Journal of Languages, Linguistics and Literary Studies (JLLLS), Vol. 4, No. 3, 2024
ISSN: 2788-4481 (**Print**) | 2788-4473 (**Online**)
Journal Website: https://journals.jozacpublishers.com/jllls
https://doi.org/10.57040/tvpzzk79

Framework for building linguistic corpora for a large language model project for the Heritage Nubian Language of Kenya

Peter Nyansera Otieno[1*]
[1]*Languages, Linguistics and Literature, Kisii University, Kenya, potieno@kisiiuniversity.ac.ke*

Received: 24 July 2024 | **Accepted:** 15 October 2024 | **Published:** 13 November 2024

Abstract: Low-resource languages face an uphill task in their documentation and preservation. Language technologies offer a way out for these beleaguered languages. However, these technologies depend on developing high-quality linguistic corpora absent in understudied and under-resourced languages like the Kisii Town Nubian. This study aims to develop a framework for constructing linguistic corpora for the Kisii Town Heritage Nubian language that can be used to develop an LLM and other language technologies. The objectives are to: 1) develop a framework for data collection and metadata labeling, and 2) Identify the main tenets to be considered in developing the language technologies. The methodology used to collect data will utilize local community knowledge experts and opinion leaders at Nyanchwa where the Nubians reside. The process will draw from linguists' knowledge of the language's terrain with necessary permissions and consents sought in the process. Diverse data will be assembled from written texts, recorded audio, web scrapings, and word lists to comprehensively view the language. This will be followed by data processing and annotation. The processed data will be trained on linguistic features such as phonology, morphology, syntax, semantics, and parts-of-speech labeling. This will then be structured into selective linguistic corpora with robust quality control guidelines. The deliverables of the project will be linguistic corpora for various domains of the Nubian language, the development of language technologies, and comprehensive documentation of linguistic corpora. The results of this project will be consequential in the field of language documentation and technological support for this endangered language.

Keywords: Annotation, Framework, Large Language Model, Linguistic corpora, Parts-of-speech labeling, Web scrapings

Biographical notes: **Peter Nyansera Otieno, Ph.D.**, (circa. 1976) is a linguist with the Department of Languages Linguistics and Literature, at Kisii University. He teaches undergraduate and graduate courses in the areas of Phonetics, Phonology, Morphology, and Communication Skills. Nyansera is an interdisciplinary scholar. He has research interests in Acoustic Phonetics, Bantu sound systems, Socio-Phonetics, lexicography, Language impairment, Natural Language Processing, Linguistic Anthropology, Language endangerment, and Language Documentation. He is a crusader for Gusii Language and Culture revitalization. He writes poetry and short stories from his rich EkeGusii oral literature background. He patronizes the Bobasi Chapter of the AbaGusii Cultural and Development Council and the Mwanyagetinge Heritage Council.

1. Introduction

One of the most critical problems facing linguists today is the endangerment and possible death of languages. This makes the preservation and documentation of world languages the most urgent undertaking. It is estimated that of the 7000 living languages today, about 40% of them are bound to die in a decade or so down the course of time (Moseley, 2010). Around 19% of the languages are no longer being passed down from parents to children. Multilingualism and other forces of modernism have exerted pressure on low-resource languages, to the extent that linguists must try and preserve the very raw material of their craft through documentation and revitalization efforts.

This project aims at leveraging language technology to boost the chances of resiliency for low-resource languages like Kisii Town Heritage Nubian language. That is the reason why LLMs are taunted to harbor prospects of ensuring that the language remains extant in the face of incursion from dominant languages. Low-resource languages like Kisii Town Nubian and destined for loss unless intervention efforts are put in place. The loss of language is also equal to the loss of a whole range of Indigenous cultural knowledge that will make our world poorer.

Recent studies have shown that using LLMs can be instrumental in preserving and helping in revitalization efforts of endangered languages (Anacka et al., 2021; Ruder et al., 2021). This is so since LLMs are AI-powered tools tuned to

a wide range of language tasks like text-to-speech translators, speech-to-text translators, and language modeling among others. Joshi et al. (2020) insist that all this is possible depending on the efficacy of linguistic corpora, which are not often available for under-studied languages like Kisii Town Nubian.

The present framework aims to address the lack of linguistic corpora in Kisii Town Nubian by creating a mechanism for collecting, annotating, and curating data for the Kisii Town Nubian language. To this end, the framework details ways of developing linguistic corpora for Kisii town Nubian, devising an LLM to support multiple uses in creating tools for language revitalization and preservation. The framework also anticipates the incorporation and full participation of local indigenous knowledge experts and members of the community to own the whole process.

1.1. Problem statement

Kisii Town Heritage Nubian is endangered and at risk of being lost in due course if efforts are not made to reverse the trend of it being edged out from nearly all domains of language use among the community members. Of greater significance is the emerging trend of lack of intergenerational transmission of the language. If intervention efforts are not undertaken the language may die. Language revitalization and preservation can be enabled and hastened by leveraging technology, especially using LLMs. The efficacy of this tool depends on the data used to build corpora which is lacking in many African understudied languages like Kisii Town Nubian. The lack of comprehensive linguistic resources for Kisii Town Heritage Nubian is a gap that this study intends to fill.

1.2. Objective

The main objective of this project was to develop a framework for gathering high-quality data and setting up a comprehensive LLM for the Heritage Nubian language. The following are the specific objectives of the study.

- **i.** To establish a robust stakeholder engagement process to collaborate with local Nubian communities, linguists, and language experts to understand the language landscape and obtain necessary permissions for data collection.
- **ii.** To devise and implement an effective data collection strategy to gather diverse linguistic resources for the Heritage Nubian language, including written texts, audio recordings, and transcripts, to capture the breadth of the language.
- **iii.** To design and implement a rigorous annotation process, including the development of detailed guidelines and quality control measures, to ensure high-quality labeling of linguistic features such as part-of-speech, morphology, syntax, and semantics.
- **iv.** To organize the annotated data into a structured and versioned linguistic corpus, with comprehensive documentation to facilitate the corpus's use and reusability by the broader research community.

1.3. Significance of the study

This study can significantly lend itself to the documentation and preservation of the endangered Heritage Nubian language, potentially the basis for developing multiple language technologies in the Kisii Town Nubian. It also contributes to the scanty literature for low-resourced languages, empowering local communities, and enabling the inclusivity of understudied languages in leveraging language technologies for even development in Natural Language Processing.

Specifically, this study is significant in the preservation and documentation of Kisii Town Heritage Nubian. This variety of Nubian is critically endangered necessitating efforts to preserve and document it for posterity. The language community operates in the context of multilingualism and modernity which has presented English and Kiswahili into all domains of life effectively eclipsing the Heritage Nubian language. This study seeks to develop language resources accessible to the community working towards its maintenance or preservation.

This framework sets out to develop a high-quality database to form a foundation for building a Large Language Model and ensuring a range of linguistic tools like translators, text generators, and language modeling. This will be a big step towards equipping a low-resource language like Kisii Town Heritage Nubian.

The framework and methodologies devised by this study form the basis for creating reference points for low-resource languages in developing linguistic tools. By identifying best practices in stakeholder engagement, data collection, annotation, and corpus curation, this project can contribute to the broader field of language documentation and the use of advanced technologies for endangered language preservation.

This project sets the stage for local community engagement and empowerment in partnership with academia to ensure that Kisii Town Nubian community members own the process of revitalizing their language. The collaboration between academia and the community ensures respect for the speakers' linguistic and cultural heritage.

A full implementation of this framework envisages the development of language technologies for Heritage Nubian that will be able to address the limitations that exist in this and other low-resource languages in the field of Natural Language Processing and Artificial Intelligence. This will help unleash the potential of leveraging technology in preserving and documenting endangered languages.

2. Literature review

This project on building linguistic corpora for Kisii Town Heritage Nubian language seeks to offer solutions to the following challenges that stare at understudied and under-documented languages in general. Language endangerment has now become a global problem that requires all stakeholders to do something to turn back this momentum to avert language loss (Simons & Lewis, 2020; Simons & Krauss, 2019). Efforts to address this absurd trend include documenting endangered languages, working on efforts to maintain these languages, and reviving the dying and

critically endangered languages. This framework seeks to bear the effort and urge urgent investment in language documentation projects, and community-led initiatives to preserve linguistic diversity (Bromham et al., 2021).

2.1. Linguistic corpora and low-resource languages

The cascaded envisaged project relies heavily on the availability of high-quality language data from different contexts. These data are the basic resources for language documentation and crafting of language tools such as machine learning, machine translation, speech recognition, and translation tools (Sinclair, 1996; Biber et al., 1998). This is challenging to achieve for low-resource languages like Kisii Town Heritage Nubian that lack sufficient data in terms of books, standardized orthographies, and detailed linguistic descriptions (Besacier et al., 2014; Anastasopoulos & Neubig, 2020).

Building of dependable corpora for Kisii Town Heritage Nubian can be enhanced using modern ways. This includes but is not limited to, using social media posts (Dunn, 2020), leveraging crowdsourcing techniques (Sadat et al., 2014), and developing semi-automated annotation workflows (Bender & Lascarides, 2019). This also must work in tandem with local communities as custodians of their language and culture apart from also being the source of the data that will be incorporated in the corpus building (Leonard & Haynes, 2010; Rice, 2011).

2.2. Large language models and language preservation

Large Language Models trained on large, high-quality language corpora have shown great potential in performing translation tasks, language generation tools, and question-answering tasks which can be utilized to support low-resourced languages and endangered cultures in preserving and revitalizing them (Brown et al., 2020; Xiong et al., 2020).

However, this great promise of developing robust LLMs can be realized with the availability of comprehensive linguistic corpora (Nekoto et al., 2020; Caswell et al., 2021). That is why today there is a great reawakening of the need to advance NLP technologies and consequent reliable linguistic corpora, robust data collection and annotation protocols, and close collaboration with community stakeholders to ensure ownership and cultural and linguistic validity of the resulting tools (Blasi et al., 2022; Bender et al., 2021).

With the advancement of LLMs like GPTs and BERT, a lot of work can be done in the preservation of languages and cultures. Even diminishing flora and fauna can be kept and sustained in the database for many generations in perspective to inquire into. Linguists and engineers can use the corpora to model a variety of tools that can be used by different professionals to enhance the agenda of language preservation (Bender et al., 2020).

3. Research methodology

This framework is firmly based on local community involvement. Researchers will engage local knowledge experts, key stakeholders, and academicians in linguistics and computer science to take stock of the linguistic landscape of Kisii Town Heritage Nubian. By working together and getting the necessary consent and permissions, the whole exercise of data collection will be legitimate, trusted, and comprehensive. This will also ensure that the project will align with the community's expectations, and goals and that the entire process respects their cultural heritage.

3.1. Data collection

After stakeholder engagement, the next step will be to collect as much data as possible for Kisii Town Heritage Nubian in the form of audio recordings, written texts, and transcripts. These should cut across diverse subject contents, genres, and styles to reflect as much as possible the entire length and breadth of the language (Besacier et al., 2014). Care should be taken to ensure that the data is representative of the linguistic variation within the community – dialects, and registers.

3.2. Data preprocessing and annotation

Preprocessing is an important stage that the collected data will undergo. This will involve data cleanups, labeling, and normalization to ensure consistency and quality. This behooves the researchers to develop detailed annotation guidelines. This can involve the research team training research assistants, most ideally, from the community to label the data with linguistic features like parts of speech, phonology, morphology, syntax, and semantics. This must be done carefully to ensure high levels of consistency and accuracy.

3.3. Corpus curation and quality assurance

All the data that has now been annotated will be compartmentalized into various linguistic corpus forms following the best practices in the literature for corpus design and management. High-fidelity corpus control measures such as inter-annotator agreement tests and automated validation checks, will follow to ensure the reliability and reusability of the corpus. The corpus will be thoroughly documented, including metadata, provenance, and licensing information, to facilitate its use by the broader research community.

4. Findings and discussions

This framework for crafting linguistic corpora for the Kisii Town Nubian acted like a pilot study with pointed outcomes. It began with engagement with local Nubian community leaders and members who allowed for a good rapport between the researcher and the participants. This engagement has yielded promising results on the envisaged

project especially insights into the language's context and requirements. It has also secured the required necessary permissions, consents, and support for data collection.

The data collection phase was able to successfully put together different types of linguistic resources that included written texts, audio and video recordings, and transcripts across various genres and styles of Heritage Nubian from the residents of Kisii Town. It was noted however that written texts like published books are rare in the language. It was seen that even the orthography was not standardized.

Data preprocessing and annotation were of prime importance to make accurate labeling of the linguistic features in the language such as parts of speech, phonology, morphology, syntax, and semantics. Quality control measures, which include inter-annotator agreement tests and automated validation checks have been carried out on the data to ensure reliability and consistency all along the annotations.

The miniature linguistic corpus for this framework represents a significant milestone for this under-resourced language. It is organized in a structured and versioned format, making it accessible and reusable for the broader research community once uploaded into GitHub. Detailed documentation, including metadata, provenance, and licensing information, further enhances the resource's accessibility and transparency.

All the above preliminary phases were done to give way for the next phase of the project which is the actual creation of a Large Language Model for Kisii Town Heritage Nubian. The comprehensive corpus will facilitate the training and fine-tuning of the LLM. In turn, the LLM can be used to model tools for various language-related tasks like speech-to-text and text-to-speech tools, machine translation, and language modeling.

The best practices adopted for this project can also be effected in reference to create linguistic resources for other low-resource languages which abound in the African context. This means that the same procedure, especially for local community engagements, data collection, data annotation, and corpus curation, can work for other language's projects to preserve or even revitalize linguistic diversity.

Despite the progress achieved, the project researcher recognizes the ongoing challenges and limitations in working with this low-resource language. The most glaring shortcoming is the inadequacy of written texts in the language. The community writers must be enabled to produce as many texts as possible for the training of language tools. Below is an excerpt of a story collected for the corpus.

> *Zaman khan pii subiana tinim khan uman soo ajana kuluu khan umon fu badu. Ahasa kulu houmon khan ebil umon misen kha umon fii ma tenebia seme umon khan glasma kalama ta ana toumo.Lagadi pui kakan garaya khan umon anasa ta aulan. Yom wai jaa umon khan giamja fii wele wai too umon ainarartas ta asue alfi ma ajana…*

(Once upon a time there lived two great friends. They did everything together. Everyone in their village loved them because they were children of good morals. They were obedient and followed diligently whatever their parents told them. Even in school, they topped in their examinations. One day, as they were going home from school, one of them saw a black shopping bag…)

Ensuring the long-term sustainability and maintenance of the linguistic corpus, as well as addressing potential biases and gaps in the data, will necessitate continuous collaboration with the Nubian community. This research was committed to tackling these challenges and refining the framework to better meet the needs of the language and its speakers.

5. Conclusion

5.1. Summary

This framework developed a roadmap to crafting a comprehensive linguistic corpora database to help in creating a Large Language Model for Kisii Town Heritage Nubian, an endangered Nilo-Saharan language spoken in Kisii, Kenya. This framework had the following main objectives: to develop a comprehensive framework for collecting, annotating, and curating linguistic data for Kisii Town Heritage Nubian; Identify best practices and key considerations that will inform the development of linguistic resources for other under-resourced languages.

The methodology used by this framework was multidimensional involving local community involvement, collection of data, preprocessing and annotating the data, and corpus curation with quality assurance measures. The researcher and his team worked closely with local Nubian indigenous knowledge experts, local leaders, general members of the community, language experts, academicians, and computer scientists to help understand and appreciate the linguistic terrain, obtain necessary permissions and consents, and utilize each person's expertise. The result is a comprehensive corpus is a wide range of genres and linguistic phenomena, with detailed documentation to facilitate its use and reusability.

5.2. Conclusion

The development of this framework for building linguistic corpora for Kisii Town Heritage Nubian was significant in laying the foundation for creating an LLM that can help model various language tools. These tools can be leveraged to support language preservation and revitalization efforts. This framework and best practices established through this work can be references for other low-resourced languages to use in preserving linguistic diversity that is under threat in the face of multilingualism and other modern trends.

However, several challenges came up that need to be surmounted. Low-resourced languages like Kisii Town Nubian lack enough text, have data biases, and have gaps in linguistic data now that the language is understudied in its phonology, morphology, syntax, and semantics. Continued collaboration with the Nubian community and a commitment to addressing these challenges will be crucial for the success and impact of this project.

5.3. Recommendations

This study made the following recommendations:

1. Adopt a Comprehensive Data Collection Approach. Every effort should be made to gather a diverse, high-quality range of linguistic data.
2. Invest in Robust Annotation Processes. More research assistants are needed to develop detailed annotations on the data which will ensure high quality and consistent labeling of linguistic features.
3. Implement Rigorous Quality Assurance Measures. To maintain the reliability and reusability of the linguistic corpus, employ a range of validation techniques, such as inter-annotator agreement tests and automated checks.
4. Ensure Long-term Sustainability.
5. Promote Knowledge Sharing.

References

Adelaar, W. F., & Muysken, P. (2004). The Languages of the Andes. Cambridge University Press.

Anacka, A., Ruder, S., & Pierrehumbert, J. B. (2021). Adapting large language models for low-resource languages. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4336-4346.

Anastasopoulos, A., & Neubig, G. (2020). Should All Cross-Lingual Embeddings Speak English? *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8658-8679).

Bani-Shoraka, H., & Matthewson, L. (2020). The Yolŋu Languages of Northeast Arnhem Land. Oxford University Press.

Bender, E. M., & Lascarides, A. (2019). Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax. Morgan & Claypool Publishers.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication*, 56, 85-100.

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2022). Human Sound Systems Are Shaped by Post-Neolithic Changes. *Science*, 372(6540), eaax0762.

Bromham, L., Dinnage, R., Skirgård, H. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nat Ecol Evol* **6**, 163–173 (2022).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Caswell, I., Kreutzer, J., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., ... & Straka, M. (2021). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 9, 1520-1544.

Crystal, D. (2000). Language Death. Cambridge University Press.

Dunn, J. (2020). Computational Learning of Language from Pockets of Stability in Sound and Meaning. *Language*, 96(1), e1-e30.

Grenoble, L. A., & Whaley, L. J. (2006). Saving Languages: An Introduction to Language Revitalization. Cambridge University Press.

Harrison, K. D. (2007). When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge. Oxford University Press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.

Leonard, W. Y., & Haynes, E. (2010). Making "Collaboration" Collaborative: An Examination of Perspectives that Frame Linguistic Field Research. *Language Documentation & Conservation*, 4, 268-293.

Moseley, C. (2010). Atlas of the world's languages in danger. UNESCO.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., ... & Prinsloo, D. A. (2020). Participatory Research for Low-Resource Machine Translation: A Case Study in African Languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2144-2160).

Nettle, D., & Romaine, S. (2000). Vanishing Voices: The Extinction of the World's Languages. Oxford University Press.

Rice, K. (2011). Documentary Linguistics and Community Relations. *Language Documentation & Conservation*, 5, 187-207.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., ... & Mielke, S. J. (2021). Transfer learning in natural language processing. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* (pp. 8-41).

Sadat, F., Kazemi, F., & Farzindar, A. (2014). Automatic Identification of Arabic Dialects in Social Media. *In Proceedings of the First International Conference on Social Networks Analysis, Management and Security (SNAMS*) (pp. 139-146).

Simons, G. F., & Krauss, M. E. (2019). The world's languages in crisis: A 20-year update. Language, 95(1), 1-16.

Simons, G. F., & Lewis, M. P. (2020). The world's languages in crisis: A 20-year update. SIL International.

Sinclair, J. (1996). Corpus to Corpus: A Study of Translation Equivalence. International Journal of Lexicography, 9(3), 199-224.

UNESCO. (2022). Atlas of the World's Languages in Danger. Retrieved from http://www.unesco.org/languages-atlas/

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 Conversational Speech Recognition System. *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5934-5938). IEEE.