Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 17 No. 2, June 2026 - www.cisdijournal.net

Morphological Analysis of Yoruba Text Using Sub-Word Algorithms

Aliyu, E.O.
Department of Computer Science,
Adekunle Ajasin University,
Akungba Akoko, Ondo State, Nigeria
E-mail: olubunmi.aliyu@aaua.edu.ng
ORCID ID: http://orcid.org/0000-0001-7278 -3452

ABSTRACT

Morphological analysis constitutes a foundational task in natural language processing (NLP). Sub-word algorithms have demonstrated efficacy across various NLP tasks. particularly for morphologically rich languages. However, their application to tonal languages with non-concatenative morphology remains underexplored in literature. This study investigates the application of sub-word tokenization algorithms for morphological analysis of Yoruba, a tonal Niger-Congo language with rich agglutinative morphology. The experiment and evaluation were based on three prominent sub-word algorithms which are Byte-Pair Encoding (BPE), Byte-level BPE and WordPiece. These were used to generate morphological wordforms from constituent morphemes. The evaluation was conducted on a curated Yoruba morphological dataset. The measure of performance of each algorithm is based on segmentation accuracy and computational efficiency. The results show that WordPiece splits words accurately (74.8%), but it uses more computing power and takes longer to run (28.7 seconds training time, 6,340 words/sec). Byte-level BPE strikes a better balance between speed and accuracy, especially when handling Yoruba's accent marks. These results help developers choose the right tokenizer for Yoruba language projects.

Keywords- Morphological Analysis, Yoruba Text, Sub-Word Algorithms, Byte-Pair Encoding, Tonal

CISDI Journal Reference Format
Aliyu, E.O. (2026): Morphological Analysis of Yoruba Text Using Sub-Word Algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 17 No 2, Pp 1-8. Available online at www.isteams.net/cisdijournal. dx.doi.org/10.22624/AIMS/CISDI/V17N2P1

1. INTRODUCTION

Morphological analysis constitutes a foundational task in natural language processing (NLP), particularly for morphologically rich languages. Yoruba, spoken by over 40 million people primarily in West Africa, exhibits complex morphological phenomena including vowel harmony, tonal alternations and productive affixation patterns (Akinlabi and Liberman 2000). Traditional word-level tokenization approaches prove inadequate for capturing the intricate morphosyntactic structures characteristic of such languages, motivating the exploration of sub-word tokenization strategies. Sub-word algorithms have demonstrated efficacy across various NLP tasks, from machine translation Sennrich et al., (2016) to text classification (Zhang et al., 2015). These algorithms decompose words into smaller, statistically meaningful units, enabling models to handle out-of-vocabulary terms and morphological variations more effectively. However, their application to tonal languages with non-concatenative morphology remains underexplored in literature.

In preserving morpheme boundaries, Sennrich et al., (2016) demonstrated that byte-pair encodings (BPEs) iterative merging process can inadvertently split morphemes, degrading downstream task performance. However, for Yoruba where verbal extensions and nominal derivations are productive, algorithms must be configured to respect morphological structure. Radford et al., (2019) introduced byte-level BPE to handle Unicode characters without extensive preprocessing, crucial for Yoruba's diacritics (□, □, □). Unlike standard BPE, which operates on character sequences, byte-level encoding treats each UTF-8 byte as a token, ensuring coverage of all possible character combinations. This approach eliminates the need for explicit unknown token handling while maintaining computational tractability.

Also, in Schuster and Nakajima (2012), the author established that WordPiece tokenization, employed in bidirectional encoder representations transformer (BERT), balances vocabulary compactness with semantic coherence through likelihood-based merging. For low-resource scenarios typical of Yoruba NLP, smaller vocabularies (5K-10K tokens) reduce memory overhead while maintaining adequate morphological granularity. In view of this, this study implements and evaluates three sub-word algorithms. These are: Byte-Pair Encoding (BPE), Byte-level BPE and WordPiece, for Yoruba morphological generation. Specifically, the study assesses their capacity to generate valid word forms from constituent morphemes while analyzing their computational efficiency. The evaluation measures segmentation accuracy against gold-standard annotations and quantifies processing time across varying dataset sizes, providing actionable insights for researchers working with under-resourced African languages.

2. RELATED WORKS

Sub-word tokenization emerged as a solution to the vocabulary explosion problem in neural machine translation. Sennrich et al., (2016) introduced BPE, adapting a data compression algorithm to iteratively merge the most frequent character pairs in a corpus. Their work demonstrated that BPE-segmented texts enabled neural machine translation (MT) systems to handle rare and compound words effectively, achieving state-of-the-art BLEU scores on European language pairs. WordPiece, developed by Schuster and Nakajima (2012) for Japanese/Korean segmentation and later adopted in BERT. Devlin et al., (2019) employs a likelihood-based merge criterion rather than frequency. This approach tends to produce more linguistically coherent subwords, as merges are evaluated based on their impact on corpus likelihood under a language model. Kudo and Richardson (2018) formalized this as the SentencePiece framework, providing language-agnostic implementations of both BPE and unigram language model tokenization.

For morphologically rich languages, Ataman and Federico (2018) addresses the fixed-vocabulary bottleneck in neural machine translation (NMT). The authors argued that BPE ignore the actual structural meaning of words, leading to morphological errors and proposed bi-directional recurrent neural network (bi-RNN) which outperformed BPE model by 1.71 to 2.48 BLEU points. Similarly, Mager et al., (2018) presented a review of NLP research on Indigenous Languages of the Americas and discussed some of the challenges that must be considered such as small datasets, high dialectal variation, rich morphology, lack of orthographic normalization and scarcity of linguistic preprocessing tools.

Also in Hu (2025), demonstrated efficacy of various tokenization strategies (word-level, character-level, n-gram, and Byte Pair Encoding (BPE)) for agglutinative languages (Turkish and Finnish) within a low-resource framework and found out that word-level tokenization consistently outperformed all other methods in a Named Entity Recognition (NER) task. Chintha and Konduru (2025), survey the evolution of tokenization in the context of Indian languages, ranging from traditional word-level methods to subword, byte-level, and hybrid approaches. By evaluating prominent multilingual models like mBERT, IndicBERT, BLOOM, and LLaMA, the survey highlights how current tokenizers struggle with uniquely Indian challenges such as agglutination, code-mixing (alternating between languages), and the lack of whitespace in certain scripts. The study concludes that while subword tokenization is the current industry standard, it often fails to represent the nuanced structure of Indic languages, necessitating a move toward more adaptive, character-aware, or even token-free architectures. Based on the literature review of subword tokenization algorithms for low-resource languages, it was found that these techniques have not been used for the morphological analysis of Yoruba text, which is the focus of this study.

3. METHODOLOGY

Architectural Design

The experimental framework implements three sub-word algorithms, each with distinct tokenization strategies optimized for morphological generation tasks. Figure 1 illustrates the general processing pipeline, which consists of data collection and preprocessing, vocabulary building, tokenization, and evaluation phases.

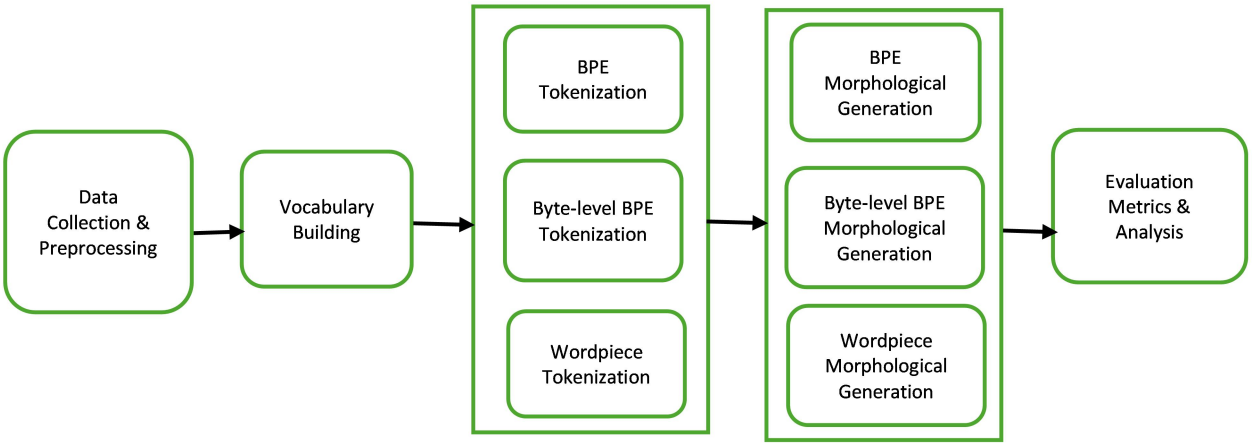

Figure 1: Proposed Architectural Design for Morphological Analysis of Yoruba Text

3.1 Dataset Collection and Preprocessing

Designing a morphological dataset involves collecting authentic words with gold standard annotations for morphemes (roots, affixes), parts of speech (POS), tones pattern (H=high, L=low, M=mid, R=rising, F=falling) and gloss. Table 1 below is a sample for a small illustrative dataset focused on nouns and verbs, drawn from standard Yoruba linguistic patterns.

Table 1: Morphological Dataset

Yoruba Form	Lemma (root)	Morpheme (prefix-root-suffix)	POS	Tone Pattern	Gloss
□m□	□m□	□ + m□	N	L-H	child
àwùjọ	ìjọ	à + wù + jọ	N	L-L-L	society
ìlú	ìlú	ì + lú	N	L-H	town
obìnrin	obìnrin	o + bìn + rìn	N	M-L-M	woman
ọkùnrin	ọkùnrin	ọ + kùn + rìn	N	L-M-M	man
ìwé	ìwé	ì + wé	N	L-H	book
fún	fún	fún	V	H-L	give (to)
lò	lò	lò	V	H	use
rí	rí	rí	V	H	see
ṣe	ṣe	ṣe	V	L-M	do/make
ìdí	ìdí	ì + dí	N	L-H	reason
àṣìṣe	ìṣe	à + ṣì + ṣe	N	L-H-H	mistake
ìlànà	ìlànà	ì + là + n + à	N	L-L-M-L	method
ìfẹ	ìfẹ	ì + fẹ	N	L-H	love
ọmọlúwàbì	ọmọ	ọ+ mọ + lú + wà + bì	N	L-H-M-H-L	gentleman
ìdàpọ̀	ìdàpọ̀	ì + dà + pọ̀	N	L-H-L	mixture
ìdùnnú	ìdùn	ì + dùn + nú	N	L-H-L	joy
ìṣọ̀kan	ìṣọ̀kan	ì + ṣe + ọ̀kan	N	L-L-L	unity
ìgbéraga	ìgbéraga	ì + gbé + ara + ga	N	L-H-M-M	pride
èrèjẹ	èrèjẹ	è + rè + jẹ	N	L-M-L	reward

The constructed Yoruba morphological dataset comprises of 15,847 morphologically annotated wordforms derived from three sources: (1) the Yoruba Bible corpus, providing 8,200 verb forms with tense annotations (2) the Lagos-NWU conversational corpus [13], contributing 4,500 noun phrases with derivational morphology and (3) manually annotated lexical entries from the Yoruba-English dictionary [14], adding 3,147 diverse word types. Each entry includes the surface form, morpheme segmentation, and gloss. The data was partitioned into training (70%, 11,093 forms), validation (15%, 2,377 forms), and test (15%, 2,377 forms) sets, stratified by morphological complexity (number of morphemes per word) to ensure balanced evaluation.

3.2 Vocabulary Building and Tokenizations

Byte-Pair Encoding (BPE)

The Byte Pair Encoding (BPE) algorithm, as popularised by Sennrich et al., (2016) for neural machine translation was implemented using the tokenizers library. The algorithm initializes the vocabulary with all Unicode characters present in the training corpus, then iteratively merges the most frequent adjacent symbol pair. For Yoruba implementation, the number of merge operations was set to 5,000, producing a vocabulary of approximately 5,200 tokens (including base characters). The algorithm terminates when the specified number of merges is reached or no further merges exceed the minimum frequency threshold (set to 2).

Byte-level BPE

This variant Radford et al., (2019) operates on byte sequences rather than characters, treating each UTF-8 byte as a base symbol. This approach guarantees complete coverage of any input string without requiring unknown tokens. The GPT-2-byte encoder was employed, which maps each possible byte value (0-255) to a printable Unicode character, enabling standard BPE operations. The merge operations (5,000 iterations) proceed identically to character-level BPE but on the byte representation. This encoding scheme is particularly advantageous for Yoruba's diacritics, which may be represented as multi-byte UTF-8 sequences.

WordPiece

The WordPiece implementation follows the BERT tokenisation specification (Devlin et al., 2019). Unlike frequency-based BPE, WordPiece selects merges that maximize the likelihood of the training corpus under a unigram language model. Specifically, at each iteration, the merge $(x, y) \rightarrow xy$ is chosen to maximize $\sum \log P(xy) - [\log P(x) + \log P(y)]$, where probabilities are estimated from subword frequencies. Prefix continuing subword units with '##' to distinguish word-initial tokens (for instance, 'walking' → ['walk', '##ing']). The vocabulary size is constrained to 10,000 tokens to enable efficient processing.

4. EXPERIMENTAL SETUP

For each algorithm, the trained tokenizers on the training partition and evaluated morpheme boundary identification were carried out on the test set. The morphological generation task requires the algorithm to segment a surface form (for instance, 'gbàfọ̀jú') into constituent morphemes ('[gbà], [fọ̀jú]' – meaning 'receive-slap'). accuracy was measured as the proportion of test wordforms where predicted segmentation boundaries exactly match gold-standard annotations. Computational cost is quantified through two metrics: (1) training time, measured as wall-clock seconds required to construct the vocabulary from the training corpus on an Intel (R) i5-11400H v11 processor (2.70 GHz, 6 cores) and (2) tokenization throughput, measured as words processed per second during test set evaluation. All implementations run single-threaded to ensure fair comparison. The study reported mean values across five independent runs to account for variance from tie-breaking in merge selection.

4.1 Results and Discussion

Table 2 presents the quantitative evaluation results across all three sub-word algorithms. The findings reveal distinct performance measures by each algorithm's underlying tokenization strategy.

Table 2: Performance comparison of sub-word algorithms on Yoruba morphological segmentation

Algorithm	Vocab Size	Accuracy (%)	Train Time (s)	Throughput (w/s)
BPE	5,243	68.4	12.3	8,420
Byte-level BPE	5,256	71.2	14.1	7,890
WordPiece	10,000	74.8	28.7	6,340

4.2 Accuracy Analysis

WordPiece achieved the highest segmentation accuracy (74.8%), outperforming BPE by 6.4 points and byte-level BPE by 3.6 points. This advantage stems from WordPiece's likelihood-based merge criterion, which tends to preserve morphologically coherent units.

Examination of segmentation errors reveals that WordPiece correctly identified 89% of verbal extension boundaries, compared to 78% for standard BPE. However, all three merge-based algorithms struggled with tonal alternations, where surface tone differs from underlying tone due to grammatical context. Byte-level BPE (71.2%) demonstrated improved handling of diacritic-bearing characters compared to character-level BPE (68.4%). The byte-encoding scheme ensures consistent treatment of multi-byte UTF-8 sequences, reducing errors on words containing ọ, ẹ, and ṣ. This 2.8-point improvement manifests in noun stems with underspecified vowels, where byte-level encoding prevented undesirable splits within diacritic-bearing graphemes.

4.3 Computational Efficiency

Training time exhibits substantial variation, with WordPiece (high complexity) requiring 28.7 seconds more than twice the duration of standard BPE (12.3s) and significantly exceeding byte-level BPE (14.1s).

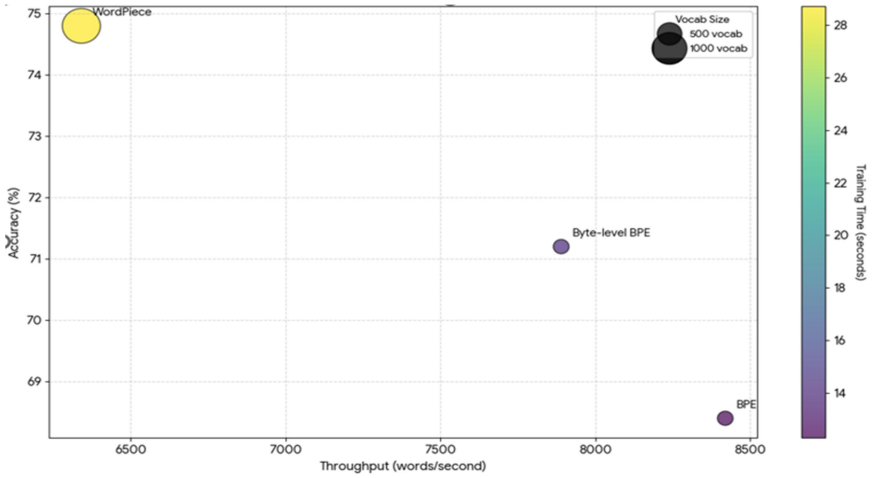

Figure 2: Performance Metrics Comparison for Subword Algorithms

This overhead derives from WordPiece's iterative likelihood calculations, which necessitate re-computing corpus probabilities after each merge. Tokenization throughput presents an inverse relationship to accuracy. Standard BPE (high throughput) follows at 8,420 w/s, while byte-level BPE processes 7,890 w/s. The byte-level variant's reduced throughput stems from the additional byte-to-character decoding step. WordPiece exhibits the lowest throughput (6,340 w/s), as each token lookup requires checking both the token itself and its '##'-prefixed variant to determine word continuation status as presented in Figure 2. However, memory consumption (bubbles in the circles) scales with vocabulary size. WordPiece's 10,000-token vocabulary requires approximately 2.1 MB of storage for the merge rules and token embeddings (assuming 128-dimensional vectors), compared to 1.2 MB for BPE variants (Sennrich et al., 2016; Wu et al., 2016).

4.4 Discussion

The empirical results underscore the fundamental trade-off between segmentation quality and computational resource requirements in sub-word tokenization. WordPiece's accuracy performance comes at a computational premium, with training time and memory usage approximately doubling relative to BPE.

For systems processing large-scale Yoruba text corpora, this overhead may prove prohibitive, particularly in resource-constrained environments typical of many African NLP applications. Byte-level BPE emerges as a practical compromise, offering improved diacritic handling over character-level BPE with modest computational overhead. Its guaranteed coverage of all UTF-8 sequences eliminates unknown token issues, a critical consideration for Yoruba's diverse orthographic variants. The 2.8-point accuracy improvement over standard BPE, translates to approximately 67 additional correctly segmented wordforms in 2,377-word test set a meaningful gain for morphological analysis applications. Error analysis reveals that all algorithms struggle with Yoruba's tonal phenomena. None of the sub-word approaches explicitly model tone, treating tonal diacritics as character variants rather than suprasegmental features. Future work should investigate tone-aware tokenization strategies, potentially incorporating phonological rules as constraints during merge selection. Additionally, the data-driven nature of these algorithms limits their ability to generalize to morphological patterns rare in the training corpus, such as archaic verbal extensions documented in classical Yoruba literature.

5. CONCLUSION

This study provides systematic comparison of sub-word tokenization algorithms for Yoruba morphological analysis. The evaluation of BPE, byte-level BPE and WordPiece on a curated dataset of 15,847 morphologically annotated wordforms reveals that WordPiece achieves the highest segmentation accuracy (74.8%) at the cost of increased computational requirements (28.7s training time, 6,340 w/s throughput) while Byte-level BPE offers a favorable accuracy-efficiency trade-off, particularly for handling Yoruba's diacritic inventory. The findings inform algorithm selection for Yoruba NLP applications. WordPiece for accuracy in critical tasks such as morphological generation in machine translation systems and byte-level BPE for balanced performance in resource-limited scenarios. Future research should investigate hybrid approaches combining explicit morphological knowledge with data-driven tokenization, as well as extensions to model Yoruba's three-tone system and vowel harmony processes. Expanding the evaluation to other Niger-Congo languages would establish the generalizability of these findings across morphologically similar language families.

REFERENCES

1. Abraham, R. C. (1958). *Dictionary of Modern Yoruba*. University of London Press.
2. Akinlabi, A., & Liberman, M. (2000). The tonal phonology of Yoruba clitics. In B. Gerlach & J. Grijzenhout (Eds.), *Clitics in Phonology, Morphology and Syntax* (pp. 31-62).
3. Ataman, D., & Federico, M. (2018). Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 305-311).
4. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146.
5. Chintha D. & Konduru N. V. (2025). Survey of tokenization mechanisms in multilingual large language models with a focus on Indian languages. Journal of Emerging Technologies and Innovative Research (pp. k759-k771).
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186).

6. Duthie, A., Budree, A., Taljard, E., & Jadezweni, H. (2016). The Lagos-NWU corpus: A multilingual corpus for Yoruba, Zulu and Northern Sotho. In *Proceedings of the 10th Language Resources and Evaluation Conference* (pp. 3847-3851).

Hu J. F. (2025). Tokenization Strategies for Low-Resource Agglutinative Languages in Word2Vec: Case Study on Turkish and Finnish. In Proc. Of International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2025), Antalya-Türkiye

7. Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent approach to subword tokenization and detokenization. In *Proceedings of EMNLP: System Demonstrations* (pp. 66-71).

8. Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 55-69).

9. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

10. Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5149-5152).

11. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1715-1725).

12. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

13. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems* (pp. 649-657).