The Current State of the NLP in Sub-Saharan Africa - A Position Paper

Anonymous ACL submission

Abstract

In this study, we present our position on the current state of the NLP in Sub-Saharan Africa. Our position comes from surveys of literature we conducted on NLP activities (research and publications) in Sub-Saharan African languages. We discussed issues of NLP research outcomes for Sub-Saharan Africa based on the results of the survey. Some of these issues include low-quality results of NLP studies, insufficient access to research funding, and lack of an interdisciplinary approach to NLP research in the region's languages. Results of the study reveal that most of the NLP work done for Sub-Saharan African languages from 2020-2023 was centered around corpus development, language modeling, and sentiment analysis. About 61% of the NLP work in sub-Saharan Africa does not have access to funding. Funding sources are mainly NGO-driven with 66.7% of work that received funding being multilingual studies. However, 64% of NLP activities in the region are monolingual. We finalize our position by providing recommendations on addressing issues raised and discovered based on the survey.

1 Introduction

Natural Language Processing (NLP) has emerged as a transformative field in artificial intelligence, revolutionizing how humans interact with machines and systems. NLP, a subfield of artificial intelligence, involves the development of algorithms and models that enable computers to understand, interpret, and generate human language. As the applications of NLP continue to increase globally, it is crucial to assess the state of this technology in different languages and geographical contexts. In this position paper, we look into the current state of NLP in Sub-Saharan Africa. We aim to highlight the challenges and opportunities in harnessing this technology for the region's linguistic diversity.

Sub-Saharan Africa, a region rich in cultural and linguistic diversity, presents a unique set of challenges and opportunities for adopting and developing NLP technologies. With a rich tradition of storytelling, poems, songs, and literature, in recent years it has seen a proliferation of communication in through digital and social media. One of the distinctive characteristics of the region is its low-resource nature, marked by limited access to digital data, computing resources, and research infrastructure. The low-resource nature of Sub-Saharan Africa poses a significant hurdle to the application of NLP technologies. Different from well-resourced languages with extensive datasets and models, many African languages, have a scarcity of linguistic resources. Despite these challenges, there is a growing recognition of the potential of NLP to address linguistic diversity, promote local language preservation, and contribute to the region's socio-economic development.

This paper aims to explore the current landscape of NLP in Sub-Saharan Africa, from the Horn of Africa in the east to the Westernmost of Africa in the west, down to the Southernmost of Africa in the south. We present the breadth and depth of NLP activities in Sub-Saharan Africa. Discuss the current challenges, areas of strength and weakness, and propose recommendations.

Our discussion is within the context of NLP Algorithms, fundamental applications, corpora, and sources of NLP research funding. By examining these, we seek to identify opportunities for the integration of NLP technologies to address societal needs and linguistic preservation. Also, to contribute to the overall development of the Sub-Saharan African region and its inclusion in the ever-rapid growth of information and communication technology usage. As we navigate through the nuances of NLP in Sub-Saharan Africa, we hope to present our position on collaborative efforts that harness the potential of technology to uplift communities, preserve languages, and foster inclusive development.

2 Fundamental Algorithms and tools for NLP

NLP algorithmic tools are as diverse as corpora (texts) used to model algorithms to achieve a specific NLP application. Algorithms are better utilized when they are applied across multiple languages (Daniel and H. Martin, 2020). Africa has over 1000 languages (at the minimum) grouped into four categories based on their ethnological origin. Language properties differ within and across these groups, most of the NLP work in Sub-Saharan Africa in the past is monolingual and tested on a single language only (C. S. and R., 2020; M. A. et al., 2019; M.A. et al., 2019; Mahloane and Trausan-Matu, 2015; Eckart et al., 2020; Malema et al., 2020; Suleiman et al., 2019; Zakari et al., 2021; Zandam et al., 2023; Rakhmanov and Schlippe, 2022; Ibrahim et al., 2022a,b; Chukwuneke et al., 2022; Awwalu et al., 2021; Ogbuju and Onyesolu, 2019; Abdulmumin and Galadanci, 2019; Tukur et al., 2019; Onyenwe et al., 2018; Ezeani et al., 2018; Rozaimee et al., 2017; Sèmiyou et al., 2013; Afolabi et al., 2013; Mohammed and Prasad, 2024; Bashir et al., 2015; Ralethe, 2020; Augustinus et al., 2016; Augustinus and Dirix, 2013; Gaustad and Eiselen, 2022; Matthew, 2015; Brokensha et al., 2023). However, this is changing, with over 45% of the recent NLP work in Sub-Saharan Africa being multilingual. (Akera et al., 2022; Alabi et al., 2022b,a; Z. et al., 2021; Muhammad et al., 2023; N. et al., 2022; Gaim et al., 2022; Gaustad and Puttkammer, 2022; Adebara and Abdul-Mageed, 2022; Adelani et al.; et ‘al., 2022; Mabokela and Schlippe, 2022; Nekoto et al., 2020; Nyoni and Bassett, 2021; Moses et al., 2022; Griesel and Bosch, 2020; Marivate et al., 2020; Ogueji et al., 2021; Oladipo et al., 2022; E. and P., 2014; Mboning et al., 2020; Rajab, 2022; Nakatumba-Nabende et al., 2024; Tonja et al., 2024; Oyewusi et al., 2021; Dossou and Emezue, 2021; Adelani et al., 2020; Mohammed and Prasad, 2023; Van Zaanen et al., 2014; Eiselen and Gaustad, 2023). (See Figure 6 in Appendix A: 6.)

NLP algorithmic tools used for language preprocessing and modeling can be divided into classical and modern tools. Tokenization and preprocessing techniques, useful algorithms such as Minimum Edit Distance, statistical Hidden Markov Model (HMM), N-gram Language Models, Naive Bayes, and Logistic regression, etc. are classical tools that integrate well with language grammatical structures and variations (Daniel and H. Martin, 2020; Dale, 2010). These have worked well with languages with NLP research history to provide fundamental NLP resource tools and are suitable for developing quality NLP resource tools for low-resource languages. Modern algorithmic tools are a suite of Neural networks and Neural language models. Modern NLP tools are data-driven and machine-learning approaches to language modeling and processing, responsible for the huge success of NLP applications among languages with NLP research history and better language resources. Most work in Africa NLP in recent times utilizes modern approaches. Even though, with the technological advancement of the Internet and social media technologies, there is sufficient raw data (texts, images, videos, and audio) for the Sub-Saharan African languages, the results of NLP research are significantly low when compared with the rich NLP resource languages and languages with a long history of NLP research.

3 NLP Applications for Sub-Saharan African Languages

The Fundamental fields of NLP applications are divided into five groups namely, Information extraction/retrieval (e.g. Name entity recognition, relation detection & classification, etc.), Question answering & Summarization (Question processing & question answering, and summarization), Speech recognition (Automatic speech recognition and text-to-speech), Machine translation and Dialogue system & Conversational agents (Basic dialogue system and chatbot) Daniel and H. Martin (2020). Within these areas of NLP applications are numerous applications or tools such as Chatbots, Sentiment Analysis, Grammar Checker, Spam Detection, Smart Assistant, Market Intelligence, Targeted Advertising, machine translation, speech recognition, automatic summarization, email filtering, social media analytics, etc. Most of the NLP work done for Sub-Sahara African languages in 2020 - 2023 centered around corpus development, language modeling, and sentiment analysis (See Figure 1 in Appendix A: 1), with little work on name entity recognition, speech recognition, part of speech tagging, machine translation, and speech recognition. Despite these efforts, the results of most studies are of low quality and lack standardization and benchmarking. However, 2023 ended with a more collaborative effort across the globe involving numerous Sub-Saharan African NLP research communities that improvised a Name Entity Recognition tool as a benchmark that works for 10 African languages Adelani et al..

4 NLP Datasets (Corpora) for Sub-Saharan African Languages

NLP datasets are known as corpora. Corpus is a linguistic term that refers to the language electronic text in its natural communicative forms without necessarily regard to the language it contains Suleiman et al. (2019). For example, text from newspapers, fiction or non-fiction books, and spoken genres like telephone conversations, and business meetings. Words are products of our conversations. Every text in its natural form is a derivative of one or more speakers or writers, using a specific dialect of a specific language, at a specific time and place for a specific function Daniel and H. Martin (2020). Language models depend on quality corpus and suitable algorithms. Many NLP applications perform better when modeled with large corpus and modern deep learning algorithms. Applications such as machine translation require large parallel corpora for good performance. Only very recently we started having a benchmark corpus for Sub-Sahara Africa NLP tasks. Except for fewer public datasets developed by individual Africa NLP researchers, most existing platforms for Africa Datasets curation are NGOs deriving, with their sources from International organizations. Notable among them are Lacuna funds and AI4D. Despite the efforts in acquiring original Sub-Saharan Africa datasets, most of the datasets have issues of access, proper documentation, and presentation format. While there are several corpora for African languages, you hardly see parallel corpora for two or more Sub-Saharan African languages. The common thing is to have parallel corpora of an African language and English or French, and most of them are lexicon-based parallel corpora curated from online multilingual dictionaries. For example Hausa - English or Kiswahili-English parallel corpora. And most of the time, they are specialized forms of parallel corpus.

5 Issues and Recommendation

1. Sub-Saharan African language NLP has a long way to catch up with global NLP research communities, particularly countries with a long history of NLP research activities. The problem of Sub-Saharan African language NLP ranges from low-quality research resources with interchangeability and interoperability issues, to the absence of multidisciplinary collaborative research involving linguistics and computational linguistics, to the issue of funding NLP research activities in universities.

To address the issue of low-quality tools, standards, interoperability, and interchangeability, we recommend adherence to the principles of software design and software engineering best practices in language modeling. This will help with developing a standard African language resources tool that is largely independent of a specific language and can easily be adapted for a large category of Sub-Saharan language families with interoperability.

2. Another big issue with Sub-Saharan African language NLP is the disconnect between NLP researchers and linguistic domain experts. Sub-Saharan African language linguistics uses methodologies different from NLP-based computing approaches for language processing. Even though, the modern machine learning approach to NLP does not require an understanding of a language's grammatical structures; origin, formation, and derivations, such knowledge is needed to produce reliable educational basic language resource tools. These tools are fundamentals in developing NLP applications that are acceptable, reusable, and deployable.

We recommend that African NLP research communities should focus on the provision of basic language resource toolkits that comprise language preprocessing techniques such as tokenization, word normalization, lemmatization/word form, and word stemming. The toolkits should also include morph-lexical tagging, language model N-grams, and maximum likelihood estimator. Such a toolkit will enhance the rapid prototyping of essential requirements for a new language. Furthermore, there is a need to engage in multidisciplinary research involving Sub-Saharan African language linguistics, computational linguistics, domain-specific experts, and NLP researchers.

3. On the issue of corpora, the study indicated a growing dataset development for Sub-Saharan African Languages. We have started to see benchmark work done. While this is a welcome development, more is still needed for sub-Saharan African languages to grow from low-resource language to rich-resource language. Our study has indicated the complete absence of languages in the Central Africa region of Sub-Saharan Africa in NLP research visibility. In short, less than 4% of Sub-Saharan African languages are currently covered in NLP research radar. This indicates the level of work that needs to be done to bring as many languages as possible.

We recommend a deliberate effort to be made to develop datasets for underrepresented languages of Sub-Saharan Africa. This means that more than the availability of online text is required to cover all kinds of text needed for NLP, specifically for languages that are not visible online. Spoken genres from conversations, interviews, and transcription of stories from late-night storytellers will provide our algorithms with text to process.

4. Another dimension regarding corpora is the lack of parallel corpora for inter-languages of Sub-Saharan African languages. Parallel corpora are required for machine translation applications.

We advocate for the collection of parallel corpora for inter-language or intra-language families of Sub-Saharan African Languages. Doing so will set a new narrative for African NLP in the area of machine translation application. Finally, we recommend the use of a datasheet to provide detailed properties of dataset development. Properties such as motivation, situation, language varieties, speaker demographics, collection process, annotation process, and distribution Daniel and H. Martin (2020) will provide answers to who produced the language dataset, in what context, and for what purpose Daniel and H. Martin (2020).

5. Finally, the problem of funding NLP research efforts for Sub-Saharan African languages. Our study indicated that more than 60% of NLP research for Sub-Saharan African languages is not funded (see Figure 7, Appendix A: 7). This has a negative setback towards achieving interoperable and interchangeable NLP resources for Sub-Saharan African languages. The few available sources of funding come from either institution of the government of some African countries (such as the South African Centre for Digital Language Resources (SADiLaR), the National Research Foundation (NRF), and the Tertiary Education Trust Fund (TetFund) of Nigeria, etc.) or International organizations intervention with the NGOs model (such as Lacuna Fund, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), and The AI4D Africa Initiative among others). Either of the two sources of funding research comes with too many bureaucratic protocols that tend to limit the fund amount, scope, and direction of research for optimal research outputs.

We advocate for more funding from African government institutions, local stakeholders corporate organizations, foreign agencies, and governments with less bureaucracy to enable access for Sub-Saharan African NLP research communities.

6 Conclusion

This study surveyed more than 100 articles written between 2013 and 2023 containing works on Sub-Saharan African Language NLP. We have reviewed and extracted fewer than 70 relevant articles on NLP activities in languages that cut across all four regions of Sub-Saharan Africa. Based on these surveys, we established our position on the thought-provoking NLP research issues of Sub-Saharan African languages. Finally, we provide recommendations on ways to address issues in the current state of NLP in Sub-Saharan Africa.

7 Limitation

Our position comes from surveys of literature we conducted on NLP activities (research and publications) for Sub-Saharan African languages. While we have covered several major publications on sub-Saharan African languages, this is by no means exhaustive.

References

I. Abdulmumin and B. S. Galadanci. 2019. hauwe: Hausa words embedding for natural language processing. *In Proceedings of the 2019 2nd International Conference of the IEEE Nigeria Computer Chapter*.

I. Adebara and M. Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers:3814–3841.

D. I. Adelani, M. A. Hedderich, D. Zhu, E. Van den Berg, and D. Klakow. 2020. Distant supervision and noisy label learning for low resource named entity recognition: A study on hausa and yorùbá. *In Proceedings of the ICLR 2020 Workshop.*

D. I. Adelani, G. Neubig, S. Ruder, S. Rijhwani, M. Beukman, C. Palen-Michel, and et 'al. [link].

A. Afolabi, E. Omidiora, and T. Arulogun. 2013. Development of text to speech system for yoruba language. *2nd International Conference on Engineering and Technology Research.*

B. Akera, J. Mukiibi, L. S. Naggayi, C. Babirye, I. Owomugisha, S. Nsumba, and et'al. 2022. Machine translation for african languages: Community creation of datasets and models in uganda. *AfricaNLP workshop at ICLR 2022.*

J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow. 2022a. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. *arXiv:2204.06487v3.*

J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow. 2022b. Multilingual language model adaptive fine-tuning: A study on african languages. *AfricaNLP workshop at ICLR 2022.*

L. Augustinus and P. Dirix. 2013. The ipp effect in afrikaans: A corpus analysis. *In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).*

L. Augustinus, P. Dirix, D. Van Niekerk, I. Schuurman, V. Vandeghinste, F. Van Eynde, and G. Van Huyssteen. 2016. Afribooms: An online treebank for afrikaans. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).*

J. Awwalu, S. E. Abdullahi, and A. E. Evwiekpaefe. 2021. A corpus-based transformation-based learning for hausa text parts of speech tagging. *International Journal of Computing and Digital Systems*, 10.

M. Bashir, A. Rozaimee, and W. M. Wan Isa. 2015. A word stemming algorithm for hausa language. *IOSR Journal of Computer Engineering*, 17.

S. Brokensha, E. Kotzé, and B. Senekal. 2023. Machine learning for document classification in an archive of the national afrikaans literary museum and research centre. *Journal of the South African Society of Archivists*, 56.

Shikali C. S. and Mokhosi R. 2020. Enhancing african low-resource languages: Swahili data for language modeling. *Data in Brief*, 31.

C. Chukwuneke, I. Ezeani, P. Rayson, and M. El-Haj. 2022. gbobert models: Building and training transformer models for the igbo language. *n Proceedings of the Thirteenth Language Resources and Evaluation Conference.*

R Dale. 2010. Classical approaches to natural language processing. *In Indurkhya, F. J. Damerau (Eds.), Handbook of natural language processing, second edition(pp. 3-7). (Chapman Hall: CRC machine learning pattern recognition series).*

Jurafsky Daniel and James H. Martin. 2020. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 3rd edition draft.

B. F. P. Dossou and C. C. Emezue. 2021. Okwugbé: End-to-end speech recognition for fon and igbo. *In Proceedings of the African NLP Workshop, EACL 2021.*

Roald E. and Martin J. P. 2014. Developing text resources for ten south african languages. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).*

T. Eckart, S. Bosch, U. Quasthoff, E. Körner, D. Goldhahn, and S. Kaleschke. 2020. Usability and accessibility of bantu language dictionaries in the digital age: Mobile access in an open environment. *Proceedings of the First Workshops on Resources for African Indigenous Languages (RAIL).*

R. Eiselen and T. Gaustad. 2023. eep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct south african languages. *n Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023).*

Barack et 'al. 2022. Kencorpus: A kenyan language corpus of swahili, dholuo, and luhya for natural language processing tasks.

I. Ezeani, M. Hepple, I. Onyenwe, and E. Chioma. 2018. Igbo diacritic restoration using embedding models. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop.*

F. Gaim, W. Yang, and J. C. Park. 2022. Geezswitch: Language identification in typologically related low-resourced east african languages. *IREC 2022.*

T. Gaustad and R. Eiselen. 2022. Exploring afrikaans word embeddings with analogies and nearest neighbors. *Proceedings of the 3rd Workshop on Resources for African Indigenous Languages (RAIL)*, 4.

T. Gaustad and M. J. Puttkammer. 2022. Linguistically annotated dataset for four official south african languages with a conjunctive orthography: Isindebele, isixhosa, isizulu, and siswati. *Data in Brief*, 41.

M. Griesel and S. Bosch. 2020. Navigating challenges of multilingual resource development for under-resourced languages: The case of the african wordnet project. in proceedings of the first workshop on resources for african indigenous languages. *In Proceedings of the First Workshop on Resources for African Indigenous Languages (RAIL)*.

U. A. Ibrahim, M. M. Boukar, and M. A. Suleiman. 2022a. Development of hausa acoustic model for speech recognition. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 13.

U. A. Ibrahim, M. M. Boukar, and M. A. Suleiman. 2022b. Development of hausa dataset: A baseline for speech recognition. *Data in Brief*, 40.

Dibitso M. A., Owolawi P. A., and Ojo S. O. 2019. Part of speech tagging for setswana african language. *In 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages pp. 1–6.

Dibitso M.A., Owolawi P.A., and Ojo S.O. 2019. Context-driven corpus-based model for automatic text segmentation and part of speech tagging in setswana using opennlp tool. *In: Bella, G., Bouquet, P. (eds) Modeling and Using Context. CONTEXT 2019. Lecture Notes in Computer Science*, 11939.

K. R. Mabokela and T. Schlippe. 2022. A sentiment corpus for south african under-resourced languages in a multilingual context. *In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*.

M. J. Mahloane and S. Trausan-Matu. 2015. Metaphor annotation in sesotho text corpus: Towards the representation of resource-scarce languages in nlp. *In 2015 20th International Conference on Control Systems and Computer Science*.

G. Malema, B. Tebalo, B. Okgetheng, M. Motlhanka, and G. Rammidi. 2020. Complex setswana parts of speech tagging. *In Proceedings of the First Workshop on Resources for African Indigenous Languages (RAIL)*.

V. Marivate, T. Sefara, V. Chabalala, K. Makhaya, T. Mokgonyane, R. Mokoena, and et'al. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *In Proceedings of the First Workshop on Resources for African Indigenous Languages (RAIL)*.

G. Matthew. 2015. Using technology recycling to develop a named entity recogniser for afrikaans. *Southern African Linguistics and Applied Language Studies*, 33.

E. Mboning, D. Baleba, J. M. Bassahak, and O. Wandji. 2020. Building collaboration-based resources in endowed african languages: Case of ntealan dictionaries platform. *In Proceedings of the First Workshop on Resources for African Indigenous Languages (RAIL)*.

I. Mohammed and R. Prasad. 2023. Building lexicon-based sentiment analysis model for low-resource languages. *MethodsX*, 11.

I. Mohammed and R. Prasad. 2024. Lexicon dataset for the hausa language. *Data in Brief*, 53.

E. E. Moses, A. S. Aminu, and Murtala S. 2022. Towards massive parallel corpus creation for hausa-to-english machine translation. *Current Issues in Descriptive Linguistics and Digital Humanities*.

S. H. Muhammad, I. Abdulmumin, A. A. Ayele, N. Ousidhoum, D. I. Adelani, S. M. Yimam, and et'al. 2023. A twitter sentiment analysis benchmark for african languages.

Claire B.and Joyce N., Andrew K.and Ronald O., Jeremy T. F., and et'al. Jonathan M. 2022. Building text and speech datasets for low resourced languages: A case of languages in east africa. *In AfricaNLP workshop at ICLR*.

J. Nakatumba-Nabende, C. Babirye, P. Nabende, J. F. Tusubira, J. Mukiibi, E. P. Wairagala, and et'al. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: Experiences and lessons. *Applied AI Letters*, 5.

W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Kolawole, T. Fagbohungbe, and et'al. 2020. Participatory research for low-resourced machine translation: A case study in african languages.

E. Nyoni and B. A. Bassett. 2021. Low-resource neural machine translation for southern african languages.

E. Ogbuju and M. Onyesolu. 2019. Development of a general purpose sentiment lexicon for igbo language. *In Proceedings of the WiNLP Workshop ACL Conference 2019*.

K. Ogueji, Y. Zhu, and J. Lin. 2021. Small data? no problem! exploring the viability of pre-trained multilingual language models for low-resourced languages. *In Proceedings of the 1st Workshop on Multilingual Representation Learning*.

A. Oladipo, O. Ogundepo, K. Ogueji, and J. Lin. 2022. An exploration of vocabulary size and transfer effects in multilingual language models for african languages. *In AfricaNLP workshop at ICLR*.

I. E. Onyenwe, M. Hepple, and U. C. I. Ezeani. 2018. A basic language resource kit implementation for the igbonlp project. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17.

W. Oyewusi, O. Adekanmbi, I. Okoh, V. Onuigwe, M. Salami, O. Osakuade, S. Ibejih, and U. Musa. 2021. Comprehensive named entity recognition for 5 nigerian languages. *In Proceedings of the AfricaNLP Workshop, EACL 2021*.

J. Rajab. 2022. Effect of tokenisation strategies for low-resourced southern african languages. *In ICLR 2022 Workshop AfricaNLP*.

O. Rakhmanov and T. Schlippe. 2022. Sentiment analysis for hausa: Classifying students' comments. *In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*.

S. Ralethe. 2020. Adaptation of deep bidirectional transformers for afrikaans language. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*.

A. Rozaimee, M. Bashir, and W. Malini. 2017. Automatic hausa language text summarization based on feature extraction using naïve bayes model. *World Applied Sciences Journal*, 35.

M. A. Suleiman, M. M. Aliyu, and S. I. Zimit. 2019. Towards the development of hausa language corpus. *International Journal of Scientific Engineering Research*, 10.

A. A. Sèmiyou, J. O. R. Aoga, and M. A. Igue. 2013. Part-of-speech tagging of yoruba standard, language of niger-congo family. *Research Journal of Computer and Information Technology Sciences*, 1.

A. L. Tonja, O. Kolesnikova, A. Gelbukh, and J. Kalita. 2024. Parallel corpus for low-resource ethiopian languages. *In Proceedings of the Fifth Workshop on Resources for African Indigenous Languages (RAIL)*.

A. Tukur, K. Umar, and A. S. Muhammad. 2019. Tagging part of speech in hausa sentences. *In Proceedings of the 2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*.

M. M. Van Zaanen, G. Van Huyssteen, S. Aussems, C. Emmery, and R. Eiselen. 2014. The development of dutch and afrikaans language resources for compound boundary analysis. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*.

Andrew Z., Evan C., and Sandy R. 2021. Text normalization for low-resource languages of africa. *AfricaNLP2021*.

R. Y. Zakari, Z. K. Lawal, and I. Abdulmumin. 2021. A systematic literature review of hausa natural language processing. *International Journal of Computer and Information Technology*, 10.

A. Y. Zandam, F. M. Adam, and I. Inuwa-Dutse. 2023. Online threat detection in hausa language. *In Proceedings of the 2023 AfricaNLP Workshop at ICLR*.

A Appendix

NLP Activities Across Sub-Sahara African Regions.

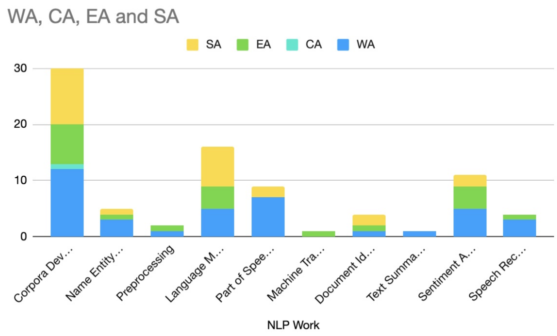

Figure 1: NLP Activities of Sub-Saharan Africa Across different NLP application areas.

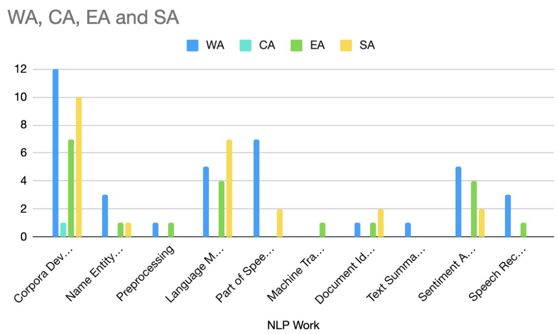

Figure 2: NLP Activities based on Regions in Sub-Saharan Africa. WA -West Africa, CA- Central Africa, EA - East Africa, and SA - South Africa regions.

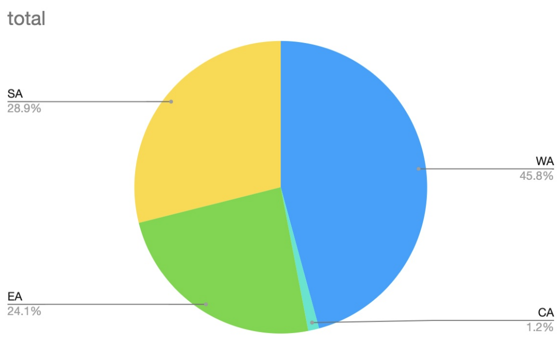

Figure 3: Percentage of NLP activities across Sub-Saharan African Regions

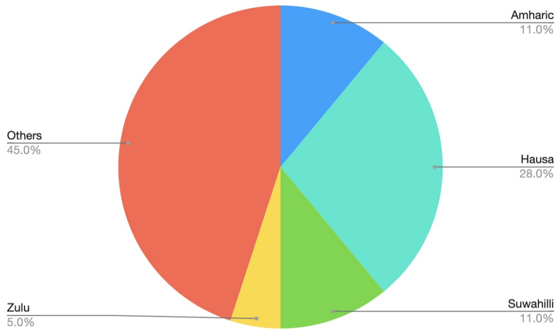

Figure 4: Percentage of NLP Activities based on the Language with the highest population in each region of Sub-Saharan Africa. Others - represent NLP activities for other languages in all four regions.

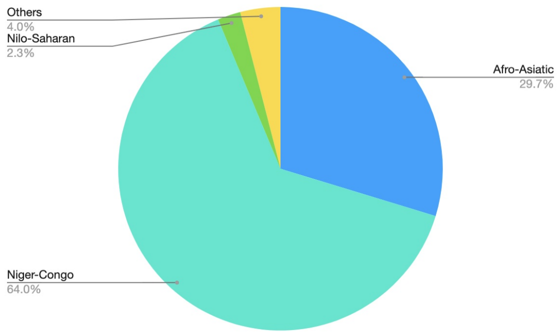

Figure 5: Percentage of NLP Activities of Sub-Saharan African Languages based on the large language family categories they belong to. Others - represent other language families with origins outside of the African Continent.

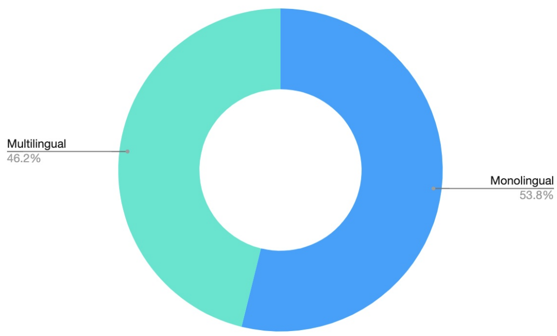

Figure 6: Percentage of NLP Activities based on single or multiple languages of Sub-Saharan Africa.

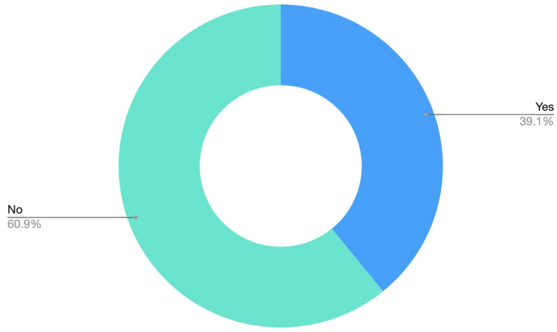

Figure 7: Percentage of NLP Activities with or without access to funding.

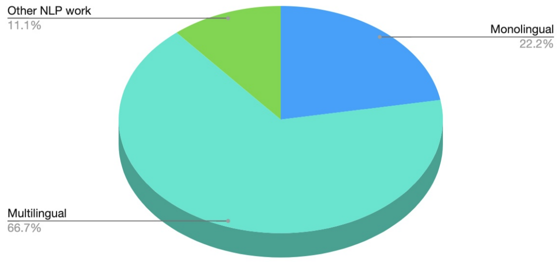

Figure 8: Percentage of NLP Activities based on single or multiple languages with access to funding. Other NLP work - represents other NLP activities in the form of a survey or review with access to funding.

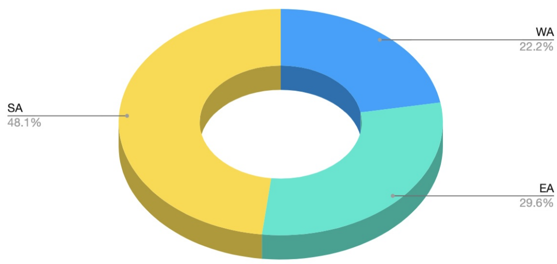

Figure 9: Percentage of NLP Activities based on regions in Sub-Saharan Africa with access to funding.