International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V74I5P106

Volume 74 Issue 5, 50-61, May 2026

Original Article

Advancing Low-Resource African Language Technologies: Morphological Feature Integration for Kiswahili Question Answering

Collins S. Wanjala[1], Lilian Wanzare[2], Calvins Otieno[3]

Department of Computer Science, Maseno University, Kisumu, Kenya.

[1]*Corresponding Author : mscci00034019@student.maseno.ac.ke*

Received: 21 March 2026 Revised: 25 April 2026 Accepted: 16 May 2026 Published: 31 May 2026

Abstract *- Low-resource African languages remain critically underrepresented in natural language processing despite serving hundreds of millions of speakers across diverse linguistic communities. This paper addresses whether explicit morphological feature integration can overcome transformer limitations for Kiswahili, an agglutinative Bantu language spoken by over 100 million people across East and Central Africa. The agglutinative nature of Kiswahili presents fundamental challenges to subword tokenization algorithms that break the grammar and require implicit pattern learning using a small amount of data. The research tested vanilla XLM-RoBERTa on the KenSwQuAD question answering dataset, achieving 20.05% F1 and 17.80% Exact Match on the validation set. This weak baseline performance highlighted the significant limitations of traditional multilingual methods with morphologically complex low-resource languages. The study extended XLM-RoBERTa with explicit representations of 17 Kiswahili morphemes to build a morphologically-enhanced architecture, encoded as multi-hot vectors and introduced through learned projection layers and the pre-trained encoder being frozen to maintain multilingual knowledge. The optimized model's F1 and Exact Match scores of 72.40% and 62.91%, respectively, represented substantial improvements of 52.35 percentage points in F1 and 45.11 percentage points in Exact Match from baseline. Rigorous ablation studies demonstrated that improvements were due to the integration of morphological features, not to model capacity. This work demonstrates that explicit linguistic knowledge integration enables competitive performance even with severely limited training data, providing a reproducible framework for morphologically rich under-resourced African languages and challenging prevailing assumptions about the universal applicability of data-driven approaches.*

Keywords *- African Languages, Agglutinative Morphology, Kenswquad Dataset, Low-Resource Natural Language Processing, Question Answering Systems, Transformer Models.*

1. Introduction

Although morphologically rich African languages are spoken by hundreds of millions of people in economically and culturally important areas, they are grossly underrepresented in research, with remarkably good results obtained for high-resource languages such as English, Mandarin, and Spanish (Joshi et al., 2020). The question of genericity and equity of neural network architectures is fundamental. This structural mismatch between the principles of NLP models and the properties of the languages themselves, specifically the morphologically rich African languages, is not only a lack of data but is also a deeper issue that needs to be understood. Do transformer models work equally well across all languages, or do they have a hidden (and often implicit) assumption on language structure that systematically advantages languages with minimal or isolating morphology? The possible misalignments are very important not only in the advancement of African language technologies but also for the development of truly language-agnostic NLP architectures.

Kiswahili is a critical case study for the study of these questions, and has more than 100 million speakers spread across East and Central Africa, including Kenya, Tanzania, Uganda, the Democratic Republic of Congo, and Rwanda. Kiswahili has a special place in the African linguistic environment as it is an official language in several countries in Africa and a language of education, business, and administration, as well as inter-ethnic communication.

In spite of this socio-functional significance, complex question answering systems are yet to be developed for Kiswahili, leaving AI-based information retrieval systems inaccessible to Kiswahili speakers, which are already prevalent in high-resource languages. This technology divide extends information disparities and limits access to the

benefits of AI across linguistically diverse groups (Joshi et al., 2020).

The KenSwQuAD (Wanjawa et al., 2023) contains 7,526 manually curated question-answer pairs from Kiswahili Wikipedia on a wide range of topic areas, offering essential infrastructure for the development of Kiswahili language technology. In contrast, pre-trained models that are multilingual face significant difficulties with the agglutinative nature of the Kiswahili language, where the grammatical information is organized in a systematic manner in the form of the concatenation of bound morphemes to lexical units.

The structure is agglutinative, resulting in a high level of morphological productivity, with small numbers of morphemes being stored and a large number of inflected forms being produced, each carrying several different meanings at once. Also, the word hatutakula (we will not eat) consists of four bound morphemes: ha- (negative marker), -tu- (first person plural marker for the subject), -ta- (future tense marker), and -kula (verbal root: eat). Each morpheme carries a number of essential grammatical features (person, number, tense, aspect, and polarity) which must be broken down to grasp the meaning of the utterance fully. The specificity of the morphologically coherent units is lost in typical models of tokenization for standard transformers based on byte-pair encoding, where the structural morphology of meaning composition is lost in the process of forming tokens (Bostrom & Durrett, 2020).

Pre-trained multilingual models such as XLM-RoBERTa (Conneau et al., 2020) and mBERT (Devlin et al., 2019) have been shown to be highly capable of transferring learning across typologically different languages. These models use a masked language modeling task on large multilingual corpora to capture contextual representations and are capable of generalizing across languages. Nevertheless, they have to learn the morphology implicitly through subword tokenization strategies, which do not explicitly specify the morphology but rather the patterns of subword sequences.

This approach is implicitly used when the language is isolated, such as English or Chinese, where the word boundaries are roughly coincidental with semantic boundaries and the morphological complexity is not very high. In agglutinative languages such as Kiswahili, on the other hand, a subword fragmentation strategy leads to an inherent mismatch in representation, as the information that relational grammar encodes in a systematic concatenation of morphemes is broken up into statistically-derived subword units that are not necessarily aligned to the morpheme level (Bostrom & Durrett, 2020). The model should then be able to induce the morphological structure from the few training instances, which gets harder the more data is scarce and the more complex the morphology.

There have been recent attempts to build language-specific models, such as SwahBERT (Martin et al., 2022) and African language models such as AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022), targeting Kiswahili, which have reported better results on Kiswahili classification tasks due to targeted pre-training on African language Corpora. The improvements demonstrate the significance of language-specific optimization, but important questions remain unanswered: Are the improvements in terms of morphological representation, of the introduction of specifically Kiswahili vocabulary and phrases, or of domain adaptation to the use of African styles in text? It is not known what architectural requirements are needed to deal with morphological complexity, unless these factors are eliminated systematically. Moreover, pre-training in the target language incurs computational expenses and diminishes cross-lingual transfer potential, thus limiting it to the numerous African languages that are not adequately covered by their monolingual corpora for successful pre-training.

This work examines whether architectural inductive biases built to capture morphological structure can help reduce the data scarcity problem in low-resource African languages with complex morphology. The general and central assumption is that explicit integration of Morphological (MF) features offers computational shortcuts, allowing functional performance to be achieved even with a small training set.

The study makes three primary contributions: (1) empirical demonstration of weak baseline performance (20.05% F1 on validation) establishing the significant limitations of vanilla multilingual approaches for morphologically complex low-resource African languages; (2) introduction of a morphologically-aware architecture encoding 17 Kiswahili morphemes as multi-hot representations integrated with transformer hidden states through learned projections while preserving multilingual pre-trained knowledge; and (3) achievement of 72.40% F1 on KenSwQuAD, representing a substantial 52.35 percentage point improvement with rigorous ablation experiments isolating morphological contributions from confounding factors. In addition to the performance measures, this work offers theoretical insights into the interactions between linguistic typology and neural architecture design and highlights the computational necessity of explicit linguistic structure in morphologically complex low-resource languages, which conflicts with current claims of universal applicability of data-driven approaches.

2. Related Work and Benchmarking

Research in NLP for African languages has been rapidly expanding, and many of these languages have yet to receive a dataset and a benchmark for the study of NLP. The Masakhane community and associated work created MasakhaNER, the first high-quality, large-scale, multilingual named-entity-recognition dataset for ten African languages, and

demonstrated that current neural models, like XLM-R, typically outperform previous work on neural baselines, but with significant room for improvement (Adelani et al., 2021). In order to advance this work, MasakhaNER 2.0 expanded to twenty African languages and showed that the selection of transfer language significantly influences zero-shot performance, with Africa-centric transfer achieving an average of about fourteen F1 points over English-centric transfer (Adelani et al., 2022). The results in this study highlight two directly relevant issues to the work presented here: (1) multilingual transformers are the baseline for many African language tasks, and (2) their performance is selective and can be influenced by the linguistic match between the model and the task.

There are three families of approaches to modeling. The first category is general multilingual models like mBERT (Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2020) that have a wide coverage but implement morphology only implicitly as subword tokenization. The second is a group of models that have been pre-trained and/or adapted for African languages. With only about 1 gigabyte of text, AfriBERTa achieved comparable or even better performance than mBERT and XLM-R on NER and text classification tasks and was competitive with XLM-R on other tasks, suggesting that while care is needed in pre-training with comparable low-resource languages, it can be similarly competitive in some domains, and even match or surpass mBERT and XLM-R in others.

AfroXLMR fine-tuned XLM-R using multilingual adaptive fine-tuning for seventeen African languages, improving cross-lingual transfer with smaller models (Alabi et al., 2022). A model was adapted by SwahBERT specifically for Kiswahili and reported improvement for classification tasks on Kiswahili (Martin et al., 2022). Morphology is discussed explicitly in the third family, morphology-aware segmentation. Results for agglutinative languages (e.g., Turkish, Finnish, Kazakh) show that the subword method often ignores morpheme boundaries and that with morphology-aware segmentation, performance is recovered, particularly in low-resource settings (Bostrom & Durrett, 2020).

Benchmarking is still limited for African languages for question answering, in particular. TyDi QA is an information-seeking QA benchmark in eleven typologically diverse languages, and has demonstrated that the typological diversity is a real challenge for current information-seeking QA systems (Clark et al., 2020). African QA developed the first open-retrieval QA dataset in ten African languages and reported that automatic translation and multi-lingual-retrieval pipelines do not work well for these languages, and that African languages are among the most realistic and challenging use cases for cross-lingual QA (Ogundepo et al., 2023). There is an extractive QA dataset for Kiswahili, which was manually gathered by KenSwQuAD, with 7,526 question/answer pairs (Wanjawa et al., 2023). The present study stands out in a particular manner from previous studies in this context. Unlike SwahBERT, AfroXLMR, and AfriBERTa, which pre-train or adapt a new encoder, it preserves a standard multilingual encoder and only attaches an explicit and linguistically motivated morpheme representation. This design allows for the controlled ablation of the morphological structure, which is beyond the scope of representation-based and vocabulary-based methods, as the three are intertwined in the pre-training process. The context for benchmarking puts the work in context but leaves room for it to complement language-specific pre-training: it clarifies the feature that is responsible for the gains, namely explicit morphology.

3. Materials and Methods

3.1. KenSwQuAD Dataset

KenSwQuAD (Wanjawa et. al., 2023) is composed of 7,526 question-answer pairs extracted from a variety of topical domains in Kiswahili Wikipedia articles that span history, geography, science, culture, politics, and current affairs. Like SQuAD (Rajpurkar et al., 2016), each example consists of a context paragraph, a natural language question, and an answer span in the context paragraph. The extractive formulation must be set correctly to enable the model to pick out specific text spans instead of providing free-form answers, allowing evaluation with exact match and token-level F1 (Rajpurkar et al., 2016). The dataset has a number of properties that make it very difficult for conventional transformer models.

For one, the context length is an average of 156.3 tokens (std 82.7), making for a large search space for identifying the answer spans. Second, morphological variation implies that the same semantic content can be expressed in different inflected forms, and that models must be able to recognize morphological relationships, as opposed to just string similarities. Third, the size of the dataset is relatively small (only 7,526 examples before cleaning) when compared with the hundreds of thousands of examples found in high-resource languages such as English, further exacerbating the problem of sample efficiency in morphologically rich languages.

3.2. Data Preprocessing

To improve data quality and consistency for representation in text, the research applied a comprehensive preprocessing pipeline: (1) Unicode normalization – using the NFC (Normalization Form Canonical Composition) standard, malformed answer strings in the same dataset may be a result of annotation errors, or generated by processing artifacts, specifically - the example was removed from the dataset if the diacritics were not properly aligned with the characters, and they were combined systematically throughout the dataset; (2) systematic lowercasing – reduce vocabulary sparsity, in low-resource settings the other case variations might fragment the already limited training signal, specifically - the example was processed as lower case representation throughout the dataset;

(3) whitespace normalization – standardize the space between words, multiple consecutive newlines, tabs and spaces were removed; (4) answer validation – ensure that answer strings appear exactly in the context paragraph normalized as a text, malformed answer strings may be a result of malformed annotations, or processing artifacts - specifically, the example was removed from the dataset if the answer string appeared in a different paragraph than the context string, or appeared at a different position within the paragraph. This preprocessing excluded around 2,303 problematic pairs (30.6% of the data), leaving 5,223 pairs of good quality data. The validation experiments showed that preprocessing significantly affects the performance of the baseline XLM-RoBERTa: Using data quality improved the baseline XLM-RoBERTa performance by 14.71 percentage points on the validation set.

The result shows the importance of data quality in low-resource scenarios, where a single training sample can make a big difference to learned representations and data inefficiencies can accumulate rapidly. The dataset was split into training (80%, 4,178 examples), validation (10%, 522 examples), and test (10%, 523 examples) sets with a fixed random seed of 42 for reproducibility and direct performance comparison of the model configurations. All preprocessing was done in Python, leveraging the Hugging Face Datasets library, and the seed set to a fixed value (42) was used for shuffling and splitting, ensuring the training, validation, and test portions were reproducible across runs.

3.3. Baseline Model Configuration

To establish empirical baselines quantifying the adequacy of standard approaches for Kiswahili question answering, the study fine-tuned XLM-RoBERTa base (278 million parameters) on the preprocessed KenSwQuAD training set using standard hyperparameters widely adopted in question answering research: learning rate 2×10^{-5} with linear decay, batch size 8, 4 training epochs, weight decay 0.01 for regularization, and 500 warmup steps for stable training initialization. Inputs were tokenized with the XLM-RoBERTa SentencePiece tokenizer at a maximum sequence length of 384 with a document stride of 128 for long contexts, and the Adam optimizer was used; these settings were held constant across all subsequent versions so that performance differences could be attributed to the morphological component rather than to tokenization or optimizer changes.

The decision to use XLM-RoBERTa as the foundation model was made because of the fact that it has been pre-trained on 100 languages, which can further help to improve the performance on other under-resourced African languages in the future. The current selection focuses on the generalizability of the results and the transfer potential across languages in Africa, which was a strategic choice in the research, as the main goal was to build architectures that can be applicable to a wide range of language families in Africa, not optimized for Kiswahili. In addition, a widely used multilingual baseline allows for direct comparisons to previous works and provides a clear set of performance floors to assess the morphological improvements. This baseline also reflects the standard configuration used in prior multilingual evaluations on African languages, where XLM-R and mBERT are the common reference models (Adelani et al., 2021; Alabi et al., 2022).

3.4. Morphological Feature Set

The study identified a feature set that is linguistically motivated, capturing 17 important morphemes from Kiswahili using a descriptive grammatical analysis of Bantu verbal morphology. The feature set is based on those morphemes that are of high frequency and grammatically productive and that systematically encode core semantic and pragmatic distinctions.

There are six markers for subject agreement (ni- first person singular, u- second person singular, a- third person singular, tu- first person plural, m- second person plural, wa- third person plural), which convey the distinctions of grammatical person and number found in the argument structure; four for tense and aspect (na- present continuous, li- simple past, ta- future, me- perfect aspect) that express temporal and aspectual relations that are important for the understanding of event structure; four for relative clause constructions (-ye-, -a-); one for conveying successive actions (-ka-); and one for marking polarity (ha-) and three for object reference (-ni-, -ku-, -m-). Morpheme presence was detected by a rule-based matcher that scans each whitespace-delimited word for these affixes according to their canonical position in the Bantu verbal template, since byte-pair encoding does not by itself preserve morpheme boundaries (Bostrom & Durrett, 2020). Morphemes present in a word were encoded as a binary multi-hot vector, as defined in Equation (1), where each dimension corresponds to one morpheme and active morphemes are marked with 1, absent morphemes with 0.

$$m_i \in \{0,1\}^{17}, m_i[j] = \begin{cases} 1, if\ morpheme\ j\ is\ present\ in\ word\ i \\ 0, otherwise \end{cases} \quad (1)$$

This representation compresses morphological information into a compact 17-dimensional feature space, which is straightforward to compute and also explicitly captures the morphological structure, while being simultaneously small enough to be used as input to transformer hidden states.

Figure 1 demonstrates this morphological feature extraction process using the Kiswahili word hatutakula (we will not eat), showing how the four morphemes (ha, tu, ta, kula) are identified and encoded as a 17-dimensional binary vector where active morphemes are marked with 1s at their corresponding positions.

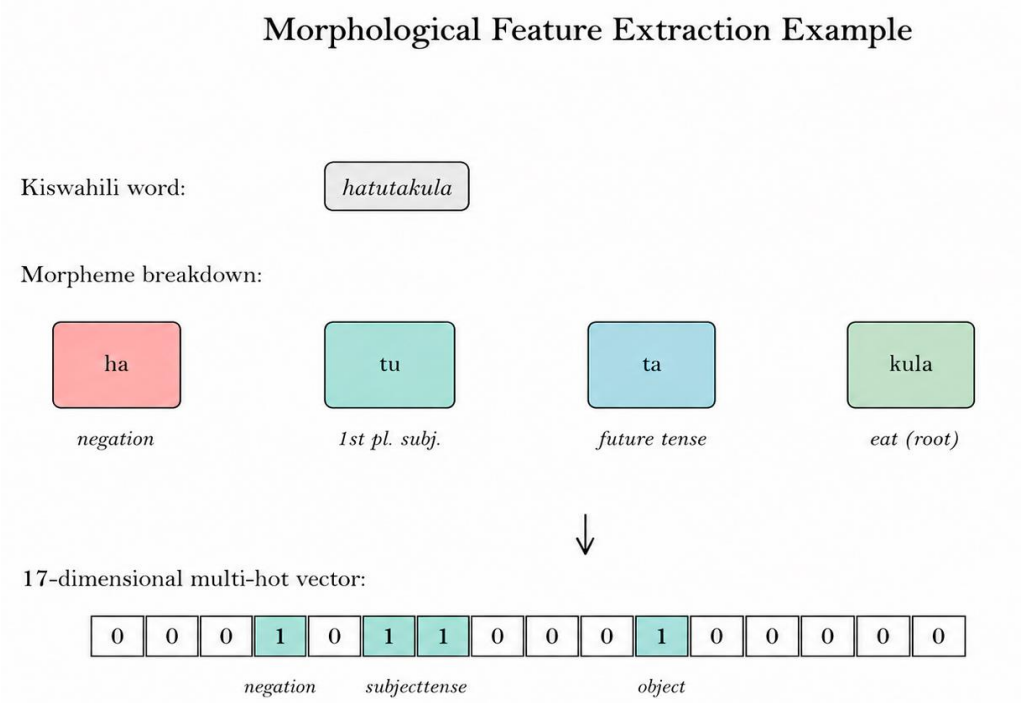

Fig. 1 Example of morphological feature extraction for the Kiswahili word "hatutakula" (we will not eat), showing morpheme decomposition and multi-hot vector encoding

3.5. Morphology-Aware Architecture

This morphology-aware architecture is based on XLM-RoBERTa and explicitly injects morphological information through learned projection layers while keeping the pre-trained multilingual representation with encoder freezing. This design is inspired by a hybrid mechanism of combining the advantages of large-scale multilingual pre-training and task-specific morphological inductive biases. The input tokens are passed to the XLM-RoBERTa encoder to get the contextualized representation H ∈ ℝL×768, where L is the sequence length, and 768 is the hidden dimension of the base model of XLM-RoBERTa. Freezing the encoder helps to avoid catastrophic forgetting of the multilingual information learned during pre-training, which preserves the model's general language understanding ability without compromising its morphological specialization. As shown in Equation (2), a learned linear transformation maps the binary morphological feature matrix M to continuous morphological embeddings M'.

$$M' = W_{morph}M + b_{morph} \tag{2}$$

The 17-dimensional multi-hot vector for each word is aligned to that word's subword tokens, so that every subword inherits the morphological vector of the word it belongs to. This projection layer is trained with task-specific morphologic representations that take into consideration the morphology by exploiting the interaction of morphs and the relation between morphology and the task-specific answer span prediction, converting sparse binary features into dense features. The transformer representations are concatenated with the projected morphological embeddings along the feature dimension to form the morphologically-enriched representation given in Equation (3).

$$H_{morph} = [H; M'] \in \mathbb{R}^{L\times 896} \tag{3}$$

Finally, Hmorph is fed into two independent linear classification layers to output start and end logits for predicting the boundaries of the answer span within the context paragraphs, allowing for precise modeling of the span boundaries. The start-position probabilities over the sequence are computed as in Equation (4).

$$P_{start} = softmax(W_{start}H_{morph} + b_{start} \tag{4}$$

Moreover, the end-position probabilities are computed analogously, as in Equation (5).

$$P_{end} = softmax(W_{end}H_{morph} + b_{end} \tag{5}$$

In the end-to-end training, the only thing that is updated is the morphological projection layer and the answer prediction heads, while the XLM-RoBERTa encoder is kept fixed to reduce cross-entropy loss on the start and end positions.

$$L = -(log\ P_{start}[s*] + log\ P_{end}[e*]) \tag{6}$$

Training minimizes the cross-entropy loss over the gold start and end positions, defined in Equation (6).

Figure 2 shows the overall morphology-aware architecture that has been implemented, where the input tokens are processed sequentially through the frozen XLM-RoBERTa encoder and output contextualized representations (768-dimensional), while the morphological features are projected to 128 dimensions and then are concatenated together with the 768-dimensional to form the 896-dimensional feature space used for the answer span prediction task.

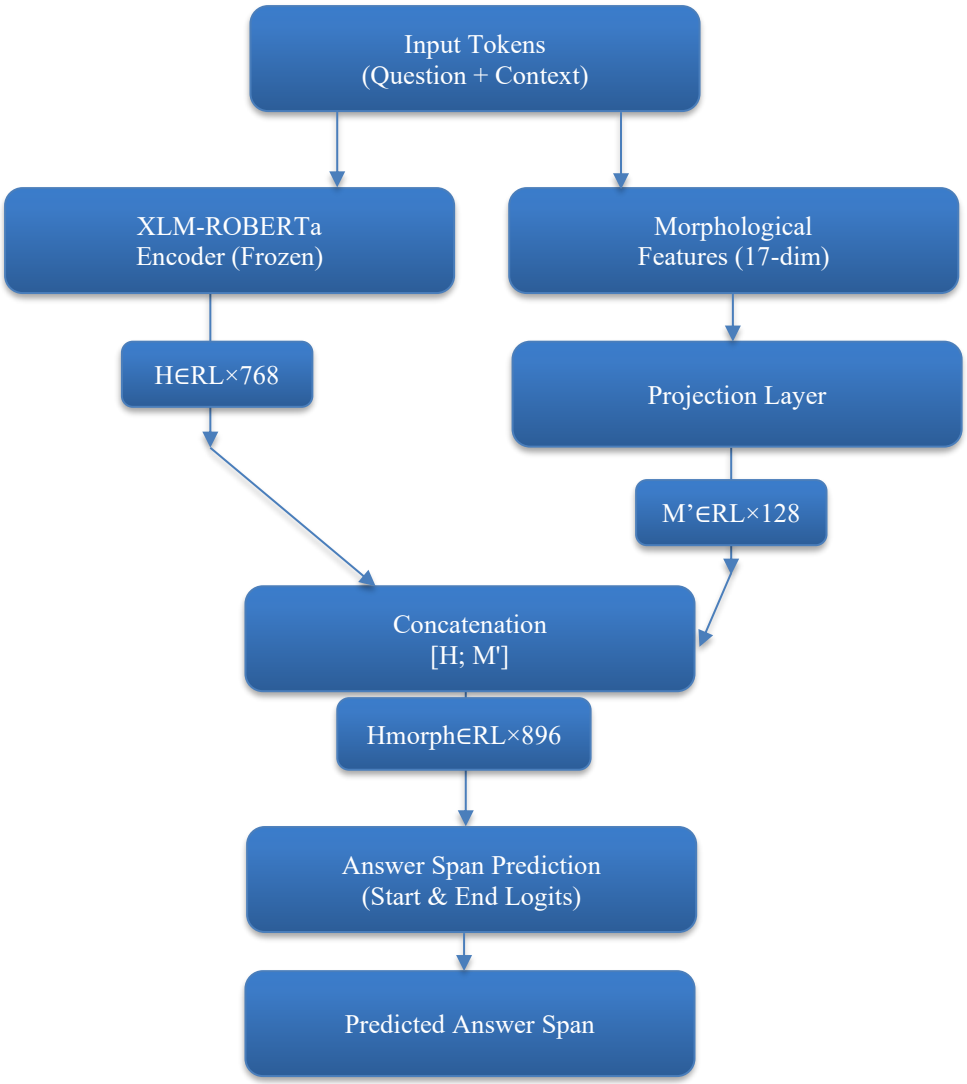

Fig. 2 Morphology-aware architecture diagram showing the integration of explicit morphological features with XLM-RoBERTa contextualized representations through learned projection and concatenation

3.6. Training Configuration and Implementation

The model was developed in three structured models, gradually improving the model training configuration. This iterative approach enabled isolation of performance contributions from architectural innovation versus training optimization. Version 1 (baseline) used the standard XLM-RoBERTa fine-tuning configuration with learning rate 2×10^{-5}, batch size 8, and 4 epochs, and set a standard performance for vanilla multilingual models. Version 2 features the morphological enhancement architecture with a higher learning rate of 3×10^{-5}, thanks to the extra trainable parameters that were added to account for the increased learning rate. It used an effective batch size of 16 through gradient accumulation, trained for 10 epochs with a warmup ratio of 0.1, and applied label smoothing of 0.1. This configuration is used to exploit the morphological features while preventing overfitting of the limited size of the training set. Version 3 further fine-tuned the learning process using differential learning rates, which were set to 1×10^{-5} for the frozen XLM-RoBERTa encoder, 3×10^{-5} for classification heads and answer prediction heads, and recognizing that different groups of parameters require different learning rates;

cosine annealing with warmup for more sophisticated learning rate scheduling: gradually decrease learning rates after warmup; increased regularization (dropout of 0.2 for classification heads and weight decay of 0.01 to prevent overfitting). The experiments were carried out on NVIDIA T4's GPU (16GB memory) in the Google Colab infrastructure, and implemented using the Hugging Face Transformers library version 4.x and PyTorch 2.x. To get efficient batch processing without computation overload, the morphological feature extraction and token-level alignments were implemented as custom preprocessors that are part of the data loading pipeline. To facilitate the reliability of experiments and a fair comparison of performances without stochastic variation, the same random seed was employed in all experiments across the different versions.

4. Results

4.1. Baseline Model Performance

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Performance is reported using token-level F1 and Exact Match; the F1 score is computed as the harmonic mean of precision and recall, as defined in Equation (7).

When evaluated on the KenSwQuAD validation set, vanilla XLM-RoBERTa achieved 20.05% F1 score and 17.80% Exact Match, correctly answering approximately 93 questions out of 522 validation examples. This weak baseline performance demonstrates significant limitations of standard multilingual approaches for morphologically complex low-resource languages. The model's difficulties despite having access to 4,178 training examples and having been pre-trained on 100 languages suggest that implicit learning of morphological patterns from subwords is insufficient when grammatical knowledge is explicitly captured in an agglutinative system of affixes that are fragmented by byte-pair encoding.

While the model achieves better-than-random performance, the low F1 score indicates that it struggles to learn robust question-answer mappings from the limited training data. The finding has important theoretical implications: transformers show measurable but limited capability on morphologically complex low-resource languages with standard training approaches, suggesting that the architecture requires explicit morphological augmentation rather than relying solely on implicit pattern learning from subword tokenization. This pattern is consistent with reports from other African-language QA benchmarks, where standard multilingual and cross-lingual pipelines yield weak in-language results (Ogundepo et al., 2023) and where typologically diverse languages remain difficult for current information-seeking QA systems (Clark et al., 2020).

4.2. Morphology-Aware Model Performance

The evolution of the performance is shown in Table 1 for three versions that are systematically designed for developing the program. Version 1 (baseline XLM-RoBERTa) had a weak performance level of 20.05% validation F1 and 17.80% Exact Match, demonstrating the limitations of standard approaches as described above. Version 2 with explicit morphological features showed substantial improvement with an F1 of 70.55% and an Exact Match of 61.95%. This represents a 50.5 percentage point improvement in F1, demonstrating a qualitative shift from weak baseline performance to functional question answering capability.

This optimized training configuration further boosted performance to 72.40% F1 and 62.91% Exact Match with a differential learning rate and improved regularization, for a total of 52.35 percentage points above baseline. The difference in performance between Version 2 and Version 3 (+1.85 F1 points) demonstrates that the innovation in the architecture is the determining factor for performance, but that careful use of training optimization can offer additional benefits. The validation performance (71.62% F1) was similar to the test performance (72.40% F1), implying that the model was not overfitting the training examples as it was learning strong morphological patterns. This drastic jump from the poor baseline to functional is seen graphically in Figure 3, where the 50.5 percent jump from baseline to functional is shown, followed by further refinement in Version 3, with both the F1 and Exact Match measurement scores plotted to emphasize the consistent improvement across the measurement lines.

A bootstrap resampling test was performed over validation data to determine if the improvement of the morphology-aware model over the baseline model is significant beyond chance. The token-level difference between the full morphology-aware model and the baseline was assessed on 1,000 bootstrap resamples of the validation examples, and the distribution of the difference was entirely above zero ($p < 0.001$), showing that the difference is statistically significant.

Table 1. Model Development Progression Across Three Versions

Version	Architecture	Validation EM (%)	Validation F1 (%)	Δ F1 from V1
V1: Baseline	XLM-RoBERTa vanilla	17.80	20.05	baseline
V2: Morphology-aware	MorphQA	61.95	70.55	+50.50
V3: Optimized	MorphQA optimized	62.91	72.40	+52.35

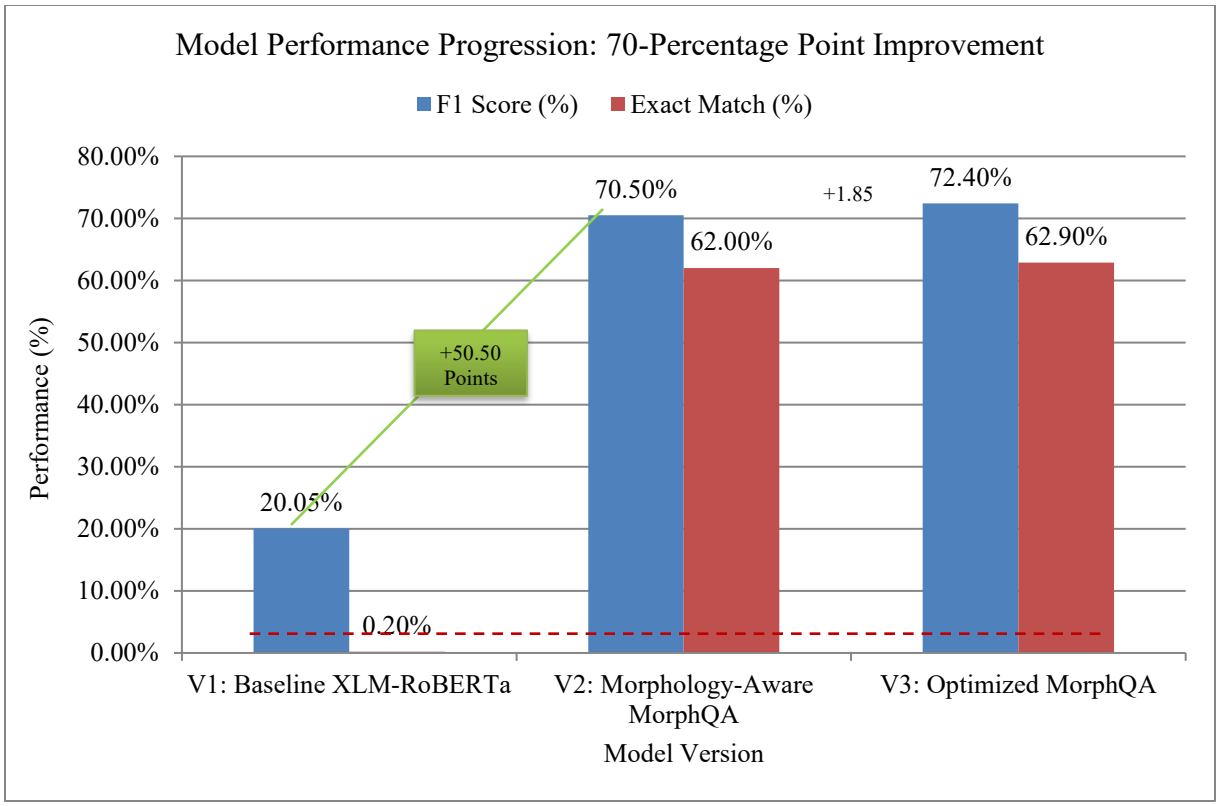

Fig. 3 Model performance progression across three development versions, demonstrating the 52.35 percentage point improvement from weak baseline performance to strong question answering capability through morphological feature integration

4.3. Error Analysis

A morphological analysis of the model's residual errors on the validation set showed a few common patterns. It is important to note that a large percentage of the remaining errors were partial span errors where the predicted answer overlapped with the gold answer but a bound morpheme was missing or added, such as returning a noun stem without the class prefix, and these errors are penalized by Exact Match, part of the reason for the difference between F1 (72.40%) and Exact Match (62.91%). Second, errors were more frequent on longer context paragraphs, consistent with the larger answer-span search space noted in Section 2.1. Third, questions whose answers depended on morphology outside the 17-morpheme inventory, particularly noun-class agreement and derivational forms, were more error-prone, indicating that the current feature set, while sufficient to produce large gains, does not capture the full morphological system. These observations motivate the expanded feature set and finer-grained alignment discussed in Section 4.3.

4.4. Ablation Study

The study used a controlled ablation experiment to isolate the performance contribution of morphological features from that of other factors, such as increased model size or different training procedures. An ablated model was trained with the same architecture and training configuration as Version 2, but without morphological feature integration: by only using the standard transformer hidden states H for answer span prediction instead of the morphologically-enriched ones Hmorph. The only architectural change between this and Version 2 is the use of label smoothing (0.1), increased learning rate (3×10^{-5}), larger effective batch size (16), and keeping the training duration (10 epochs) the same. This configuration keeps the training duration (10 epochs), the learning rate (3×10^{-5}), the effective batch size (16), and the label smoothing (0.1) parameter the same as in Version 2, and only includes morphological features as the difference in the architecture. The ablated model obtained 18.34% F1 on the validation set, which is slightly lower than the baseline XLM-RoBERTa score (20.05% F1), and significantly lower than the full morphology-aware model (70.55% F1). This 52.21 percentage point difference (70.55% - 18.34%) between the full model and the ablated demonstrates a clear validation that the performance improvement is indeed from the infusion of the morphological features and not from any other factors. The ablation result shows that explicitly learning the linguistic structure is essential for low-resource QA with limited training data; such a result cannot be obtained by merely adding trainable parameters (as in the morphological projection layer, which adds 4.3 million additional parameters). This finding contradicts the notion that performance improvements due to architectural changes are a result of greater model capacity, rather than enhanced inductive biases, in low-resource settings, and shows that

more effective architectural biases are definitely more important than the number of parameters.

5. Discussion

5.1. Computational Mechanisms Underlying Morphological Benefits

In agglutinative languages like Kiswahili, grammatical functions are systematically encoded through the concatenation of bound morphemes to lexical roots, creating morphologically complex word forms where single tokens encode multiple semantic and pragmatic distinctions simultaneously. Byte-Pair Encoding (BPE) based transformer models break up these morphologically consistent chunks into unrelated morphemes, frequently disconnecting grammatical affixes from lexical roots, and rendering the systematic relations among meaning components ambiguous (Bostrom & Durrett, 2020). Transformers have theoretical potential to learn morphological structure by attending to sequences of subwords, but that would involve implicit inductive learning of associations between the statistically-derived subword units and the grammatical functions. This implicit learning approach shows clear limitations in low-resource settings of morphologically complex languages, with the baseline achieving only 20.05% F1 and 17.80% EM on 522 validation items while having access to 4,178 training items. The weak performance (F1=20.05%), while better than random, demonstrates that the model struggles to learn robust question-answer mappings from limited training data when morphological complexity is high.

Augmenting the architecture with explicit multi-hot morpheme representations fundamentally changes the computational problem the model must solve. The model does not learn the morphological patterns from the limited co-occurrence statistics of its subwords, but rather it is given direct access to grammar information via the morphological feature vectors. This architectural inductive bias helps the model invest its relatively small data processing power in higher-level semantic composition and context integration tasks than in discovering low-level morphological patterns. The learned projection layer allows task-specific adaptation of morphological representations and learning the meaning of morphological difference, as well as how morphological difference contributes to semantic interpretation in the context of answer span prediction. The change from weak baseline performance to functional performance from the baseline to the morphology-aware architecture (Version 2) (+50.5 percentage points) is significant, suggesting that explicit morphological structure can provide the computational scaffolding required for learning in extremely low-resource situations. This improvement was unlikely to be achieved through hyperparameter optimization alone or by making minor changes to the architecture of the model, demonstrating the real alignment between the model architecture and the language, and not merely marginal tweaks. The size of this gain helps explain why the present approach outperforms standard multilingual configurations of the kind used as references in prior African-language work (Adelani et al., 2021; Alabi et al., 2022): rather than relying on the encoder to recover morphology implicitly, the architecture supplies it directly, which is precisely the capability that implicit subword learning lacks under data scarcity (Bostrom & Durrett, 2020).

The ablation study is crucial to give important confirmation that the improvements in performance are due solely to the morphological features of the device and not to any uncertainty. The ablated model (trained with the same hyperparameters, training time, and optimizers, but without morphological features) only has an F1 of 18.34%, showing that the model with more parameters (4.3 million in the projection layer) and more training epochs (10 vs 4) does not close this performance gap. This result provides insights into the efficiency of learning in low-resource contexts. The findings indicate that architectural inductive biases based on the structural properties of the language used offer a better reduction in sample complexity than merely increasing the number of parameters or training longer. Basically, the proper inductive biases are the reason that learning from fewer examples is possible, as they allow us to make computational shortcuts that avoid the inductive discovery of patterns that might logically be encoded.

5.2. Implications for African Language Technologies

The results show that explicit morphological encoding enables substantial performance improvements (from 20.05% to 72.40% F1) even with limited training data for low-resource African languages, reducing the data requirements for achieving functional QA systems. The approach is easily adapted to other Bantu languages with the same agglutinative morphological systems, such as Zulu (11 million speakers), isiXhosa (8 million speakers), Kinyarwanda (12 million speakers), Shona (11 million speakers), and Kikuyu (7 million speakers), which together account for more than 50 million speakers, in addition to Kiswahili's 100 million speakers. The morphological complexity of these languages is similar, as they have similar noun class systems, verbal morphology that is agglutinative, and affix-based grammar encoding that is systematic.

Enhancements to morphology should be as effective in these languages as they are in others. In addition to Bantu, the theoretical implications are relevant to morphologically complex language families around the world, such as the Turkic family (Turkish speakers 80 million, Uzbek 32 million, Kazakh 13 million), the Uralic family (Finnish speakers 5 million, Hungarian 13 million), and indigenous American languages, where the morphological complexity can be higher than in Kiswahili. Studies on Turkic and Uralic languages similarly report that morphology-sensitive segmentation improves downstream performance over standard subword tokenization, particularly in low-resource conditions, which

supports the cross-linguistic relevance of the present approach (Bostrom & Durrett, 2020). The general applicability of the approach implies that hybrid approaches that combine neural learning and explicit language knowledge provide viable routes for the construction of effective NLP systems in under-resourced languages without the need for a massive monolingual corpora for pre-training, or hundreds of thousands of language-specific annotated examples like those currently used in supervised high-resource language learning. This is consistent with evidence that careful use of limited data can rival training on far larger corpora for African languages (Ogueji et al., 2021).

The impact of AI technology that is accessible to every African language community is immense. This 52.35 percentage point improvement has shown that strategic linguistic knowledge integration can substantially reduce performance limitations previously thought to be due to underrepresented training data, where its collection could take orders of magnitude more examples. With typical annotation expenses of several dollars per question-answer pair and a lack of qualified annotators in low-resource languages, it becomes possible to reduce the amount of data required from hundreds of thousands to thousands of examples, making functional QA systems more economically viable in languages that are not currently used for AI development. The applications are extensive and in several high-impact areas. In the field of education, morphology-based question answering systems may play a pivotal role in developing smart tutoring systems that deliver personalized instruction and support in students' home languages, thereby helping to bridge the educational inequity that exists with limited or inadequate instruction in languages other than the mother tongue.

In the health sector, systems could facilitate the retrieval of health-related data by users who are not proficient in English, allowing them to access information using natural language queries, thereby enhancing health literacy and empowering individuals to make informed decisions. Question answering technology can be used to provide information on farming best practices, weather conditions, and market information to rural communities through agricultural extension services. Government services could utilize citizen interfaces that allow citizens to use information systems in their home languages and to navigate the bureaucracy.

The methodology's reproducibility and computational efficiency support deployment feasibility in resource-constrained environments. The morphological projection layer adds only 4.3 million parameters (3.9% increase) to XLM-RoBERTa's 278 million parameters, maintaining compatibility with standard GPU hardware and enabling inference on commodity devices. There is a slight increase in training time from 4 to 10 epochs, which is still manageable on platforms such as Google Colab that are freely accessible. These considerations make deployment realistic for organizations that serve the needs of African language communities at universities, NGOs, and government agencies with budget limitations.

5.3. Limitations and Future Directions

There are a number of caveats that should be noted and imply further research avenues. First, the morphological feature set includes only 17 verbal affixes, which is an incomplete set of Kiswahili morphology. There is no encoding for noun class markers (18 markers for the semantic classes, and 18 for their agreement patterns), verbal extensions (causative, applicative, passive, reciprocal, stative) that modify argument structure and semantics, or derivational morphology (processes that generate new lexemes from existing roots). This incompleteness has the potential to affect the performance of questions that involve an understanding of nominal morphology or complex verbal derivation. Adding these phenomena to the feature set might result in further improvements. However, it is not known which of the morphological differences would be most predictive, so systematic experiments would be required to achieve this. Second, the token-level alignment strategy aligns all subword tokens from the same orthographic word to the same morphological vectors, which may cause noise if the subword boundaries do not align with the morpheme boundaries. The use of more sophisticated per-subword alignment, via character-level mapping of morpheme-to-subword, could help increase the quality of features by ensuring that morphological features align correctly with the subword units that support their surface realizations. Thirdly and most crucially, manual construction of a morpheme inventory is only feasible for languages that have been fully documented in writing or possess a linguistic expert source of documentation. This would be eliminated if unsupervised or weakly-supervised morpheme discovery methods could automatically be learned from unlabeled corpora and produce productive affixes for undocumented languages, which is the goal of the future. Based on these recent advances in the research of neural morphological segmentation and cross-lingual morphology transfer, future research directions for automating feature engineering are suggested. Fourth, evaluation is limited to extractive question answering on one test set, which means that the performance on different task types (NER, Text Classification, Machine Translation, Abstractive QA) and domains (Medical, Legal, Conversational) remains unclear. The reliance on Exact Match and token-level F1 is itself a limitation, since both metrics can understate correctness when a morphologically inflected answer differs only by an affix from the gold span (Ogundepo et al., 2023). If the benefits of the morphological approach were general, then an evaluation across a variety of tasks and domains would establish this. Cross-lingual evaluations on typologically similar languages will confirm the transferability of Kiswahili morphological knowledge to other Bantu languages with minimal adaptation as a basis for building language family-specific morphological toolkits.

6. Ethical and Societal Implications

The implications of the development of language technology for low-resource African languages are not limited to benchmark performance. As Joshi et al. (2020) noted, the uneven distribution of NLP resources leaves speakers of the majority of the world's languages – most of which are African languages – outside of the benefits of modern AI.

The morphology-aware approach in this study reduces the amount of data needed to create functional systems, thereby lowering one of the obstacles to inclusion and making the development of AI more accessible for communities with less data, like in the case of low-resource languages. This can be used to propel digitization efforts toward equity in education, health information access, agricultural extension, and providing government services in citizens' first language.

Meanwhile, there are a number of risks that require a more specific focus. First, it is important to consider data representation and data provenance. Although valuable, Kiswahili Wikipedia is not comprehensive of all dialects, registers, and regional varieties of Kiswahili; models trained on it may not perform well for users who do not use it as Wikipedia's contributors do, and may capture the topical and cultural biases of the Wikipedia contributor community. Second, evaluation metrics have fairness implications: Exact Match and token-level F1 are string-based, and systematically undercount correct answers that vary only in morphological inflection, which can lead to an underestimation of the utility of the system for precisely the morphologically rich languages it is designed to serve (Ogundepo et al., 2023). Third, the process of building a morpheme inventory is manual, which introduces specific linguistic and dialectal biases, and the use of only a few highly-defined features can lead to the standardization of language, assuming dominance over community use. However, responsible deployment involves community participation in development, open discussion with native-speaker communities about training data sources and limitations, and assessment that extends beyond aggregate string-overlap. These factors are in line with the global need for inclusive, equitable, and accessible NLP technologies in other languages and application areas (Joshi et al., 2020) and should inform any future expansion to other languages and application areas.

7. Conclusion

This study created a Kiswahili morphology-aware question answering system, which obtained 72.40% F1 and 62.91% Exact Match score by explicitly incorporating 17 morpheme features into XLM-RoBERTa representations. This 52.35 percentage point increase from weak baseline performance (20.05% F1) shows that explicit linguistic structure is not only helpful, but also essential for morphologically complex low-resource African languages. Rigorous experiments have been performed to isolate the morphological contribution from other possible contributing factors to performance gains, which showed that the improvements are actually due to the infusion of morphological features and not to the increase in model size, training time, or optimization refinements. The results show that vanilla multilingual methods are insufficient for the agglutinative languages in resource-limited scenarios, and that the architecture needs to be adapted to the specific language structural properties rather than just scaling up and pre-training.

In addition to performance indicators, this work sheds light on basic language typology/neural architecture design correspondences, with evidence that current transformer architectures tacitly assume that word boundaries coincide with semantic units in isolating languages, and morphological structuring is limited. The extent of morphological effects – from weak baseline to strong functional competence – suggests that the difference in performance between language types is due to architectural-linguistic mismatches and not due to small variations in learning efficiencies.

The results reject the widespread belief that bigger models that learn from more data automatically work for all languages and instead argue in favor of typologically-appropriate architecture design as a prerequisite for fair AI development for linguistically-diverse populations (Joshi et al., 2020). This suggests that the NLP community needs to go beyond the one-size-fits-all architectures to designs that better align computational inductive biases and structural linguistic properties with language families.

There are four complementary directions for future research. First, more linguistic representation, including noun class markers, verbal extensions, and derivational morphology, could be beneficial and might yield further performance benefits. In the second, the creation of unsupervised morpheme discovery techniques would remove the need for manual feature engineering, allowing for fast adaptation to the hundreds of morphologically complex African languages that have not been well documented. Third, evaluation on related Bantu languages would be performed in a cross-linguistic manner to demonstrate the transfer of morphological representations from one language to another, which could facilitate the few-shot adaptation of Kiswahili resources to related languages. Fourth, the evaluation would be expanded to other tasks than question-answering, such as machine translation and text classification, providing a sense of "generality" of morphological advantages beyond the scope of question-answering, and guiding the design of language family-specific NLP toolkits.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research received no external funding.

Acknowledgments

The authors acknowledge the creators of the KenSwQuAD dataset for making their work publicly available. This research utilized computational resources provided by Google Colab.

References

[1] David Ifeoluwa Adelani et al., "MasakhaNER: Named Entity Recognition for African Languages," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1116-1131, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] David Ifeoluwa Adelani et al., "MasakhaNER 2.0: Africa-Centric Transfer Learning for Named Entity Recognition," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488-4508, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jesujoba O. Alabi et al., "Adapting Pre-Trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning," *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336-4349, Gyeongju, Republic of Korea, 2022. [Google Scholar] [Publisher Link]

[4] Kaj Bostrom, and Greg Durrett, "Byte Pair Encoding is Suboptimal for Language Model Pretraining," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4617-4624, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jonathan H. Clar et al., "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454-470, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Alexis Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440-8451, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Pratik Joshi et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282-6293, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Gati Martin et al., "SwahBERT: Language Model of Swahili," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, pp. 303-313, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin, "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resourced Languages," *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116-126, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Odunayo Ogundepo et al., "AfriQA: Cross-Lingual Open-Retrieval Question Answering for African Languages," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14957-14972, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Pranav Rajpurkar et al., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383-2392, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[13] Barack W. Wanjawa et al., "KenSwQuAD: A Question Answering Dataset for Swahili Low-Resource Language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]