Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 4(2) March 2026
DOI: https://dx.doi.org/10.4314/jobasr.v4i2.21

Preservation of Low Resource Languages through Natural Language Processing: Challenges, Opportunities, and the Case of Tiv

Jerome Aondongu Achir[1] & Matthew T Ogedengbe[2]
[1,2]Joseph Sarwuan Tarka University, Makurdi, Nigeria
*Corresponding Author Email: achir.jerome@uam.edu.ng

ABSTRACT
The research progress made in Natural Language Processing (NLP) still cannot hide the fact that low-resource languages are well underrepresented in computational research. Low-resource languages face underrepresentation of computational resources in natural language processing (NLP) tools. Natural Language Processing tools such as machine translation, morphological analysis, and speech recognition are computational systems designed to document and revitalize low-resource language and are seen as an opportunity to document and revitalize them. However, the successful application of NLP tools requires adequate digital resources and linguistic dataset. Current NLP for low-resource languages has been seen to lean heavily on methods such as transfer learning, multilingual pre-trained models and data augmentation but still has shortcomings stemming from the scarcity of data and linguistic resources as well, the existing models are not suited to the morphological and tonal nature of Tiv. This work undertakes a careful review of research on NLP in low-resource Nigerian languages to ascertain the key challenges associated with Tiv language as pertain NLP. Fifty-four (54) publications were selected through a search and filter procedure which employed inclusion/exclusion rules to filter results from large scholarly databases and other NLP dataset sources. The analysis of the result exposes the near none existence of NLP resources for Tiv language. This research work contributes to identifying challenges in NLP research for Tiv language through creating a consolidated overview to provide a context-specific analytical framework and a a roadmap for the development of Tiv NLP, which spans resource development, method adaptation, and application use.

Keywords:
Low-resource Languages,
Natural Language Processing,
Tiv language,
NLP strategies,
African languages,
Computational Linguistics

INTRODUCTION

Natural language processing (NLP) has rapidly evolved in artificial intelligence that allows humans to interact with computers and other digital systems in a more natural feel. NLP allows computers to process and generate written and spoken language, and to translate between languages (Kaur & Kumar, 2025). The benefits of NLP currently vary across the world's languages due to the fact that majority of the best NLP systems are trained on vast amounts of text from a small number of high-resource languages, such as English, Mandarin Chinese, and Spanish. This implies that low-resource languages (LRLs) are currently underrepresented in the digital world (Xiang et al., 2024; Yusufu et al., 2025).

The high under representation of LRLs from the digital space is a technical problem and also a unfair exclusion since languages with limited textual corpora, lexicons, and speech data tend to have limited NLP capabilities, which restrict access to automated tools that can aid in education, knowledge acquisition, and communication in the digital realm (Ye et al., 2025). Furthermore, speakers of these languages are also excluded from fully participating in the digital revolution (Ge et al., 2024; Li et al., 2024).

Studies have shown that multilingual language models like mBERT and XLM-R have enhanced cross-lingual capabilities, their effectiveness is limited for languages that have scarce data since these language models are trained on high-resource language data (Conneau et al., 2020; Devlin et al., 2019; Nyalang, 2025). In addition; lack of diverse and annotated data also makes it difficult to measure the development of NLP research for these languages (Maheshwari et al., 2025).

How to cite this article: Jerome A. A. & Matthew T O. (2026). Preservation of Low Resource Languages through Natural Language Processing: Challenges, Opportunities, and the Case of Tiv. *Journal of Basics and Applied Sciences Research*, *4*(2), 207-218. https://dx.doi.org/10.4314/jobasr.v4i2.21

In Africa and Nigeria in particular, there has also been progress in the case of larger languages such as Hausa, Yoruba, and Igbo, there is still a big problem of the lack of depth and breadth in the available resources in comparison to major languages of the world (Adelani, 2025; Muhammad et al., 2025). Also, there is a problem of fragmented, incorrectly, and insufficiently developed resources in the case of most African languages, which has impeded the development of critical NLP technologies such as automatic speech recognition (ASR), machine translation, and morphological analysis (Imam et al., 2025).

The case of Tiv language is an example of the general trends in the field of NLP as it relates to African languages. Tiv is an under-resourced language with minimal available resources in the form of digitized resources and computational tools (Ishima & Agaji, 2025). This has serious consequences for the general NLP research in the under-resourced languages, Tiv inclusive even as there are millions of speakers of these languages in their domains.

In Nigeria, despite the growing body of research on NLP local languages, existing studies have largely focused on the widely spoken languages such as Hausa, Yoruba, and Igbo, with little or no attention given to low resourced languages such as Tiv. Aside that, most research focuses on multilingual models without addressing language-specific linguistic differences such as tonal complexity and morphological variation. As such, there is no harmonised work that assesses the application of existing NLP techniques to Tiv language, thereby leaving a significant gap in the development of NLP products for the language.

This study therefore intends to carefully review existing NLP approaches for Nigerian languages, identify challenges relevant to the development of NLP for Tiv language, assess the possible application of current NLP techniques to the Tiv language, and propose a systematic roadmap for the adaptation of NLP research and applications for Tiv language.

The rest of research work is structured as follows: Section 2 reviews related literature on NLP Nigerian languages. Section 3 presents the methodology adopted in this study; Section 4 discusses the findings and analysis while Section 5 proposes a roadmap for Tiv NLP resources development, while Section 6 is the conclusion on the research paper and future research suggestions.

Low-Resource Languages and NLP

Low-resource languages are defined by computational linguistics as languages with scarce digital texts, linguistic resources, and computational tools. NLP applications such as machine translation, morphological analysis, POS tagging, question answering, and sentiment analysis depend heavily on data availability (Xiang et al., 2024). Scarce resources affect representation, functionality, and scalability of NLP systems (Ye et al., 2025).

This paper contributes in three key ways:

1. Synthesizes current NLP strategies for low-resource languages within the African context.
2. Provides one of the first structured computational analysis of Tiv linguistic features relevant to NLP.
3. Proposes a phased development roadmap for Tiv language technology.

Core Challenges in NLP for Low-Resource Languages

Data Scarcity and Annotation Bottlenecks

Data scarcity and lack of annotated corpora are major barriers for LRL NLP (Ge et al., 2024). Small datasets hinder even basic NLP tasks such as morphology extraction and syntactic parsing (Li et al., 2024).

Linguistic Complexity and Diversity

Many African LRLs have complex morphology, variation in tone, and many differences in dialect, which cannot be handled by standard NLP pipelines (Nyalang, 2025). Standard pipelines methods are mostly generic and often perform poorly without language-specific adaptations (Yusufu et al., 2025).

Evaluation and Benchmarking

Due to the absence of standardized benchmarks, it results to a complicated progress assessment (Maheshwari et al., 2025). Proper evaluation requires annotated datasets and defined metrics, which are largely missing for LRLs (Gao et al., 2025).

Literature Insights

Recent research provides a comprehensive overview of NLP applications for languages with limited resources. Studies emphasize that, even today, one of the major bottlenecks that researchers face is that there is still a scarcity of large, well-annotated datasets available (Xiang et al., 2024). They also emphasize that it is a time-consuming and expensive task to obtain annotated datasets, and without such datasets, data-driven approaches are unable to generalize to rare or scarce languages (Ye et al., 2025).

In a related systematic review, research analyses strategies to address data scarcity specifically in generative language modeling, drawing on studies that span multilingual training, data augmentation, and back-translation techniques (Yusufu et al., 2025). They show that even transformer-based models depend on creative usage of limited corpora and that consistent evaluation standards for annotation quality are lacking, further compounding the challenge of meaningful resource comparison (Ge et al., 2024).

Recent reviews discuss Nigerian languages that have low resource status, including Hausa, Igbo, and Yoruba, despite their wide use (Adelani, 2025; Muhammad et al., 2025). The review indicates that a small proportion of the studies present new linguistic resources, whereas the majority of the studies reuse existing corpora instead of creating annotated corpora from scratch (Imam et al., 2025).

Importantly, no peer-reviewed NLP studies that focus on Tiv were readily available. Linguistic documentation exists (Ishima & Agaji, 2025), but computational corpus development is absent, revealing a significant research gap.

MATERIALS AND METHODS

This study adopts a systematic literature review design to look at state-of-the art Natural Language Processing research into LRLs and to do an analytical case study on the Tiv language. Whilst a systematic review allows for the most comprehensive meta-analytic synthesis of evidence, a structured narrative review approach provides transparency in methodology, as well as analytical depth in terms of interpretation when the domain of inquiry is nascent or fragmented (Grant & Booth, 2009; Snyder, 2019). In the specific context of NLP for LRLs in the African landscape, this method provides for a critically synthesized overview of computational approaches, techniques, methodologies, and limitations of infrastructure for this field (Snyder, 2019). This review synthetically integrates the systemically obtained literature with qualitative thematic analysis in regard to challenges, resource issues, and strategies for adaptation in the realm of Tiv language technology development.

Search Strategy and Data Sources

The literature search was systematically designed and performed between January and February 2026 to ensure breadth across the various domains contributing to this research question: Scopus, Google Scholar, IEEE Xplore, ACL Anthology, SpringerLink, and ScienceDirect were used for comprehensive retrieval. These platforms were selected based on the richness and diversity of their content concerning computational linguistics, artificial intelligence, digital humanities and African language studies (Snyder, 2019).

Boolean operators and keyword combinations were designed to retrieve maximum recall, as well as achieve precision. Some of the most salient search queries were "low-resource languages" AND "natural language processing", "African languages" AND "NLP", "Nigerian languages" AND "computational linguistics", "Hausa NLP", "Yoruba NLP", "Igbo NLP", "Tiv language" AND "computational", "transfer learning" AND "low-resource", "multilingual models" AND "African languages". Keywords were augmented with backward and forward citation tracing, as recommended by systematic review authors (Grant & Booth, 2009; Snyder, 2019), and repetitive use of search terms.

Inclusion and Exclusion Criteria

To define a precise and rigorous analytical scope for the study, the following criteria were applied:

Inclusion Criteria

i. Publications from 2015 to 2025; these years capture state-of-the-art techniques leveraging modern transformer architectures (Devlin et al., 2019).
ii. Publications in peer-reviewed journals, conference proceedings or preprints in highly-regarded digital archives.
iii. Publications on developing NLP tools, computational approaches, resources, and/or evaluation mechanisms for LRLs (Xiang et al., 2024).
iv. Publications specifically on Nigerian languages including Hausa, Yoruba, Igbo and Tiv (Adelani, 2025)
v. Publications contributing empirical evidence, computational techniques, or infrastructure development for LRL NLP.

Exclusion Criteria

i. Pure linguistic theoretical papers lacking computational applications.
ii. Ii. Publications lacking technical details or rigorous methodology.
iii. Iii. Blogs, opinion pieces, and other forms of online content not subject to review.
iv. Iv. Duplicate entries.

These selection criteria were derived from accepted systematic review practice and standard-setting works (Grant and Booth, 2009; Snyder, 2019).

Study Selection

A total of 180 papers were initially retrieved through the search queries. Duplicate articles and preliminary articles were removed after title and abstract screening, resulting in a list of 94 papers for full-text review. Following the exclusion/inclusion criteria review, 54 papers were retained for the qualitative synthesis of the findings. These retained papers were thematically grouped and reviewed under four distinct categories which were used to structurally guide the analysis: Mitigation strategies for low-resource NLP, corpus and lexical resource development for low-resource languages, evaluation frameworks and benchmarks, and applications to Nigerian languages as listed in Table 1.

Table 1: Distribution of Selected Studies by Analytical Theme

Analytical Theme	Number of Studies (54)	Percentage (%)
Data Scarcity Mitigation Strategies	18	33.3
Corpus and Lexical Resource Development	14	25.9
Evaluation Frameworks and Benchmarking	10	18.5
Nigerian NLP Applications	12	22.2
Total	54	100

Table 1 shows that 18 resources (33.3%) focus on data scarcity mitigation strategies, which is the central central problem in low-resource NLP research (Ye et al., 2025). Furthermore, only 22.2% of the reviewed studies specifically address Nigerian languages, underscoring underrepresentation of these languages (Adelani, 2025).
Importantly, very few peer-reviewed studies were identified that specifically addressed computational modelling or corpus development for the Tiv language. This near absence buttresses the research gap already stated in this study and underscores the need for structured resource development for Tiv language (Ishima & Agaji, 2025).

Resource Classification Procedure

To provide a comparative analysis of Nigerian language resource availability, computational infrastructure for Hausa, Yoruba, Igbo, and Tiv was evaluated using publicly verifiable evidence. Resource status was determined through:

i. Peer-reviewed documentation
ii. Institutional repositories
iii. Open-source platforms (e.g., GitHub)
iv. Publicly accessible corpora and benchmark datasets

Languages were categorized using predefined labels ("Yes," "Limited," "Very Limited," "Not Identified," "None Publicly Available") based on publicly accessible documentation as of January 2026 as captured in Table 2 (Joshi et al., 2020).

Table 2: Comparative Resource Classifictionof Major Nigerian Languages and Tiv

Language	Parallel Corpus	POS Tagset	MT System	ASR	Public Dataset Availability
Hausa	Yes	Yes	Yes	Limited–Moderate	Yes
Yoruba	Yes	Yes	Limited–Yes	Limited	Yes
Igbo	Limited	Yes	Limited	Very Limited	Limited
Tiv	Very Limited	Not Identified	Very Limited	None Reported	None Publicly Available

The comparative classification of the resources as captured in Table 2 shows a clear digital divide between Nigerian languages (Adelani, 2025). Hausa and Yoruba have moderate to high levels of NLP resource availability (Muhammad et al., 2025). Igbo is in the middle with some structured linguistic resources like POS tagsets, but with a lack of parallel corpora, machine translation systems, and speech technologies. Tiv, however, is in a critical state with a lack of the most basic computational tools like standardized tagsets, parallel corpora, ASR systems, and publicly available datasets (Ishima & Agaji, 2025). This shows the need to digitize the Tiv language and create computational resources for the language.
This resources classification approach aligns with meta-analytical assessments of low-resource in NLP, which emphasize transparency in resource documentation, availability, and accessibility (Joshi et al., 2020).

By integrating linguistic analysis with computational strategy evaluation, the methodology bridges descriptive language documentation and applied NLP system design and directly informs the phased development roadmap proposed for Tiv.

NLP Strategies for Supporting Low Resource Languages

Transfer Learning and Multilingual Models

For languages like Tiv, which do not have significant parallel corpora, part of speech tagsets, or speech data, it is extremely challenging to develop and train natural language processing models. Transfer learning using multilingual pre-trained models such as XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019) offers a solution. This approach uses cross-lingual knowledge transfer to leverage the learned knowledge

from high-resource languages such as Hausa, Yoruba, and English for under-resourced languages (Conneau et al., 2020; Yusufu et al., 2025).

For Tiv, even small datasets in specific domains, for instance, the Tiv UKC Lexicon, can be leveraged for fine-tuning the multilingual models for downstream tasks. This strategy is effective in extending the frontiers of computational capabilities for Tiv without the need for large datasets. This is a practical and cost-effective approach for initiating natural language processing for Tiv (Ge et al., 2024; Li et al., 2024).

Corpus Creation and Community Engagement

Corpus creation is the backbone for the development of natural language processing for Tiv. Due to the lack of Tiv in the digital space in the form of texts and speech data, it is necessary to engage the community and Tiv speakers in the creation of corpora (Gao et al., 2025). This will involve the creation of written texts and speech data. Community engagement is crucial in the development of corpora for Tiv. This will not only promote the development of natural language processing for Tiv but will also encourage the community to take ownership of the development and preservation of the Tiv language (Masoka et al., 2025). For instance, the community can be engaged in the development of speech data for automatic speech recognition and translation. This will encourage the development of oral texts for Tiv.

Tool and Resource Development

For effective NLP development in Tiv language, there is a need to develop language-specific tools and resources that include but are not limited to Tokenizers, Morphological analyzers, Lexicons, and Speech corpora (Nyalang, 2025).

Even basic tools and resources can play an important role in promoting and advancing research and applications in the language. For instance, the development of a Tiv language morphological analyzer using collected lexical resources can be used as a basis for various applications and research, including but not limited to parts-of-speech tagging, machine translation, and text-to-speech systems. Initial research on basic tools and resources for other African languages has shown that investing in these tools and resources significantly contributes to and advances computational research in low-resource settings (Adelani, 2025; Imam et al., 2025).

Implications for This Research

All these strategies and implications justify the methodology used in the research. This research combines cross-lingual transfer, community-driven corpora creation, and basic tool development, which collectively address the critical issue of the lack of digital resources in the Tiv language. This research provides a framework for the digital preservation of the language and its applications, thus laying the foundation for the sustainable computational use of the language in future research and technology development.

NLP Research in the Nigerian Context

Overview of Nigerian LRL NLP

Nigeria's linguistic diversity of over 500 indigenous languages has attracted some NLP research attention, particularly for major languages like Hausa, Yoruba, and Igbo, which are also classified as low resource due to limited computational resources (Adelani, 2025). Systematic reviews of NLP advancements for Nigerian languages emphasize challenges such as limited annotated datasets and the need for region-specific resources (Muhammad et al., 2025). However, only a minority of studies contribute novel linguistic resources as many rely instead on repurposing existing data (Imam et al., 2025).

Research on Individual Languages

Work on sentiment analysis---such as language-adaptive fine-tuning for Hausa using AfriBERTa---illustrates how pretrained models can be adapted to a specific low-resource language (Muhammad et al., 2025). The overall sentiment analysis performance was significantly increased with fine-tuning on different corpora, which showed that even in low data scenarios, pre-trained multilingual models would have a good potential in such NLP tasks for Nigerian languages (Sani et al., 2025). Although there have been studies that further language technology for Nigerian languages, explicit peer-reviewed papers for Tiv has not been documented yet; indicating a substantial gap in the body of work.

RESULTS AND DISCUSSION

Tiv Language: Linguistic Features and Implications for NLP Development

Tiv has peculiar characteristics that makes processing difficult at the computational level. Based on the findings from the works analyzed above on low resource languages this section will delve into the features that make Tiv a complex language to be used in computational development. Hence, it is crucial to explore features of Tiv language.

Tonal System

Tiv is a tonal language. In tonal languages, contrasts in pitch can have lexical and grammatical significance. For example, these tonal contrasts may make otherwise segmentally identical words carry distinct meanings. Unfortunately, the use of tone marking in written Tiv is often omitted in casual texts (Ishima & Agaji, 2025). This aspect of written Tiv complicates computation in several ways: first, acoustic models in automatic speech recognition must encode tone, because the pitch

differences that make words semantically distinct may not be marked, causing a system to miss lexical distinctions. Second, text to speech systems must model tone to maintain understandability. For machine translation, it's likely that lexical ambiguity arising from lack of tonal information may be widespread, especially in contexts with low pragmatics. Transformer models that have only been trained on non-tonal corpora have underperformed in situations requiring tone modeling, as observed in studies of other low-resource tonal languages (Nyalang, 2025). Thus, tone aware acoustic modeling and preprocessing would be prerequisites to developing speech systems for Tiv.

Morphological Structure and Inflection

Tiv morphology is very rich with respect to noun marking and verbal constructions (Ishima & Agaji, 2025). Inflection and morphosyntactic agreement create multiple surface forms for each lexeme. For computational modeling, this causes a great amount of data sparsity, which is already a problem for morphologically rich low-resource languages (Yusufu et al., 2025). In data driven NLP, word level embeddings may be rendered less useful for morphological variations of the same word because they may be modeled as separate items. Without either morphological normalization or subword segmentations, limited corpora cannot achieve effective generalization. There has been previous work in low-resource settings showing rule based morphological parsers or subword techniques such as Byte Pair Encoding and SentencePiece are necessary to solve this sparsity (Ge et al., 2024). This is particularly problematic for Tiv, as no computational morphological lexicon exists (Ishima & Agaji, 2025).

Noun Inflection and Agreement Systems

Noun inflection and agreement patterns shape Tiv sentence structure and syntactic relations. Specifically, subject-verb agreement and noun-modifier agreement are crucial to tasks like POS tagging and dependency parsing. High-resource languages have robust syntactically annotated corpora which allows for evaluation within systems like Universal Dependencies. However, currently no syntactically annotated corpus of Tiv has been published (Ishima & Agaji, 2025). Lack of such a corpus prevents benchmark evaluation and both rule based and neural systems. It is the lack of syntactic annotated resources which "remains one of the greatest bottlenecks to progress" in low resource NLP research (Maheshwari et al., 2025; Xiang et al., 2024).

Orthographic Standardization and Variation

Although standardized orthography does exist for Tiv, practical use varies according to domain (e.g., religious texts versus casual texting versus web content). The mark of tone on lexical items may appear inconsistently, and dialect specific spellings may appear in written Tiv (Ishima & Agaji, 2025). Orthographic variation creates problems in terms of data fragmentation and sparseness. Normalized spellings will be necessary to create unified corpora of any kind of linguistic data. There is general evidence across low resource NLP that the quality of preprocessing is strongly correlated with model performance (Nyalang, 2025), making it critical for any attempt at Tiv NLP.

Dialectal Variation

There exists a number of Tiv dialects with differences in vocabulary, pronunciation and grammar (Ishima & Agaji, 2025). Models trained on an insufficient representation of Tiv's dialects risk internal digital marginalization, leaving a majority of speakers unrepresented in training data. The nature of speech technologies means that differences in accent and pronunciation will matter far more than they will for NLP systems; models trained on limited dialects will fail to generalize for the overall population. Ensuring equitable representation of dialects will require deliberate sampling in the training corpus design process.

Computational Challenges

There is a doubly low-resource scenario for Tiv, where its structural complexity is compounded with data limitations. Unlike some low-resource languages, which may be simpler than Tiv but still possess little data, Tiv has tonality, complex morphology, dialectal and orthographic variation within a severely limited digital corpus (Ishima & Agaji, 2025). These conditions result in four mutually compounding challenges:

1. Data Scarcity compounded with structural richness in morphology. Limited data resources have been seen to exacerbate sparsity in complex morphological systems (Ye et al., 2025).
2. Tone-Sensitive Semantics: Standard pretrained transformer models may inadequately encode tonal distinctions (Nyalang, 2025).
3. Absence of Benchmarks: No publicly available annotated datasets exist for POS tagging, parsing, or translation (Maheshwari et al., 2025).
4. Minimal Parallel Resources: The lack of Tiv--English aligned corpora constrains machine translation development (Ishima & Agaji, 2025).

A similar picture emerges from a comparative stance. Targetted efforts have been made in Hausa, Yoruba and Igbo such as building corpuses, benchmark, using pre-trained multilingual models. These efforts indicate that development of resources step by step alongside using transfer learning has led to improved results for NLP, and as such Tiv's problem is much more basic and needs to be tackled step-by-step.

Quantifying the Computational Resource Gap

This study evaluates reported corpus sizes and benchmark results for the three most resourced Nigerian languages; Hausa, Yoruba, and Igbo, and compares them with Tiv. Table 3 presents an approximate publicly reported resource scales compiled from major multilingual datasets and public repositories such as Masakhane Research Foundation, JW300, MENYO-20k, Common Voice, and GitHub.

Table 3. Comparative NLP Resource Availability for some selected Nigerian Languages

Metric	Hausa	Yoruba	Igbo	Tiv
Estimated Parallel Sentences	~300k+	~200k+	~150k+	<1k (fragmented)
Public POS / Morphological Datasets	Available	Available	Available	None documented
Public ASR Speech Hours	100+ hrs	80+ hrs	60+ hrs	0 hrs
Reported MT BLEU Range	18–28	15–25	10–18	No baseline

This shows that Tiv exhibits an approximately 100 to 300 times lower parallel corpus availability relative to Hausa and Yoruba.

This indicates that the Tiv language currently lacks a documented parallel corpus of meaningful scale, baseline machine translation evaluations, and POS or morphological annotation resources. Hence, this prevents Tiv from participating in multilingual benchmarks, being included in pretraining corpora for large language models, and receiving fair representation in cross-lingual evaluation studies.

While Hausa, Yoruba and Igbo have all improved in all NLP applications namely, MT, ASR and POS tagging, Tiv is absent from this field of research.

These computational issues directly contribute to the architecture of the proposed development roadmap. For example, the lack of annotated corpora and benchmark indicates the initial need for corpora creation and annotation, and morphological issues necessitate the creation of language-specific preprocessing tools and hybrid modeling strategies later.

A Phased NLP Development Roadmap for Tiv

The reviewed literature shows that Tiv NLP research is a relatively immature field with small corpora, scarce resources of lexicons and poor computational tools. All the studies surveyed point to issues of morphology, tones and dialectal varieties in Tiv language. Based on this, the presented road map systematically structures development into three phased approaches, which begins from the requirement of necessary resources, moves to the necessity of basic tools and ends in the implementation of real-world applications. This clearly shows how it fills the gap in previous studies and the framework takes into account sustainable generation of resources, development of tools and building of the community

Phase 1: Developing Essential Resources

The first phase centres on linguistic resources. It is proposed by the author that a parallel corpus between Tiv and English be developed using religious and academic books, and government documentation along with manually verified crow sourced translations (Gao et al., 2025). Development of about 50k - 100k aligned sentence pairs for this corpus would allow initial training of neural machine translation models. At the same time, a dedicated Tiv POS tagset needs to be designed based on the guidelines for universal dependencies framework (e.g.) by hand-labeling 5k - 10k Tiv sentences to achieve a baseline set of tagged data. The author also suggested for a digitization of existing lexical resources and a morphological rules system to be developed to achieve a computational lexicon for Tiv (Ishima & Agaji, 2025).

Phase 2: Fundamental NLP Tools

Once, baseline resources are created, an initial set of foundational tools for Tiv could be developed. The researcher suggested that a Tiv-specific tokenization tool considering prefix and suffix structures of the language be built. This would not only address sparsity issues by carrying out orthographic normalization of the text (Nyalang, 2025) but also build an initial tool to be used in subsequent tasks. Leveraging pretrained multilingual language models like mBERT or AfriBERTa can assist in achieving similar base results as obtained for Hausa in the POS tagging tasks (Muhammad et al., 2025). The researcher also proposed for hybrid rule-based and neural networks morphological analyzers as a suitable approach given low data constraints and such examples available from research on similar low-resourced languages (Yusufu et al., 2025).

Phase 3: Applications

Once a set of tools are developed, complex applications can be implemented. For Machine Translation, multilingual fine-tuned transformer models could be further fine-tuned using back-translation methods to maximize data availability (Ge et al., 2024). In the case of ASR, tone-awareness for Tiv is extremely important and the researcher recommended building tonally aware

models based on available pre-trained models like wav2vec fine-tuned on Tiv (Imam et al., 2025). In addition, educational applications may be developed to incorporate the toolkits produced for language learning, promoting its conservation among speakers (Masoka et al., 2025). This application is captured in figure 1.

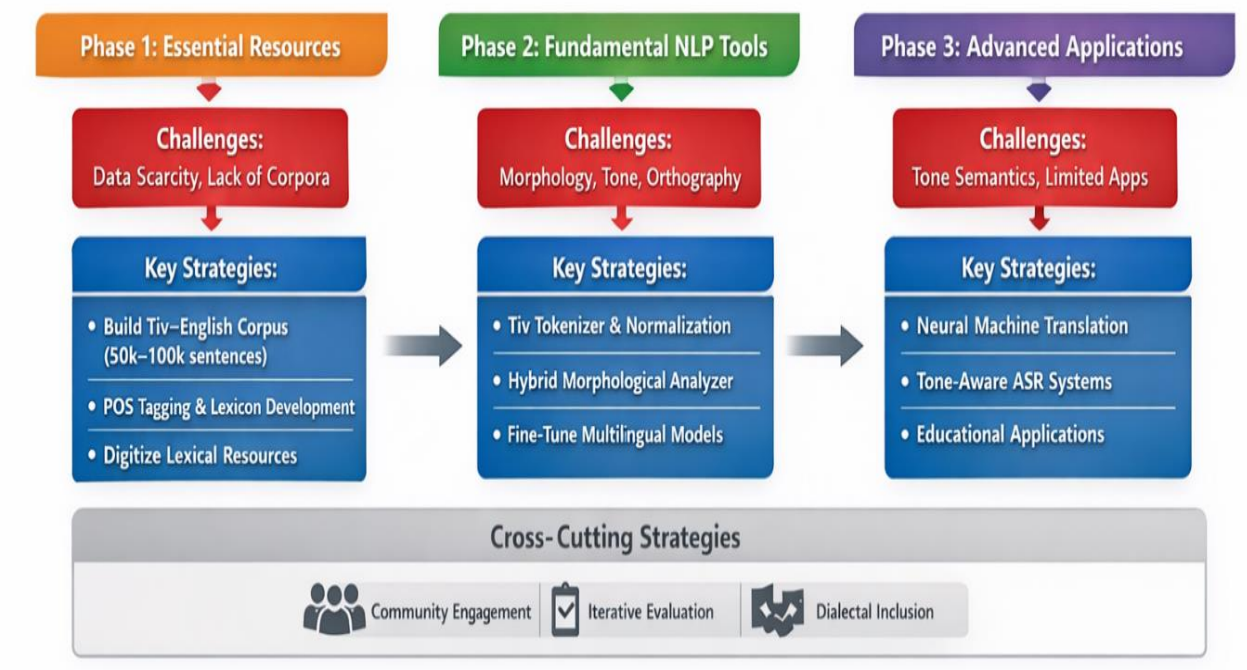

Figure 1: Tiv NLP development roadmap

Societal and Strategic Impact

These suggestions do not merely focus on technical gains but would also help in the preservation of the language, improving the quality of education for speakers and enabling access to digital platforms. Although various indigenous African languages like Hausa, Yoruba and Igbo have gained some visibility within the research community; Tiv's absence in these discussions indicate a missing piece in the overall field of NLP research on indigenous African languages, therefore such suggestions should endeavour towards building a bridge toward integrating Tiv into computational linguistics frameworks. In global pursuit of computational linguistics, Tiv is surely a significant inclusion for closing the digital language gap (Xiang et al., 2024).

It is important to treat Tiv not as an under-resourced language but as one with immense potential to make way for a new perspective on the inclusive use of computational linguistics (Gao et al., 2025).

Solutions outlined above are geared towards providing computational support for language conservation and inclusivity.

CONCLUSION

Natural language processing can play an important role in the preservation and revival of low-resource languages through like Tiv Language through the creation of tools for language maintenance thereby expanding access to digital resources and technology for these languages. While several indigenous African languages like Hausa, Yoruba and Igbo are gaining ground, current research trends lack empirical study regarding the NLP for LRLs like Tiv, indicating a gap in the available computational resources across these languages. Bridging this gap for Tiv language requires a joint effort between language experts and computational researchers to develop and enhance the language data development and tool deployment. The importance of addressing the technical limitations and requirements is of great significance in using NLP to conserve linguistic diversity and bridge the digital divide among speakers of Tiv language. The lack of mention of Tiv within computational linguists indicates a research void and therefore, developing a framework towards enhancing its datasets and enabling its computational power seems to be a reasonable approach. For effective and efficient development of resources for low-resource languages like Tiv, researchers should prioritize resource building, language-community involvement in data development and verification, model adaptation to cope with scarcity of data, and standardized benchmark evaluation framework for the task in Tiv.

REFERENCE

Adelani, D. I. (2025). Natural language processing for African languages [Conference presentation]. *Sixth Workshop on African Natural Language Processing*.

ACL Anthology. https://doi.org/10.18653/v1/2025.africanlp-1.0s

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., & Awokoya, A. (2020). MENYO-20k: A multi-domain English–Yorùbá corpus for machine translation [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.4297448

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Gao, Z., Yousufi, M., & Smith, J. (2025). Collective narrative grounding: Community-coordinated data contributions to improve local AI systems [arXiv preprint]. arXiv. https://arxiv.org/abs/2601.04201

Ge, L., Wang, H., & Zhang, Y. (2024). Cross-lingual knowledge transfer for low-resource languages. *Journal of Artificial Intelligence Research, 79*, 345–378. https://doi.org/10.1613/jair.1.15678

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Imam, S. H., Sani, B., Gete, D. K., Ahmed, B. Y., Ahmad, I. S., Abdulmumin, I., Yimam, S. M., Bello, M. Y., & Muhammad, S. H. (2025). Automatic speech recognition for African low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2025.africanlp-1.15

Ishima, J. L. J., & Agaji, S. C. (2025). Verb valency in Tiv: An X-bar approach. *Journal of African Languages and Linguistics, 46*(1), 112–134. https://doi.org/10.1515/jall-2025-0008

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Kaur, P., & Kumar, P. (2025). Bridging the gap: A survey of document retrieval techniques for high-resource and low-resource languages. *Computer Science Review, 57*, 100756. https://doi.org/10.1016/j.cosrev.2025.100756

Li, W., Zhang, M., & Chen, Q. (2024). Comprehension is greater than generation in low-resource languages: A comparative study of PLMs and LLMs. *Transactions of the Association for Computational Linguistics, 12*, 456–473. https://doi.org/10.1162/tacl_a_00645

Maheshwari, A., Singh, R., Kumar, S., & Gupta, V. (2025). IndicParam: Benchmark to evaluate LLMs on low-resource Indic languages [arXiv preprint]. arXiv. https://arxiv.org/abs/2512.00333

Masoka, H., Chikwava, T., & Nyoni, P. (2025). Advancing conversational AI with Shona slang: A dataset and hybrid model for digital inclusion [arXiv preprint]. arXiv. https://arxiv.org/abs/2509.14249

Muhammad, S. H., Abdulmumin, I., Yimam, S. M., Ahmad, I. S., & Sani, B. (2025). Yankari: A large-scale monolingual dataset for Yoruba language. In *Proceedings of the Sixth Workshop on African Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2025.africanlp-1.2

Nyalang, B. (2025). Why multilingual transformers fail for Khasi: A linguistic analysis of low-resource Austroasiatic AI gaps. In *Proceedings of the International Conference on Smart Systems and Social Management* (pp. 62–68). Atlantis Press. https://doi.org/10.2991/978-2-38476-533-1_6

Sani, B., Muhammad, S. H., & Jarvis, S. (2025). Language-adaptive fine-tuning for Hausa sentiment analysis using AfriBERTa. In *Proceedings of the Sixth Workshop on African Natural Language Processing* (pp. 52–63). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.africanlp-1.8

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Xiang, Y., Li, J., & Wang, L. (2024). Addressing the uneven distribution of language resources in multilingual NLP. *Computational Linguistics, 50*(2), 321–348. https://doi.org/10.1162/coli_a_00489

Yahaya, U., Iliyasu, U., and Sule S (2026). An optimization-driven artificial neural network framework with reinforcement learning for intelligent phishing email detection.. *JOURNAL OF BASICS AND APPLIED SCIENCES RESEARCH, 4*(1), 152-163. https://doi.org/10.4314/jobasr.v4i1.17

Ye, F., Chen, X., & Zhang, H. (2025). Data augmentation strategies for low-resource machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 33,* 112–128. https://doi.org/10.1109/TASLP.2024.3512345

Yusufu, M., Wang, L., & Liu, Y. (2025). Cross-lingual transfer learning for morphologically rich languages: Challenges and solutions. *Natural Language Engineering, 31*(1), 78–104. https://doi.org/10.1017/S1351324924000234

Appendix A: Accessed Nigerian Language NLP Resources

S/ No	Language	Resource Name	Type	Public URL	Citation
1	Hausa	NaijaSenti Twitter Sentiment Corpus	Text Corpus (Sentiment)	https://zenodo.org/record/6538055	Muhammad, S. H., Ahmad, I. S., Bello, B., et al. (2022). NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. DOI:10.5281/zenodo.6538055
2	Hausa	Hausa-Yoruba Sentence Pairs	Parallel Corpus	https://huggingface.co/datasets/michsethowusu/hausa-yoruba_sentence-pairs	michsethowusu (2025). Hausa-Yoruba Sentence-Pairs. HuggingFace
3	Hausa	MasakhaNER (Hausa subset)	NER Dataset	https://huggingface.co/datasets/masakhane/masakhaner	Adelani et al. (2021). MasakhaNER Dataset.
4	Hausa	HausaNLP HausaCorpus	Text Corpus	https://hausanlp.org/	HausaNLP Project (ongoing)
5	Hausa	NaijaVoices Speech Dataset (Hausa splits)	Speech Corpus (ASR)	https://huggingface.co/datasets/naijavoices/naijavoices-dataset	Emezue, C., & NaijaVoices Community (2025). The NaijaVoices Dataset.
6	Hausa	Nigerian Common Voice Dataset (Hausa)	Speech/ ASR Dataset	https://huggingface.co/datasets/benjaminogbonna/nigerian_common_voice_dataset	Benjamin Ogbonna (2025). Nigerian Common Voice Dataset.
7	Hausa	African Voices Speech Dataset (Hausa)	Speech Corpus (ASR/Voice)	https://africanvoices.io/	Data Science Nigeria (2025). African Voices: Multilingual Speech Dataset.
8	Hausa	MasakhaPOS (Hausa subset)	POS Tagset	https://lacunafund.org/datasets/language/index.html	Dione et al. (2023). MasakhaPOS.
9	Hausa	JW300 Parallel Corpus (Hausa–English)	Parallel Corpus	https://opus.nlpl.eu/JW300.php	Tiedemann (2012). JW300.
10	Yoruba	NaijaSenti Twitter Sentiment Corpus	Text Corpus	https://zenodo.org/record/6538055	Muhammad, S. H., Ahmad, I. S., Bello,

			(Sentiment)		B., et al. (2022). NaijaSenti… DOI:10.5281/zenodo.6538055
11	Yoruba	MasakhaNER (Yoruba subset)	NER Dataset	https://huggingface.co/datasets/masakhane/masakhaner	Adelani et al. (2021). MasakhaNER Dataset.
12	Yoruba	NaijaVoices Speech Dataset (Yoruba splits)	Speech Corpus (ASR)	https://huggingface.co/datasets/naijavoices/naijavoices-dataset	Emezue, C., & NaijaVoices Community (2025). The NaijaVoices Dataset.
13	Yoruba	Nigerian Common Voice Dataset (Yoruba)	Speech/ ASR Dataset	https://huggingface.co/datasets/benjaminogbonna/nigerian_common_voice_dataset	Benjamin Ogbonna (2025). Nigerian Common Voice Dataset.
14	Yoruba	African Voices Speech Dataset (Yoruba)	Speech Corpus (ASR/Voice)	https://africanvoices.io/	Data Science Nigeria (2025). African Voices.
15	Yoruba	MENYO 20k Yoruba–English Parallel Corpus	Parallel Corpus	https://huggingface.co/datasets/masakhane/menyo-20k	Adelani et al. (2021). MENYO 20k.
16	Yoruba	BibleTTS (Yoruba)	Speech Corpus (TTS)	https://www.openslr.org/151/	Meyer et al. (2022). BibleTTS.
17	Yoruba	MENYO 20k English–Yorùbá Parallel Corpus	Parallel Corpus	https://github.com/uds-lsv/menyo-20k_MT	Adelani, D., et al. (2021). MENYO 20k: A Multi domain English–Yorùbá Corpus for MT.
18	Yoruba	Yankari Monolingual Yoruba Dataset	Monolingual Text Corpus	https://arxiv.org/abs/2412.03334	Akpobi, M. (2024). Yankari: A Monolingual Yoruba Dataset.
19	Yoruba/Igbo	Yoruba–Igbo Parallel Corpus	Parallel Corpus	https://figshare.com/articles/dataset/Yoruba-Igbo_parallel_Corpus/28615031	Yoruba–Igbo Parallel Corpus (2025).
20	Igbo	NaijaSenti Twitter Sentiment Corpus	Text Corpus (Sentiment)	https://zenodo.org/record/6538055	Muhammad, S. H., Ahmad, I. S., Bello, B., et al. (2022). NaijaSenti… DOI:10.5281/zenodo.6538055
21	Igbo	BBC Igbo–Pidgin Gold Standard NLP Corpus	Text Corpus (Multi-task)	https://huggingface.co/datasets/Bytte-AI/BBC_Igbo-Pidgin_Gold-Standard_NLP_Corpus	Bytte AI (2026). BBC Igbo–Pidgin Gold Standard NLP Corpus.
22	Igbo	MasakhaNER (Igbo subset)	NER Dataset	https://huggingface.co/datasets/masakhane/masakhaner	Adelani et al. (2021). MasakhaNER Dataset.
23	Igbo	NaijaVoices Speech Dataset (Igbo)	Speech Corpus (ASR)	https://huggingface.co/datasets/naijavoices/naijavoices-dataset	Emezue, C., & NaijaVoices Community (2025). The NaijaVoices Dataset.

24	Igbo	Nigerian Common Voice Dataset (Igbo)	Speech/ ASR Dataset	https://huggingface.co/datasets/benjaminogbonna/nigerian_common_voice_dataset	Benjamin Ogbonna (2025). Nigerian Common Voice Dataset.
25	Igbo	African Voices Speech Dataset (Igbo)	Speech Corpus (ASR/Voice)	https://africanvoices.io/	Data Science Nigeria (2025). African Voices.
26	Igbo	MasakhaPOS (Igbo subset)	POS Tagset	https://lacunafund.org/datasets/language/index.html	Dione et al. (2023). MasakhaPOS.
27	Igbo	JW300 Parallel Corpus (Igbo–English)	Parallel Corpus	https://opus.nlpl.eu/JW300.php	Tiedemann (2012). JW300.
28	Igbo	African & Celtic ST Track Dataset	Speech Corpus (ASR)	https://huggingface.co/datasets/McGill-NLP/african_celt_dataset	McGill NLP (2026). African & Celtic ST Track Dataset.
29	Hausa/Yoruba/Igbo	NaijaVoices Speech Dataset	Speech Corpus (ASR/Text)	https://huggingface.co/datasets/naijavoices/naijavoices-dataset	Emezue, C., et al. (2025). NaijaVoices Dataset.
30	Hausa/Yoruba/Igbo	TaTa Multilingual Dataset	Table to Text Dataset	https://github.com/google-research/url-nlp	Gehrmann, S., et al. (2022). TaTa Dataset.
31	Hausa/Yoruba/Igbo	Offensive Language Dataset	Classification Corpus	https://arxiv.org/abs/2406.02169	Aliyu, S. M., et al. (2024). Multilingual Offensive Language Dataset.
32	Tiv	Tiv UKC Lexicon	Lexicon / Semantic Resource	https://datascientiafoundation.github.io/LiveLanguage/datasets/tiv-ukc-lexicon-/	DataScientia Foundation (2023). Tiv UKC Lexicon.