no. QA pairs set on the unique story_IDs	**7,526**	At least 5QAs per story
Total no. of texts that were not annotated	215	12.9% of provided

KenSwQuAD dataset therefore consists of **1,445 unique stories** each annotated with a minimum of 5 QA pairs, giving a total corpus of **7,526 QA pairs**.

The results of the quality control check done on the sampled data from the collected QA data is shown on Table 5.3 below, where only some of the QA annotated texts were cross-checked by different annotators to confirm that the answers provided by the initial annotators were the same as that provided by a second annotator. This set had both the QAs with paragraph numbers (13%) and those without (87%). 53.9% of the answers provided to QA pairs by second annotators were exactly the same word-by-word responses as those provided by the first annotator. However, some answers were not the exact words used by the original annotators though the context or reasoning was the same e.g. 'Mama yake' and 'mamake' (his/her mother) or 'Miaka 25' and '25' ('25' years as a number, the other being '25' but also

qualified by the words 'years old'). These were deemed to be the same since they would naturally be answered in those two versions. Other cases where the original and second annotator were deemed to have agreement is when there was an obvious typo during the data entry by either of the two annotators. We had 13 stories where each of the 5 QAs in the texts were answered by second annotator exactly as was answered by the first annotator (485 answers). The analysis is shown on Table 5.3 below.

Table 5.3 – Results of quality control check on some KenSwQuAD stories

Aspect	No.	Proportion
Total no. of texts finally annotated for KenSwQuAD	1,445	100.0%
Total no. of texts Sampled for quality control check	180	12.5%
Total no. of questions in the check dataset	900	100.0%
Total no. of answers that were as exact word-for-word as initial annotation	485	53.9%
Total no. of answers that were as similar to initial annotation	341	37.9%
Total no. of answers that were as deemed comparable	34	3.8%
Total no. of answers that were not as per initial annotation	40	4.4%

The question that had the least agreement between annotators was Q5, that required some reasoning. The general thought of the annotators was the same, but the words used in the answer set were quite different. We considered the answers to be similar in cases where the same reasoning (not necessarily the words) was used in the how/why type of questions e.g. 'name one of the characters' would be correct for any qualifying answer, even if the annotators provided different answers.

As indicated on Table 5.3, there was no agreement between the original and the checking annotators for some 40 QA pairs. We analyzed the results of this set of 40 cases as per Table 5.4 below. The final correct or wrong verdict was made by the third reviewer, being the researcher.

Table 5.4 – Analysis of cases of disagreement between annotators for KenSwQuAD answers

Aspect	No.	Proportion
Total no. of answers with disagreement	40	100.0%
Total no. of answers where initial annotator was correct	9	22.5%
Total no. of answers where initial annotator was wrong	1	2.5%
Total no. of answers that needed reconciliation of annotators 1 and 2 to final confirmation that both were correct	30	75.0%

This one wrong case by the initial annotator that was considered incorrect was however not out of context. The initial annotator indicated the answer as 'tree', while the second annotator indicated the answer as 'fruit'. The context of the question would however strictly restrict the answer to 'fruit'.

5.2 KenSwQuAD proof of concept

For proof of concept on a machine learning task, some stories from the QA dataset, such as story_ID 3830 (five paragraphs with 354 words) was converted to a semantic network (SN). The partial reproduction (paragraph 1 only) of the story text in Swahili is shown on Table 5.5 below, with a translation to English also provided for purposes of illustration.

Table 5.5 – Original and Translation of story_ID 3830 from KenSwQuAD dataset

Original	*Kilimo katika nchi yetu ya Kenya ni muhimu na kinafaa kuzingatiwa kwa manufaa yake mengi. Moja ni ufugaji wa mifugo ambao hutupa protini kupitia kwa nyama. Hii protini ndiyo inayoupa nguvu mwili na kuujenga. Mifugo hawa kama kuku hutupa mayai ambayo yaweza yakauzwa na kuimarisha maisha ya mfugaji na familia yake kwa kumpa pesa za kukidhi mahitaji yake.*
Translation	Agriculture is important for Kenya hence needs attention due to the many benefits. First benefit is the protein obtained from meat. This protein provides energy to our body and builds it up. Animals such as chicken provides us with eggs which can be sold to benefit the keepers and their families by providing them with cash that is used for their wellbeing

A partial semantic network created from story_ID 3830, part of which presented in Table 5.5, is shown in Fig. 5.1 below. The figure is the visual representation of the resource description framework (RDF) formatted file as produced using online tools (*RDF Grapher*, 2021). The interrelationships between the various subjects and objects can be visualized as contained on the RDF triples data store.

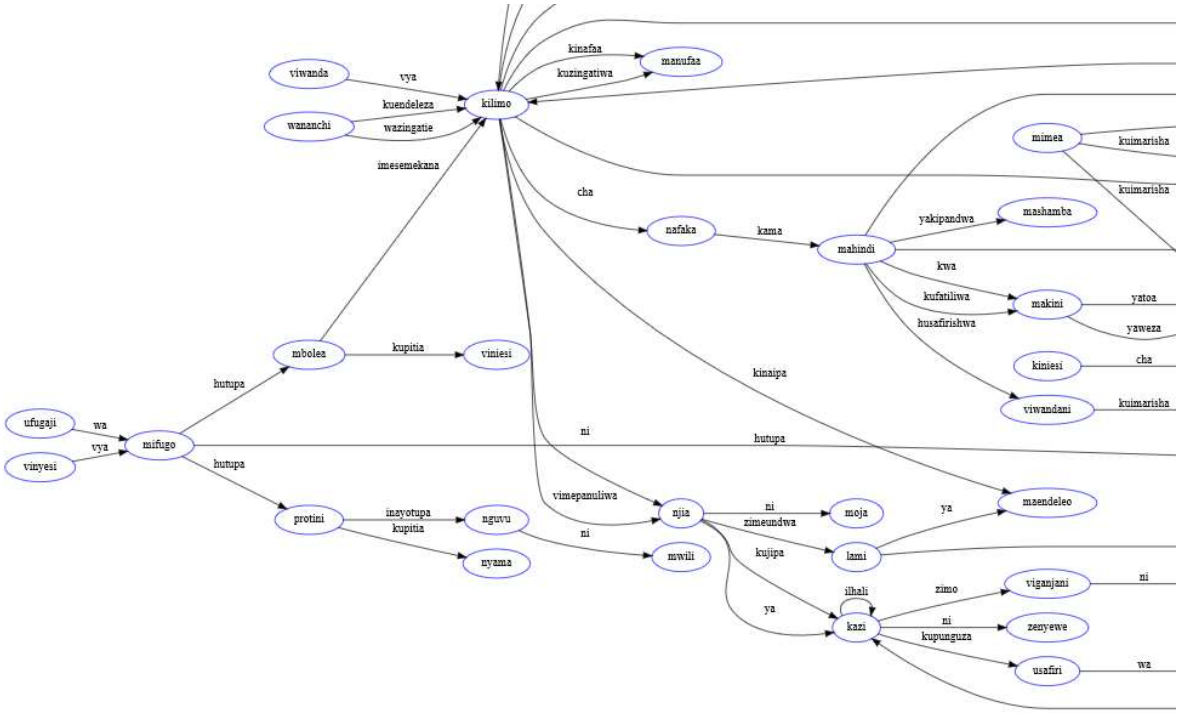

Fig. 5.1 – Visualization of RDF created from Kencorpus story_ID 3830 (source: author)

The 5 questions annotated under KenSwQuAD and subjected to the network are shown on Table 5.6, with a translation to English for purposes of illustration:

Table 5.6 – QA set for KenSwQuAD story_ID 3830

	Question	**Answer**
1	*Kilimo ni muhimu katika nchi gani* (In which country is agriculture important)	Kenya (Kenya)
2	*Mifugo hutupa nini* (What do animals provide)	*Mbolea* (fertilizer)
3	*Mahindi huuzwa wapi* (Where is maize sold)	*Ng'ambo na nchini* (Both within the country and abroad)
4	*Ni aslimia gani ya vifaa vinavyouzwa ng'ambo* (What percentage of good are sold out abroad)	80 (80%)
5	*Vipi maisha huweza kuimarishwa kupitia kwa kuku* (How do chicken improve the wellbeing of keepers)	*Kwa kuuza mayai* (By the sale of eggs)

Carefully formatted query language (SPARQL) queries subjected to the network to enquire on the subjects and objects could easily provide responses to questions 1, 2 and 4 based on the network. The SN answered Q3 as 'ng'ambo' (abroad) and 'Kenya', since that is what is explicitly stated in the story. The response of the annotator is however 'abroad' and 'within the country'. Of course, by implication, 'within the country' would mean 'Kenya', but that is not explicit from the annotator's response. This is an issue that differentiates the explicitness that computer algorithms require, versus the generalization/assumptions of human reasoning.

The object enquiry questions (1,2,3,4) were therefore all answered correctly based on the SN, with only the reasoning question being unanswerable (Q5/A5). Formulating a SPAQRL query to address this type of question is difficult as a start, and the nearest that could be done was to ask for a relationship between 'chicken' and 'livelihood'. This relationship gave a different result 'eggs', and not 'sale of eggs', which is the monetary value derived from this transaction, as indicated and required by the annotator. This was always the intent of Q5 of all the annotated QA pairs – a reasoning question that should not be directly picked from the text but needed some reasoning, despite the facts on the text.

5.3 Final KenSwQuAD corpus

The results of this project is the Kencorpus Swahili Question Answer Dataset, **KenSwQuAD**. This is dataset of **1,445 stories**, each annotated with at least 5 QA sets, hence a total of **7,526 QA pairs**. The stories are provided in a corpus while the QA set is provided on a comma separated values (CSV) format. Both the dataset of texts and QA set in CSV format have been released to the public domain on the project website below:

www.kencorpus.co.ke/kenswquad

6. Discussions

The results show that it is possible to develop a question answering dataset for a low resource language such as Swahili, the product of which is Kencorpus Swahili Question Answer Dataset, KenSwQuAD. This is possible when the data of the language is collected then an annotation guide is developed to guide the QA annotation process. Checking on the work being done by annotators through continued monitored ensures that the expectations of the annotation are achieved. Those who use and understand the language should be ideally selected to do the annotation for purposes of getting the best gold standard set. However, it is essential to check on the quality of the work done, as was done by sampling the annotated work and asking for different annotators to reconfirm the answers.

The KenSwQuAD set managed to annotate 7,526 QA pairs from 1,445 story texts by annotating a standard number of 5 questions per story. This set was developed from 55.9% of the whole collection of 2,585 Swahili story texts in the Kencorpus data collection. However, the sampling criteria that we adopted meant that only 2,168 texts of the 2,585 were shortlisted as being eligible for annotation, with 76.6% (1,660 stories) from the shortlisted texts being availed to the QA annotators. The annotators therefore managed to set QA pairs for 87.0% of the texts (1,445 stories) provided to them for annotation. In terms of the annotation work itself, the annotators were able to work on most of the story texts that they were allocated. The story texts that were skipped were those that had issues such as being illegible or those that were not making sense based on what was written by the source author. Some annotators were however not able to finish the batch of stories provided to them by the end of project. However, all annotators managed to achieve at least 80% of their targets. This shows that careful selection of annotators is essential to achieve QA project objectives such as that of KenSwQuAD.

The quality control set of 12.5% of the texts in the final annotated corpus confirmed that the initial annotators had done the correct annotations, since only 1 of the 900 tested QA cases was incorrect. The extent of the incorrect response was however not completely out of context. However, getting exact answer words from both annotators was only possible in 54% of the cases, mostly for object or one-word responses. The rest of the answers had a slight variation e.g. in spelling, word inflections or formation, use of synonyms or use of phrases that mean the same thing. The KenSwQuAD QA set is therefore reliable and is confirmed to contain questions that would elicit the expected responses regardless of the annotator. Ambiguity in language is however an issue to contend with when dealing setting QA pairs.

There are many different types and range of QA systems, and one QA annotation may not cover all aspects of QA. It is for that reason that the project developed an annotation guide that spelt out what could or could not be done. For example, our QA annotation did not cover aspects such as the unanswerable questions or the cloze type of QAs. This is even true for most public domain QA systems which usually pick a particular type of QA setup and stick with that type only. This research followed suit by annotating the machine reading comprehension (MRC) type of Swahili QA.

The KenSwQuAD set is therefore a good starting point for the different natural language tasks that require question-answer sets. These include QA systems, internet search and dialogue

systems. The data can be a starting point for such systems. With continued data collection and development of similar annotations as KenSwQuAD, machine learning of Swahili and other low resource languages shall start being more prevalent. This shall lead to more interests in such languages, which is likely to lead to their resourcing e.g. by more datasets and more model testing.

A proof of concept on the applicability of KenSwQuAD was done by trying the data on a machine learning system of semantic network (SN) generation from annotated text of the corpus. The method tried was that used in other related research (Wanjawa & Muchemi, 2021) that had been proven to work in cases where there is no data to train a model, such as was tested using the Swahili part of the TyDiQA dataset (Clark et al., 2020). The results confirm that the natural language text can be formatted as an SN and that the SN subjected to the gold standard QA pairs of KenSwQuAD can provide the expected answers. That means that research that needs QA datasets can now benefit from KenSwQuAD set because it can work in such instances. This research demonstrated that the use of SN can format the data in such a way that computer systems can understand this low resource language of Swahili and answer some questions correctly.

This QA dataset is of benefit to computer science and machine learning community who want to build models that understand language and then undertake useful tasks such as internet search, dialog systems and chat bots. Such systems need the corpus to be restructured in a way that machines can understand the text, be it by word embeddings or deep learning. Other methods of language modeling such as semantic networks (SN) can also structure the data for machines to understand and hence process and provide useful outputs. Apart from language modeling, KenSwQuAD can also benefit language communities who can access such a corpus that has texts of varying themes and sizes for their enjoyment and learning. The associated QA set can be used to test understanding of the language for reading enthusiasts who are learning the Swahili language.

This QA dataset of 1,445 Swahili text stories and its related collection of 7,526 QA pairs is the first of its kind for the low resource language of Swahili that employed our methodology. We annotated data from a corpus of narratives collected from Kenyan institutions of learning or local information sources such as current affairs news stories. Our total number of questions is however lower than that of other high resource languages such as SQuAD (Rajpurkar et al., 2016), but this can be expected, based on our limited corpus that targeted only the stories collected from the Kencorpus project. We believe that annotating 56% of the whole of Kencorpus Swahili text corpus (87% of texts availed to annotators) was a good starting point for this gold standard set that is a first of its kind.

Challenges noted during annotation included the raw text being illegible. This was because the annotation was being done on the original primary data as collected from the field by the Kencorpus project (Wanjawa et al., 2022). The field data comprised of scanned documents such as PDF or image documents taken by phones as JPG or PNG. Some of these texts were unclear either due to low resolution processing or inadequacy of the equipment that captured the images. Some handwritten texts were difficult to read, even when scanned properly. We advised the annotators to skip out any raw text that were unclear or illegible. We missed out a number of stories through skipping. These were about five skipped stories per annotator

over the annotation project, though we had also done a purposive selection of stories to only get texts that were as clear as possible. The initial intention was to do QA annotation on texts already retyped to computer format, but these retyped texts were not yet ready by the time of annotation and we were working on a tight timeframe.

Preprocessing the images to text (TXT) format first then doing annotation thereafter would have been the preferred approach if this opportunity presented itself. This would have dealt with the quality of the raw data, but probably not the issues of grammar or illegible handwriting by the original authors. Nonetheless, the corpus was a true reflection of the stories from the field with the author's grammatical constructs being maintained. This could still be a good thing for research on natural language construct by authors and probably even the influence of language construct on machine learning. Of course, editing the original stories to provide better grammar is possible, though this would not be true to original. This is also a possible research direction of creating a revised corpus which can then be used as a comparator with the original copy for linguistic and machine learning tasks. Typing out texts would also likely lead to introduction of errors during the retyping, though quality control checks can be introduced to address this. Such issues present a research opportunity for future work. The Kencorpus project has already presented these retyped texts.

The other challenges were noted when dealing with the annotation process and tools. Some annotators indicated text data on numerical fields e.g. where we asked for paragraph numbers, the annotators would write the full text such as 'paragraph 3' instead of just '3'. This was easy to resolve at data cleaning stage. We also set the annotation data collection form to have all data fields as compulsory, ensuring that at least we got data on each of the required fields. The issue of repeated story_IDs featured in 177 annotation cases. We however reconciled this set and come up with 75 unique stories from this set. Some stories were repeated but different questions had been set regardless. This meant that we increased the number of QA pairs collected in some cases to be more than the standard 5 QA pairs per story.

These repetitions likely occurred due to the methods of keeping track of work progress that the annotators employed. Some did not have an effective system to remind them of what they had already annotated, since all stories were posted on one location on their individual repositories. This mainly occurred during the initial weeks of the project. The errors became fewer with time as we continued to sensitize annotators on better methods of tracking work done, include moving such finalized works to different folders within their collections. However, we also lost 22 annotated stories (1.3% of annotations) whose story_IDs were erroneous, and it was not possible to determine the correct story_IDs. It would take lots of time and effort to determine the relationship between the QAs in this set and the correct story_ID from the collection of over 1,600 uniquely numbered texts. This was a small loss of potential QA pairs that could have gone into our dataset.

Our annotation was also based on raw data as collected from the field. Some texts had been collected from both lower and higher-level school students. Some texts had grammatical or logical errors. These rendered some stories difficult to follow, hence difficult to annotate with QA pairs that made sense of the story. We advised the annotators to skip any of such stories in case they felt that setting 5 QA pairs were not possible from such stories. We also had

already resolved not to correct any errors on the raw text that was collected by the Kencorpus project. We were to remain true to source due to copyright and consent restrictions that the project had adopted.

Budget and time constraint meant that we could not undertake full annotation of all the texts on the Swahili dataset from Kencorpus. It would have been desirable to annotate as much of the available text as possible, including the shorter or even longer stories, so as to benefit the researchers who are keen at testing machine learning models that need such data aspects. KenSwQuAD however is a good start and more work can still be done either on the unannotated part of the Kencorpus Swahili data, or any other Swahili or low-resource language data collections elsewhere using our methodology.

7. Conclusion

This research developed Kencorpus Swahili Question Answer Dataset, KenSwQuAD, a Swahili language question answer (QA) dataset of 7,526 QA pairs from a corpus of 1,445 stories, having annotated at least 5 QA pairs for each story. The dataset is the first of a kind that is created specifically for the low resource language of Swahili, which is a major language of communication in Eastern Africa and is also spoken in many other countries such as USA, Australia and Europe. This dataset is meant to spur interest in low resource languages for enthusiasts who may want to learn and test their understanding, as well as the members of the machine learning community, who may need a gold standard dataset to test their models in machine reading comprehension (MRC). The results of such modeling are practical end user systems such as internet search, dialog systems and chatbots, which can be done directly in the low resource languages such as Swahili, instead of relying on systems that are tailored for high resource languages such as English. Users prefer to use their usual languages of communication and are usually just forced by circumstances to switch to other languages due to lack of tools of usage in their usual languages (Orife et al., 2020).

The KenSwQuAD dataset was developed by human annotators who got their primary data from the larger Kencorpus project (Wanjawa et al., 2022), the Kenyan languages corpus that collected data in three low resource languages of Africa, namely Swahili, Dholuo and Luhya. The Swahili data collected was both speech and text, with KenSwQuAD being created from 56% of the Swahili text stories in that project. This means that our QA dataset annotated more than half the collected texts of Swahili in the corpus. The annotators used a guide that spelt out all aspects to guide their work. A quality check mechanism was employed to confirm that the annotated questions were matching with the provided answers. Additionally, we built a proof-of-concept machine learning system based on semantic networks which confirmed that the QA set is capable of usage in a typical machine learning system as a gold standard set.

This research has provided insights on how to develop any typical QA dataset for a low resource language and should therefore be useful for research interest in the area of QA annotation. Challenges in developing such QA sets for low resource languages included the readability of the raw texts themselves and the language constructs of the raw corpus. In some cases the authors of the original stories may not have used language that was grammatically correct or the stories themselves did not make sense. However, such can be

resolved by an exclusion criteria, though this is done at the expense of reducing on the total number of the annotated data. The purposive sampling employed assisted in addressing such issues. The sampling was also restrictive to particular types of texts i.e. short prose texts, hence longer stories that may probably have been challenging in machine learning systems were not considered. These are opportunities for further research.

This QA dataset of Swahili, KenSwQuAD, is now available as both a collection of 1,445 story texts and a separate comma separated values (CSV) file with the 7,526 QA pairs with at least 5 QA pairs per story. It is released as an open-source dataset to the research community for further review, research and exploration. Future works are possible in areas such as expanding this corpus by inclusion of more stories that were not annotated due to sampling, time and budget constraint. Our research methodology can also be used to annotate other low resource languages such as Dholuo and Luhya, which was already collected as part of the Kencorpus project, or any other low resource language that has a data repository. Availing such datasets and annotated QA sets in the public domain shall continue to be of real value to researchers. Such sets as KenSwQuAD also provide researchers with many opportunities to test their machine learning models on a gold standard QA dataset. KenSwQuAD is publicly available on the project website at the link below:

www.kencorpus.co.ke/kenswquad

Acknowledgments

This research was made possible by funding from Meridian Institute's Lacuna Fund under grant no. 0393-S-001

We acknowledge the inputs of the following research assistants who did the annotations: Alice Muchemi, Eric Magutu, Henry Masinde, Naomi Muthoni, Patrick Ndung'u and Rose Nyaboke

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, September). *Kiswahili Part-of-Speech Tagger - Demo AfLaT.org*. https://www.aflat.org/swatag

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues peu dotées*.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. In *Speech Communication* (Vol. 56, Issue 1). Elsevier B.V. https://doi.org/10.1016/j.specom.2013.07.008

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).

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*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

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

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

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

Libovický, J., Rosa, R., & Fraser, A. (2019). How Language-Neutral is Multilingual BERT? *ArXiv Preprint ArXiv:1911.03310*.

Novák, V., Hartrumpf, S., & Hall, K. (2009). Large-scale semantic networks: annotation and evaluation. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (DEW '09)*, 37–45.

omniglot. (2020). *Swahili alphabet, pronunciation and language*. https://omniglot.com/writing/swahili.htm

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., & others. (2020). Masakhane–Machine Translation For Africa. *ArXiv Preprint ArXiv:2003.11529*.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543.

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

RDF Grapher. (2021). https://www.ldf.fi/service/rdf-grapher

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/

Selamat, A., & Akosu, N. (2016). Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University - Computer and Information Sciences*, *28*(4), 457–469. https://doi.org/10.1016/j.jksuci.2014.12.004

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

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

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). *Kencorpus - Kenyan languages corpus*. https://kencorpus.co.ke/

wikipedia. (2020). *Swahili language - Wikipedia*. https://en.wikipedia.org/wiki/Swahili_language

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.