BUILDING TEXT AND SPEECH DATASETS FOR LOW

RESOURCED LANGUAGES : KAMBA, KISII, KALENJIN, LUO, SWAHILI

Dr. Andrew Kipkebut Peter Rogendo

Kabarak University Kabarak University

AKipkebut@kabarak.ac.ke progendo@kabarak.ac.ke

**ABSTRACT**

Africa has around 2144 living languages (Eberhard et al,. 2019) ; however, those languages are not well represented in the existing Natural Language Processing ecosystem, the main reason for this being lack of data resources for training models. Kamba, Kisii, Luo and Kalenjin, all low resourced languages from Kenya, are such languages that cannot benefit from modern natural language processing techniques due to the absence of high-quality datasets. By collecting, annotating, and validating a significant corpus of texts in these languages, this project aims to create, curate and annotate high-quality parallel dataset to enhance NLP applications for these low resourced languages to enhance NLP applications, including machine translation and sentiment analysis. This initiative will address the linguistic diversity and technological inclusivity needs in East Africa, providing valuable resources for academic and practical applications.

**INTRODUCTION**

Language is the foundation on which communication rests, allowing us to share ideas and interact with one another (Adebara & Abdul Mageed, 2022). One of the NLP applications is machine

translation (MT), which helps facilitate human-machine and human-human communications (Abate et al., 2019). Data availability is one of the criteria to categorize one language as a high or low resource language (Ranathunga et al., 2021). Recently, interest in low-resource MT has been increasing both within the MT research community (Haddow et al., 2022), as well as in native speaker communities (Nekoto et al., 2020). Modern NLP technologies, however, have primarily been developed in Western societies (Adebara & Abdul-Mageed, 2022). The current state-of-the-art(SOTA) MT models were trained on enormous datasets, including sentences in a source language and their corresponding target language translations, which is the most effective of these systems(Tonja et al., 2023).

To date, there is no publicly available MT system for Kamba, Kisii, Kalenjin or Luo languages and are not represented in the commercial MT systems such as Lesan , Google Translate , Microsoft Translator , and Yandex Translate . These languages are also not included in large-scale pre-trained multilingual models like NLLB(Teamet al., 2022), MT5(Xue et al., 2021), ByT5(elalliance, 2022), and M2M-100(Fan et al., 2020). This makes it harder for people to learn and use the languages, even more so to our brothers living abroad. So, by focusing on Kamba,Kisii, Luo and Kalenjin, this dataset aims to bridge the gap between historical linguistic treasures and modern technological advancements. Through this work, we aim to provide a dataset for researchers and technologists aiming to advance machine translation capabilities, linguistic studies, and cultural preservation efforts. Furthermore, by enriching the available resources for Swahili, we contribute to the broader goal of advancing low-resource languages.

Language resources for underrepresented languages like Kamba, Kisii, Kalenji and Luo are scarce, hindering the development of effective NLP applications (Babirye et al., 2022). This project focuses on creating a Swahili-Kamba-Kisii-Kalenjin-Luo parallel dataset to fill this gap. The dataset will be used to train and evaluate NLP models, contributing to better communication tools and technological solutions for these linguistic communities.

**RESEARCH OBJECTIVES**

The research objectives are:

1. To gather and collect parallel texts in Swahili, Kamba, Kisii, Kalenjin and Luo.
2. To ensure high-quality alignment and labeling of the dataset.
3. To Train NLP models for sentiment analysis and Machine Translation using the dataset and assess their performance.
4. To curate and validate the annotated data to create a reliable and standardized parallel corpus.

**JUSTIFICATION OF THE STUDY**

The purpose of this project is to collect and annotate Swahili-Kamba-Kisii-Kalenjin-Luo sentence pairs and digitize the languages. With African languages being significantly underrepresented in the field of Natural Language Processing, majority of NLP advancements have centered around languages with abundant data resources, such as English, French, Chinese, and Spanish. African languages like Kamba, Kisii, Kalenjin and Luo languages, which are spoken by millions of people, lack the necessary datasets for effective NLP applications. This disparity limits the potential for technological innovations that could benefit speakers of these languages.

At the heart of this disparity sits google translate and other commercial machine translation systems

and they do not support these local languages. This omission reflects the broader lack of resources and tools available for this language. Developing a Swahili-Kamba-Kisii-Kalenjin-Luo parallel dataset would be a crucial step in addressing this gap.

Documenting and digitizing these languages will help preserve these languages for future generations. This is particularly important as many African languages face the threat of extinction as with the price of advancing academically is the abandonment of the culture, traditions and traditional language perceived as for the illiterate. Providing technological support for these languages ensures that the cultural identity and heritage of their speakers are represented in the digital world.

Importantly, this research has a strong emphasis on cultural sensitivity and local relevance. It utilizes data specific to the Kenyan context, ensuring that the findings and solutions are directly applicable to the local population. Additionally, improving NLP tools can boost economic opportunities by facilitating better communication in local languages, enhancing customer service, and providing new avenues for content creation in Swahili, Kamba, Kisii, Kalenjin and Luo.

**SIGNIFICANCE OF THE STUDY**

This study holds significant potential for impact, particularly in Kenya and the broader African context in the following ways:

* **Improving Access to Information:** This research will help enhanced NLP applications and will make it easier for speakers of Swahili, Kamba, Kisii, Kalenjin and Luo to access information online, bridging the digital divide and promoting digital literacy.
* **Promoting Inclusivity:** By addressing the needs of underrepresented language communities, this study promotes inclusivity and ensures that technological advancements benefit a broader range of users.

By engaging with the local community for data collection and validation, we will foster a sense of ownership and pride in the project. The dataset will also contribute to preserving the linguistic and cultural heritage of the Swahili, Kamba, Kisii, Kalenjin and Luo-speaking communities.

* **Supporting Research and Development:** The dataset will serve as a valuable resource for researchers and developers, fostering further advancements in NLP for low-resource languages and encouraging more inclusive AI development.

### EXPECTED OUTCOMES

In this research study, we anticipate several key outcomes:

1. **Collection, Annotation, and Verification of Data:**
   * We expect to collect a substantial amount of Swahili, Kamba, Kisii, Kalenjin and Luo sentence pairs from the local communities. This process will involve rigorous annotation and verification to ensure the quality and accuracy of the data. The collected dataset will be a valuable resource for various NLP applications and further research.
2. **Development of NLP Models:**
   * Using the collected dataset, we aim to train a simple sentiment analysis model and a text classifier for these local languages. These models will serve as foundational tools for more advanced NLP tasks and demonstrate the practical applications of the dataset.
3. **Research Publications:**
   * We anticipate producing several research papers based on the application of the Swahili-Kamba-Kisii-Kalenjin-Luo dataset to real-world problems such as sentiment analysis, text classification, and machine translation (MT) model training. These publications will contribute to the academic body of knowledge and highlight the importance of developing resources for low-resource languages.
4. **Dissemination and Stakeholder Engagement:**
   * Besides producing research publications, we plan to present our findings at major Artificial Intelligence symposiums and conferences. Engaging with different stakeholders in the NLP sector, including researchers, developers, and policymakers, will help in disseminating the results and fostering collaborations for further advancements in the field.

**METHODOLOGY**

The primary aim of this study is to create, design, curate, and annotate a high-quality Swahili, Kamba, Kisii, Kalenjin and Luo parallel dataset to support NLP applications. This project will be conducted in three main phases; data collection phase, data processing and annotation phase, and dataset validation and dissemination phase.

#### 1) Data Collection Phase

The data collection phase will be conducted through three distinct processes. The first approach involves gathering a set of Swahili sentences to be translated into Kamba, Kisii, Kalenjin and Luo and then allocating the translation tasks to the translators. The second process focuses on the validation of the translated sentences to ensure accuracy and quality. The third process involves voice recoding the translated sentences.

**Steps:**

* **Text Sources:** Identify and compile Swahili texts from books, newspapers, websites, and other written materials with no copyright infringements.
* **Collaborations:** Collaborate with local authors, publishers, and content creators to access a wide range of texts.
* **Information Sessions:** Organize informational sessions to explain the project’s objectives, benefits, and data collection process to the community members.
* **Job Advertisements:** Advertise job openings for translators proficient in Swahili, Kamba, Kisii, Kalenjin and Luo through local media, community centers, and social media platforms.
* **Selection Criteria:** Preference will be given to individuals with prior translation experience and linguistic knowledge.
* **Training Guidelines Provision Workshops:** Conduct training sessions for the recruited translators to ensure they understand the project’s goals, annotation standards, detailed guidelines and tools for consistent and accurate translations.

#### 2) Data Processing and Annotation Phase

**This phase is characterized by** translating and annotating the collected texts to create a high-quality Swahili, Kamba, Kisii, Kalenjin and Luo parallel dataset.

**Steps:**

* **Assignment of Sentences:** Assign the extracted Swahili sentences to the recruited translators for translation into Kamba.
* **Peer Review:** Establish a peer review system where translators review each other’s work to identify and correct errors.
* **Consistency Checks:** Develop and apply automated quality checks to identify inconsistencies, duplicates, and other potential issues in the dataset.
* **Statistical Analysis:** Use statistical methods to analyze the quality of the translations and annotations, ensuring they meet predefined criteria.
* **Voice Recordings:** Once the text translation is complete, the last step is recording voice data so as to collect the speech data of the translated sentences.

#### 3) Dataset Validation and Dissemination Phase

To ensure the dataset's accuracy, we compile it into a structured format, and make it accessible to researchers and developers.

**Steps:**

* **Data Cleaning:** Perform data cleaning to remove any errors, inconsistencies, and irrelevant content from the collected dataset.
* **Standardization:** Standardize the format of the sentence pairs to ensure uniformity and ease of use.
* **Metadata:** Add relevant metadata to each sentence pair, including source information, context, and any additional annotations useful for NLP tasks.
* **Secure Storage:** Store the cleaned and formatted dataset in a secure and accessible database.
* **User Access:** Provide access to the dataset for researchers, developers, and other stakeholders through a user-friendly interface or API.
* **Research Papers:** Publish research papers based on the collected dataset and its applications in reputable journals and conferences.
* **Presentations:** Present the findings at major Artificial Intelligence symposiums and NLP conferences to engage with the wider research community.
* **Repository Creation:** Create an open access repository for the Swahili, Kamba, Kisii, Kalenjin and Luo dataset, allowing researchers and developers to freely access and use the data.
* **Documentation:** Include detailed documentation and guidelines to facilitate the use of the dataset in various NLP tasks.

By following this methodology, we aim to create a high-quality Swahili-Kamba-Kisii-Kalenjin-Luo parallel dataset of 30,000 sentence pairs, that will significantly contribute to the development of NLP applications for these languages, promoting linguistic diversity and technological inclusivity.

**BUDGET**

The details of the budget are outlined in Table 1. It is to be noted that the salaries of the

personnel are based on a delivery of a task and not the monthly salary rate.

|  |  |  |  |
| --- | --- | --- | --- |
| **ITEM** | QUANTITY | UNIT PRICE | **TOTAL (Ksh)** |
| Sentence Translation | 30,000 sentences x 4 | 120,000 x10/= | 1,200,000 |
| Swahili Sentence collection | 30,000 sentences | 30,000 x 10/= | 300,000 |
| Voice annotators | 30,000 sentences x 4 | 120,000 x 10/= | 1,200,000 |
| Travel and Training | 6 | 15,000 | 85,000 |
| Research Lead | 1 | 100,000 x 6 months | 600,000 |
| Research Assistant | 2 | 50,000 x 6 months | 300,000 |
|  |  |  |  |
| **TOTAL** |  |  | 3,685,000 |

*Budget distribution*

**Budget Justification**

The total estimated expenditure for this research project, summing up all the categories, is Ksh 3,685,000. The total is derived from the individual allocations to each category, ensuring a comprehensive and balanced distribution of funds across all critical aspects of the project as show in the table above. The following is the justification for each individual category.

* **Sentence Translation –** This category, allocated Ksh 300,000, covers the essential sentence translation of the 30,000 Swahili sentences to Kamba, Kisii, Kalenjin and Luo. At a price of Ksh 10 per every translation, it totals up to Ksh 1,200,000.
* **Swahili Sentence collection-** This category, receiving Ksh 300,000, is allocated to cover for the 30,000 Swahili sentence generation with each sentence costing Ksh 10.
* **Voice annotators –** This segment, receiving Ksh 300,000, is set aside to cater for the capture and entry of the Kamba, Kisii, Kalenjin and Luo voice dataset, comprising of the 120,000 (30,000 x 4) translated sentences. A single voice entry costing up to Ksh 10.
* **Travel and Training -** Receiving Ksh 85,000
* **Research Lead -**
* **Research Lead Assistant -**

**REFERENCES**

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, Armand Joulin. 2020. [Beyond English-Centric Multilingual Machine Translation](https://arxiv.org/pdf/2010.11125)

Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh, Jugal Kalita. 2023. [Enhancing Translation for Indigenous Languages: Experiments with Multilingual Models](https://aclanthology.org/2023.americasnlp-1.22.pdf)

Barry Haddow, Rachel Bawden, Antonio V. M. Barone, Jindřich Helcl, Alexandra Birch. 2022 [Survey of Low-Resource Machine Translation](https://aclanthology.org/2022.cl-3.6.pdf).

Claire Babirye, Joyce Nakatumba-Nabende, Jeremy Franics Tusubira & Jonathan Mukiibi,Andrew Katumba & Ronnie Ogwang, Medadi Sentanda, Lilian D. Wanzare, Davis David. 2022. Building Text and Speech Datasets For Low Resourced Languages: A Case Study of Languages in East Africa.

David M Eberhard, Gary F. Simons, and Charles D. Fenning. 2019. [Ethnologue: Languages of th](https://www.ethnologue.com/region/Africa)e

[worlds. Twenty-second edition](https://www.ethnologue.com/region/Africa)

[Ife Adebara](https://arxiv.org/search/cs?searchtype=author&query=Adebara,+I), Muhammad Abdul-Mageed. 2022. [Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go.](https://arxiv.org/pdf/2203.08351)

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel.2021. [mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer](https://aclanthology.org/2021.naacl-main.41.pdf)

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel. 2022. [ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models](https://aclanthology.org/2022.tacl-1.17.pdf)

Surangika Ranathunga, En-Shiun A. Lee, Marjana P. Skenduli, Ravi Shekhar, Mehreen Alam, Rishemjit Kaur. 2021. [Neural Machine Translation for Low-Resource Languages: A Survey](https://arxiv.org/pdf/2106.15115)

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, Abdallah Bashir. 2020. [Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages](https://aclanthology.org/2020.findings-emnlp.195.pdf)