

Developing Natural Language Processing Tools for Nigerian Languages: Opportunities, Challenges, and Future Directions

¹Musa Jatau, ²Dorothy Aye Okpanachi, ³Andiyaryau Yohanna, ⁴Ali Abubakar Kadir

¹College of Education Zing, ²Kogi State College of Education, Ankpa,

⁴Taraba State University, Jalingo

DOI: <https://doi.org/10.60787/jolan.vol6no1.481>

Abstract

This study explores the feasibility of developing Natural Language Processing (NLP) tools for Nigerian languages using the low-resource NLP framework (Joshi et al.) and the Language Resource Management Framework (LRMF) (Ndubuisi & Marivate) as analytical lenses. It aims to identify existing linguistic resources, assess the challenges hindering computational development, and propose strategies for enhancing technological adaptation within Nigeria's multilingual context. Using descriptive and exploratory research designs, data were synthesized from secondary sources, including NLP repositories such as Masakhane, PanLex, AI4D Africa, and published linguistic corpora for selected Nigerian languages. The synthesized evidence indicates that resource limitations, orthographic inconsistencies, fragmented language datasets, and inadequate institutional support remain major constraints to NLP development. However, the reviewed studies consistently identify community-driven corpus development, multilingual transfer learning, and collaborative open-source initiatives as promising approaches for improving computational support for under-resourced Nigerian languages. The study therefore recommends increased government investment in corpus development, greater integration of computational linguistics into higher education, and strengthened collaboration among researchers, language communities, and technology developers. By synthesizing current evidence, the study contributes to ongoing scholarship on language technology in Nigeria and Africa while informing future efforts toward inclusive and culturally responsive AI.

Keywords: *Natural Language Processing, Nigerian Languages, Low-Resource NLP, Computational Linguistics, Digital Inclusion.*

1. Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that enables machines to understand, interpret, and generate human language. By integrating computational linguistics with machine learning and deep learning, NLP facilitates tasks such as machine translation, sentiment analysis, and speech recognition. Its applications are pervasive, enhancing digital communication, automating customer service, and supporting multilingual interactions in various sectors.

Globally, NLP advancements have predominantly focused on high-resource languages like English, Chinese, French, Arabic, and Spanish. These languages benefit from extensive digital corpora, standardized orthographies, and robust computational resources, leading to the development of sophisticated NLP tools. In contrast, many African languages, including those spoken in Nigeria, are classified as low-resource languages due to limited digital presence, lack of annotated corpora, and insufficient linguistic resources.

Nigeria, with over 500 indigenous languages, presents a complex linguistic landscape (Eberhard, Simons, & Fennig, 2023). While languages such as Hausa, Yoruba, and Igbo are widely spoken, they remain underrepresented in NLP research and applications (Adelani, Abbott, & Neubig, 2021). A recent survey indicated that only about 25% of studies on these languages contribute new linguistic resources, highlighting a reliance on repurposed data and underscoring the need for comprehensive resource development (Ezeani, Okegbemi, A., & Adeyemi, M 2023). Additionally, challenges such as tonal variations, diacritic dependencies, and dialectal diversity further complicate the development of effective NLP tools for Nigerian languages (Orife, Marivate, & Muhammad 2020; Adelani et al., 2022).

The study aims to explore the feasibility of developing NLP tools tailored for Nigerian languages, focusing on (i) identifying existing resources, (ii) assessing linguistic challenges, and (iii) proposing strategies to enhance computational support. By addressing these issues, it seeks to contribute to the digital inclusion of Nigerian languages, promote linguistic preservation, and support the development of AI technologies that are culturally and linguistically appropriate for Nigerian communities. The significance of this study lies in its potential to bridge the digital divide, ensuring that Nigerian languages are represented in the evolving digital and AI landscapes. By fostering the development of NLP tools for these languages.

2. Literature Review

2.1. Conceptual Review

Natural Language Processing (NLP)

Natural Language Processing (NLP) is an interdisciplinary field that integrates linguistics, computer science, and artificial intelligence to enable machines to understand, interpret, and generate human language meaningfully. Bird, Klein, and Loper (2023) define NLP as the “computational analysis and synthesis of human language through algorithms that process both its structure and semantics,” while

Jurafsky and Martin (2024) describe it as a bridge between human linguistic capability and machine computation, emphasizing its role in facilitating natural communication between humans and digital systems. Building on these perspectives, this study defines NLP as the systematic use of computational models to decode, represent, and simulate human linguistic behaviour across diverse communication contexts. NLP has evolved through three major paradigms: Rule-Based, Statistical, and Neural or Deep Learning-Based approaches. Rule-based NLP relies on predefined linguistic rules and grammars to parse and interpret text, exemplified by early systems such as ELIZA (1966) and SHRDLU (1972).

Bird, Klein, and Loper (2023) note that such systems are suitable for low-resource languages lacking annotated corpora, as observed in Yoruba morphological analyzers (Adewumi , Marivate & Akinola, 2023). The Statistical NLP paradigm, emerging in the 1990s, leverages large annotated datasets and probabilistic models such as Hidden Markov Models (HMMs), Naïve Bayes, and Conditional Random Fields (CRFs) for tasks like part-of-speech tagging and machine translation. According to Young, Hazarika,& Cambria (2022), this data-driven approach marked a significant shift toward empirical modeling, exemplified by English–French machine translation systems like IBM Model 3. The Neural or Deep Learning-based NLP era, driven by Recurrent Neural Networks (RNNs), Transformers, and Large Language Models (LLMs), have revolutionized the field by enabling context-sensitive and multilingual understanding. Jurafsky et al (2024) highlight systems like BERT, GPT-4, and Masakhane’s AfricanBERT as transformative innovations, while Orife et al., (2023) demonstrate their adaptation for African languages such as Hausa, Yoruba, and Amharic. Collectively, these paradigms illustrate NLP’s evolution from rule-based precision to data-driven and neural adaptability, supporting both global and Nigerian linguistic development.

Artificial Intelligence (AI)

Artificial Intelligence (AI) is a core branch of computer science dedicated to creating systems capable of performing tasks that traditionally require human intelligence, such as learning, reasoning, and problem-solving. Kaplan and Haenlein (2022) define AI as “a system’s ability to correctly interpret external data, learn from such data, and apply those learnings to achieve specific goals and tasks through flexible adaptation,” while Russell and Norvig (2023) describe it as the “design and study of intelligent agents that perceive their environment and act rationally to achieve desired outcomes.” For this study, AI is redefined as the computational simulation of human cognitive functions through adaptive algorithms that enable machines to perform linguistic, analytical, and

creative tasks autonomously. AI is broadly categorized into Narrow AI, General AI, and Superintelligent AI, each representing different levels of functionality and cognitive capability. Narrow AI (Weak AI) encompasses systems designed for specific, well-defined tasks such as image recognition, language translation, and speech processing - examples include Siri, ChatGPT, Grammarly, and Google Translate.

According to Russell, et al (2023), nearly all contemporary AI systems, including Natural Language Processing (NLP) models, belong to this class, especially in domains like Nigeria's fintech and education sectors where chatbots and automated tutors are employed. General AI (Strong AI) aspires to replicate human cognition, enabling reasoning, creativity, and adaptability across diverse contexts; however, Kaplan, et al (2022) note that it remains a theoretical construct, as no existing system has yet achieved human-level intelligence. Superintelligent AI, on the other hand, refers to advanced systems that surpass human intelligence in all domains, posing ethical and existential questions about control and accountability (Floridi & Cowls, 2023). These classifications reflect AI's evolution from narrow, task-specific automation to speculative forms of higher cognition. In the Nigerian context, AI is increasingly being adopted in fintech, education, healthcare, and language technology, as seen in Masakhane's multilingual translation models for Yoruba, Hausa, and Igbo, which demonstrate how localized AI research promotes linguistic inclusivity and cultural preservation. As such, AI serves not only as the foundation for NLP development but also as a transformative framework for understanding how computational intelligence can foster socio-cultural advancement in emerging digital economies.

Low-Resource Languages

Low-resource languages refer to linguistic systems with limited digital, textual, and computational data available for linguistic or technological processing. Such languages often lack essential resources like annotated corpora, comprehensive dictionaries, or digitized audio datasets required for Natural Language Processing (NLP) development. According to Joshi, Santy , & Budhiraja, (2020), a language is considered "low-resource when it lacks sufficient data to train or fine-tune computational models effectively," while Onose. Orife, & Muhammad, (2023) observe that African languages remain critically underrepresented in digital corpora, resulting in exclusion from AI-driven technologies. In the study, low-resource languages are defined as those with minimal structured datasets, inadequate computational tools, and insufficient representation in digital or institutional frameworks.

The classification of low-resource languages can be viewed from three interrelated dimensions: data availability, technological support, and institutional representation (Ndubuisi & Marivate, 2024). In terms of data availability, low-resource languages are those that lack adequate textual, audio, or parallel translation corpora necessary for computational model training. Nigerian languages such as Igala, Mumuye, Tiv, Kanuri, and Fulfulde exemplify this group, as they possess rich oral traditions but minimal digitized linguistic materials (Eberhard et al., 2023; Adelani et al., 2021; Nekoto et al., 2020). The technological support classification emphasizes the absence of software tools such as tokenizers, parsers, and morphological analyzers that enable NLP applications. For example, Yoruba has a developing textual corpus but still lacks sufficient open-source computational tools, whereas Hausa benefits from some speech recognition projects due to its broader regional presence (Onose et al., 2023). Lastly, institutional representation focuses on the extent of governmental and academic involvement in language digitization and resource development. Ndubuisi et al., (2024) argue that inadequate institutional investment in corpus linguistics and computational infrastructure perpetuates the marginalization of African languages.

Nigeria's linguistic landscape comprises languages with varying degrees of digital representation. While major languages such as Hausa, Igbo, and Yoruba have benefited from increasing NLP research and are supported by digital platforms such as Google Translate and language-specific Wikipedia editions, many other Nigerian languages including Igala, Mumuye, Tiv, and several minority languages remain severely under-resourced in terms of digitized corpora, annotated datasets, lexical databases, and NLP tools (Adelani et al., 2021; Nekoto et al., 2020). Although languages such as Kanuri and Fulfulde have recently received growing research attention, the quantity and quality of available linguistic resources remain insufficient compared with high-resource languages, limiting the development of robust AI and NLP applications (Adelani et al., 2022; Joshi et al., 2020). However, recent collaborative initiatives like the Masakhane Project and AI4D Africa are making significant progress by creating open-access corpora and translation models, fostering digital equity and ensuring that these languages gain visibility in global NLP research.

2.2. Empirical Review

Joshi et al. (2020) examined the global representation of languages in Natural Language Processing (NLP) and exposed the overwhelming dominance of English and a few high-resource languages, which they termed a “systemic bias” in computational linguistics. Their study aimed to quantify linguistic diversity and advocate inclusion

for underrepresented languages. Using large-scale analyses of ACL publications and NLP datasets, they found that most of the world's 7,000 languages are missing from computational research and benchmarks, with African languages being the most neglected. Grounded in the theory of linguistic equity, the study concluded that claims of language-agnostic models are empirically false, calling for community reorientation toward inclusive resource development. The authors recommended more collaborative data creation, open benchmarks, and capacity building for local researchers. However, the study's weakness lies in its macro-level approach; it lacks specific country or regional insights into linguistic structures, resource types, and institutional engagement. The present study addresses this gap by focusing on Nigerian low-resource languages, examining corpus availability, orthographic standardization, and academic participation to provide a grounded, context-specific understanding of digital linguistic marginalization.

Nekoto et al. (2020) (Masakhane Project) presented a participatory framework for developing machine translation (MT) resources for African languages through community-driven collaboration. The study aimed to demonstrate that sustainable NLP development is achievable when local linguists and native speakers are empowered to co-create datasets and models. Guided by participatory research theory and Afrocentric methodologies, the authors coordinated multilingual data collection, translation validation, and model training across numerous African communities. Findings revealed that community-led efforts significantly improved corpus quality, expanded linguistic coverage, and promoted ethical research practices in low-resource settings. It concluded that participatory NLP fosters ownership, inclusivity, and long-term sustainability. Nonetheless, the research overlooked detailed analysis of dialectal variations, orthographic harmonization, and long-term corpus maintenance. The present study builds on this limitation by localizing the participatory framework to Nigeria's multilingual environment, focusing on languages such as Yoruba, Fulfulde, Igala, Mumuye and Kanuri, while evaluating dialectal diversity, orthographic consistency, and sustainable data stewardship for future NLP applications.

Leong et al. (2023) investigated strategies to overcome the "low-resource double-bind," where languages suffer both data scarcity and computational limitations. Their study aimed to test whether low-compute techniques, such as adapter-based fine-tuning and lightweight model architectures, can yield efficient NLP performance for African languages. Using empirical experiments on limited datasets, the researchers demonstrated that low-compute methods achieve competitive accuracy compared to high-resource models, offering practical solutions for researchers in underfunded

institutions. The study is grounded in the theory of transfer learning and efficient modeling. It concluded that resource-constrained environments can still participate in NLP innovation through parameter-efficient approaches. The findings implied a promising direction for digital inclusion and accessible language technology development. However, its limitations include narrow language coverage and a lack of exploration into speech-based or morphologically complex languages. The present study addresses this weakness by applying low-compute modeling to Nigerian low-resource languages, particularly tonal and agglutinative systems like Yoruba, Igala, Mumuye and Fulfulde thereby expanding both linguistic diversity and methodological depth in low-resource NLP research.

Therefore, Joshi et al. (2020) investigated global language inequality in NLP, revealing English's dominance and the exclusion of African languages from computational research. Using linguistic equity theory, they called for inclusive data creation but failed to provide country-specific insights. This study addresses that gap by focusing on Nigerian low-resource languages and their digital representation. Similarly, Nekoto et al. (2020), through the Masakhane Project, demonstrated how community-driven collaboration can enhance machine translation for African languages, emphasizing participation and sustainability. However, their work lacked analysis of dialectal and orthographic challenges, which this study aims to explore within Nigeria's multilingual context. Finally, Leong et al. (2023) examined low-compute NLP methods, showing that efficient models can perform well even in data-scarce environments. Though impactful, their research excluded morphologically complex languages. The present study builds on these gaps by extending low-compute approaches to Nigerian languages such as Yoruba and Fulfulde, strengthening linguistic inclusivity and computational adaptability

3. Theoretical Framework

The study is anchored in the low-resource NLP framework articulated by Joshi et al. (2020), which examines the challenges of developing computational models for languages with limited annotated data, digital corpora, and linguistic resources. The framework highlights disparities in global language technology by categorizing languages according to the availability of linguistic resources and NLP support rather than proposing a formal theory. It emphasizes multilingual modelling, transfer learning, cross-lingual knowledge transfer, and community-driven resource development as practical strategies for improving language technologies for under-resourced languages, including many Nigerian languages.

Several scholars have adopted this framework to address linguistic inequality in NLP. Adebara and Abdul-Mageed (2022) utilized it to improve multilingual translation between Hausa and English through transfer learning and data augmentation. Orife et al. (2020) applied it in their work with the Masakhane Project, demonstrating collaborative model training for African languages using limited data. Nekoto et al. (2021) further extended the framework to create open-access corpora and translation datasets for over 30 African languages, showing the power of community-based approaches in low-resource settings.

The theory's relevance to this study lies in its alignment with the Nigerian linguistic context, where most indigenous languages lack digital representation. It guides the exploration of data-efficient methods and collaborative frameworks to develop NLP tools that promote inclusivity and linguistic preservation.

4. Methodology

The study adopted a qualitative and exploratory research design aimed at examining the feasibility of developing Natural Language Processing (NLP) tools for Nigerian languages. The qualitative approach facilitated an in-depth understanding of the linguistic, technological, and sociocultural factors influencing NLP development. Secondary data were obtained through a systematic review of publicly accessible NLP repositories and linguistic databases, purposively selected based on their relevance, accessibility, documentation quality, and inclusion of Nigerian or African language resources, including PanLex, Masakhane, and available language corpora. In addition, observations and consultations were conducted with computational linguists and local language experts to obtain insights into practical challenges and resource gaps. Data were analysed through a comparative examination of existing NLP tools, frameworks, and linguistic resources to assess their adaptability to Nigerian languages. Furthermore, the study evaluated the feasibility of developing such tools by analysing orthographic complexity, resource availability, community participation, and the technological infrastructure supporting language digitization in Nigeria.

5. Result and Discussion

This section presents and analyses evidence obtained from the systematically reviewed NLP repositories, linguistic databases, and published studies on Nigerian languages. The findings are organized according to the study objectives.

Evidence extracted from the reviewed repositories, including PanLex, Masakhane, AI4D Africa, and publicly available corpora, indicates considerable

variation in the availability of digital linguistic resources across Nigerian languages. The reviewed datasets show that Yoruba and Hausa possess relatively larger text and speech corpora, whereas languages such as Igala, Mumuye, Tiv, and several minority languages have substantially fewer annotated datasets and computational resources. Interpreted through the low-resource NLP framework of Joshi et al. (2020), these findings demonstrate that limited parallel corpora and labelled datasets constrain the development and evaluation of robust NLP models. The reviewed literature further revealed that existing language resources are dispersed across independent institutional projects, with limited interoperability and no centralized national repository. Evidence from the documented development of Swahili NLP (Onose et al., 2023) illustrates how coordinated resource aggregation has improved language technology. Collectively, these findings support the establishment of a national open linguistic resource framework through collaboration among government, academia, and open-source communities.

Evidence synthesized from the reviewed studies consistently identifies tonal complexity, morphological richness, orthographic variation, and dialect diversity as major computational challenges affecting Nigerian languages. The reviewed literature documents how tonal distinctions in Yoruba and Igala, diglossia in Hausa, and the agglutinative morphology of Fulfulde increase annotation complexity and reduce model performance. Using the Language Resource Management Framework (LRMF) (Ndubuisi, et al, 2024), these findings are interpreted as language resource management challenges requiring language-specific computational solutions rather than generalized modelling approaches. Furthermore, evidence reported by Masakhane (2024) views orthographic inconsistency with a substantial proportion of annotation errors in low-resource corpora. Across the reviewed studies, adaptive morphological segmentation, standardized orthographies, and tone-aware normalization were consistently identified as effective mitigation strategies. These findings underscore the importance of integrating linguistic expertise with computational methods during corpus development and model training.

Evidence synthesized from the reviewed repositories and empirical studies indicates that sustainable NLP development for Nigerian languages requires coordinated resource expansion, capacity building, and multilingual model adaptation. Drawing on the low-resource NLP framework (Joshi et al., 2020) and the LRMF (Ndubuisi, et al. , 2024), the reviewed studies consistently support community-based corpus development, transfer learning, and collaborative annotation as practical approaches for under-resourced languages. Published evidence from AfricanBERT and

AfriNLP demonstrates that multilingual language models achieve improved performance when adapted using locally annotated datasets and open-access resources. Similarly, findings reported in PanLex and Masakhane indicate that community participation contributes to increased corpus size and improved annotation quality. Based on the converging evidence across the reviewed studies, partnerships among universities, research institutes, and local language communities are essential for orthography standardization, dataset expansion, and the development of translation, speech recognition, and sentiment analysis systems suited to Nigeria's multilingual context.

5.1. Discussion

The findings synthesized from the systematically reviewed repositories and published studies indicate that the development of Natural Language Processing (NLP) tools for Nigerian languages is both feasible and necessary but constrained by structural, linguistic, and technological limitations. Evidence obtained from repositories such as Masakhane, PanLex, AI4D Africa, and related empirical studies shows that, although Yoruba, Hausa, and Igbo have comparatively more digital resources, the majority of Nigeria's indigenous languages remain inadequately represented in annotated corpora and computational datasets. The reviewed studies consistently attribute this gap to limited linguistic corpora, orthographic inconsistency, insufficient computational resources, and inadequate research investment. Evidence further demonstrates that tonal complexity, morphological diversity, and dialectal variation reduce the effectiveness of generalized NLP models, supporting the adoption of language-specific computational approaches. Interpreted through the low-resource NLP framework (Joshi et al., 2020) and the Language Resource Management Framework (Ndubuisi et al., 2024), the synthesized evidence indicates that community-driven corpus development, multilingual transfer learning, and open-source collaboration improve resource availability and model performance. Collectively, the reviewed evidence suggests that sustained institutional investment, coordinated linguistic resource development, and active participation of native language experts are essential for advancing inclusive NLP technologies and preserving Nigeria's linguistic diversity in the digital ecosystem.

6. Conclusion and Recommendations

The study concludes, based on evidence synthesized from the reviewed repositories and empirical literature, that developing NLP tools for Nigerian languages is

achievable through systematic resource development, technological adaptation, and collaborative engagement among linguists, computer scientists, policymakers, and local language communities. Although challenges such as limited annotated data, orthographic inconsistency, and inadequate institutional support persist, the findings interpreted through the low-resource NLP framework (Joshi et al., 2020) and the Language Resource Management Framework (Ndubuisi et al., 2024) indicate that community-driven corpus development, multilingual transfer learning, and open-source collaboration provide practical pathways for strengthening language technologies. The study further underscores the importance of integrating indigenous linguistic expertise with computational innovation to improve digital inclusivity and resource quality. Overall, the synthesized evidence suggests that sustained institutional investment and coordinated language resource management are essential for preserving Nigeria's linguistic diversity and enhancing its representation within emerging AI ecosystems.

The recommendations of this study are as follows:

- i. Government and academic institutions should fund corpus creation for Nigerian languages.
- ii. Linguists and computer scientists should collaborate to develop standardized orthographies.
- iii. Open-source platforms like Masakhane should be expanded to include more Nigerian languages.
- iv. Universities should introduce NLP and computational linguistics programs.
- v. Community participation should be encouraged in language data collection and annotation.
- vi. Local tech industries should invest in AI applications for indigenous languages.
- vii. Policies should promote digital inclusion and preservation of minority languages.
- viii. Researchers should develop language-specific NLP models for tonal and morphologically rich languages.

References

- Adebara, I., & Abdul-Mageed, M. (2022). Towards African Language NLP: Low-Resource Methods and Multilingual Models. *Transactions of the ACL*.
- Adelani, D. I., Abbott, J., Neubig, G., (2021). MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9 1116–1131.
- Adelani, D. I., Alabi, J. O., Kreutzer, J., et al. (2022). The AfroXLMR multilingual language model for African languages. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 11202–11226.
- Adewumi, T., Marivate, V., & Akinola, S. (2023). Low-Resource NLP for African Languages: Progress and Challenges. *Language Resources and Evaluation Journal*.
- Bird, S., Klein, E., & Loper, E. (2023). *Natural Language Processing with Python* (2nd ed.). MIT Press.
- Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2023). *Ethnologue: Languages of the World* (26th ed.). SIL International.
- Ezeani, I., Okegbemi, A., & Adeyemi, M. (2023). Resource creation and reuse in NLP research for Nigerian languages: A systematic review. *Journal of African Language Technology*, 5(2), 45–61.
- Floridi, L., & Cowls, J. (2023). The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. *AI & Society*, 38(1), 75–91.
- Joshi, P., Santy, S., Budhiraja, A., (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of ACL 2020*, 6282–6293.
- Jurafsky, D., & Martin, J. (2024). *Speech and Language Processing* (4th ed.). Prentice Hall.
- Kaplan, A., & Haenlein, M. (2022). “Rulers of the World, Unite! The Challenges and Opportunities of AI.” *Business Horizons*, 65(4), 457–466.

Leong C., Shandliya H., Dossou B.F.P., et al. (2023). Adopting to the Low- Resource Double- Bind: Investigating Low-Resource African Languages. Africa NLP workshop. [Httpss://doi.org/10:48550;19:25:43;UTC](https://doi.org/10.48550/19:25:43;UTC) – Accessed 07/07/2026

Ndubuisi, F., & Marivate, V. (2024). Building Corpora for African Low-Resource Languages: Challenges and New Directions. *Journal of African Language Technology*, 2(1),15-28.

Nekoto, W., Marivate, V., Matsila, T., et al. (2020). Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. *Findings of EMNLP 2020*, 2144–2160.

Nekoto, W., et al. (2021). Participatory Research for Low-Resourced Machine Translation: The Masakhane Case Study. *Computational Linguistics Journal*.

Onose, A., Orife, I., & Muhammad, S. (2023). Empowering African Languages through Open - Source NLP Tools. *Transactions on African Language Technologies*, 1(3), 44–60.

Orife, I., et al. (2020). Masakhane: Building for Africa in Africa. *Proceedings of EMNLP Workshop on NLP for African Languages*.

Orife, I., Kreutzer, J., & Martus, T. (2020). Challenges of developing NLP tools for tonal and diacritic-rich African languages. *Proceedings of the Workshop on African Language Technology (AfriLTA)*, 12–20.

Orife, I., Marivate, V., & Muhammad, S. (2023). Masakhane: Machine Translation for African Languages. *Transactions on African Language Technologies*, 1(2), 22–37.

Russell, S., & Norvig, P. (2023). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Young, T., Hazarika, D., & Cambria, E. (2022). “Recent Trends in Deep Learning for NLP.” *IEEE Computational Intelligence Magazine*, 17(2), 12–29.

¹Musa Jatau, ²Dorothy Aye Okpanachi, ³Andiyaryau Yohanna, ⁴Ali Abubakar Kadiri

¹Department of Mumuye Language, College of Education Zing, Taraba State

²Department of English Language, Kogi State College of Education, Ankpa , Kogi State
Department of Languages and Linguistics, Taraba State University, Jalingo, Taraba State
Jataumusa57@gmail.com; ayehhcj@yahoo.com; andiyaforsesus@rocketmail.com;
abubakarali943@gmail.com