Comparing Source Language Selection Strategies for Multi-Source Cross-Lingual Transfer to African Languages

Tewodros Kederalah Idris[1], **Roald Eiselen**[2], **Prasenjit Mitra**[1]
[1]Carnegie Mellon University Africa, Kigali, Rwanda
[2]Centre for Text Technology, North-West University, Potchefstroom, South Africa
tidris@andrew.cmu.edu, Roald.Eiselen@nwu.ac.za, prasenjm@andrew.cmu.edu

Abstract

Cross-lingual transfer learning enables building NLP systems for low-resource languages by leveraging data from higher-resource languages. A critical but understudied question for African languages is: which source languages should be selected for multi-source transfer? We present a systematic comparison of four source language selection strategies: random selection (baseline), genetic distance based on language family trees, geographic distance based on speaker locations, and embedding similarity from multilingual models. We evaluate these strategies on Named Entity Recognition, Part-of-Speech tagging, and sentiment analysis across five typologically diverse African target languages (Hausa, Yoruba, Swahili, Igbo, Kinyarwanda) using three multilingual models. We further investigate how the number of source languages affects transfer performance. Our experiments reveal that no single strategy dominates across tasks: geographic distance leads on sequence labeling tasks while embedding similarity is most effective for sentiment analysis, and all informed strategies consistently outperform random selection.

Keywords: cross-lingual transfer, source language selection, African languages, multilingual NLP, low-resource languages

1. Introduction

Building natural language processing systems for low-resource languages remains a significant challenge, particularly for the over 2,000 languages spoken across Africa (Eberhard et al., 2024). These languages exhibit high typological diversity across multiple language families (Niger-Congo, Afro-Asiatic, Nilo-Saharan, Khoisan), with many languages having limited linguistic proximity to higher-resource languages (Ogueji et al., 2021; de Vries et al., 2022). This diversity makes African languages an ideal testbed for evaluating source selection strategies, as transfer often requires crossing language family boundaries where traditional typological features provide limited guidance. Cross-lingual transfer learning offers a promising solution: leveraging labeled data from higher-resource languages to build systems for languages with limited or no training data (Pires et al., 2019; Wu and Dredze, 2019). Recent work has demonstrated the effectiveness of multilingual pretrained models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020b), and African-focused models like AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022) for transferring knowledge across languages. These models learn language-agnostic representations that enable transfer even between typologically distant languages (Conneau et al., 2020a; de Souza et al., 2024).

While much attention has focused on improving multilingual model architectures, a fundamental practical question remains understudied: *which source languages should practitioners select for cross-lingual transfer?* This question becomes particularly important in multi-source transfer settings, where combining data from multiple source languages can outperform single-source transfer (Lim et al., 2024; Ansell et al., 2021). For practitioners working with African languages, source selection decisions have direct implications for data collection efforts, annotation costs, and downstream system performance.

Prior work on source language selection has explored various strategies. Typological approaches leverage linguistic features such as language family, word order, and morphology (de Vries et al., 2022; Rice et al., 2025), often using databases like URIEL/lang2vec (Littell et al., 2017). More recently, embedding-based methods have shown promise by computing similarity directly from multilingual model representations (Idris et al., 2026b; Ebrahimi et al., 2025), though their effectiveness varies by task and language pair (Rice et al., 2025). Multi-source approaches combine data from multiple languages (Lim et al., 2024; Ansell et al., 2021), though optimal source combinations remain underexplored.

However, most source selection research has focused on European and Asian languages. Recent work on African NLP has developed specialized models (Ogueji et al., 2021) and benchmarks (Adelani et al., 2022; Dione et al., 2023), demonstrating that source language choice can improve performance by 14 F1 points over default English transfer (Adelani et al., 2022). Yet systematic comparison of source selection strategies specifically for African

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 31–40
12 May 2026.

languages remains lacking. Existing studies often transfer from English by default (Thangaraj et al., 2024; Ogundepo et al., 2023), leaving open the question of whether alternative source languages or combinations might be more effective.

In this paper, we present a systematic comparison of four source language selection strategies for African languages: random selection (baseline), genetic distance based on language family relationships (Littell et al., 2017), geographic distance based on speaker locations, and embedding similarity computed from multilingual models. We evaluate these strategies across three tasks (Named Entity Recognition, Part-of-Speech tagging, and sentiment analysis), five typologically diverse target languages (Hausa, Yoruba, Swahili, Igbo, and Kinyarwanda), and three multilingual models (AfriBERTa, AfroXLMR, and Serengeti (Adebara et al., 2023)). We further investigate how the number of source languages affects transfer performance.

Our contributions are threefold. First, we provide the first systematic comparison of source selection strategies specifically for African languages, evaluating how genetic, geographic, and embedding-based approaches perform across diverse typological scenarios. Second, we investigate the effect of the number of source languages in zero-shot settings, providing guidance on how many existing annotated datasets practitioners should transfer-learn from when no target language training data is available. Third, we analyze selection strategy effectiveness across different tasks and model architectures, revealing that the optimal strategy depends on the task type, providing actionable guidance for practitioners building NLP systems for low-resource African languages.

2. Related Work

2.1. Source Language Selection Strategies

Cross-lingual transfer relies on selecting appropriate source languages to maximize knowledge transfer. Prior work has explored three main selection strategies. **Genetic distance**, based on language family relationships, has been widely used through typological databases like URIEL (Littell et al., 2017). Large-scale studies show that genealogical distance reliably predicts transfer performance (de Vries et al., 2022; Rice et al., 2025), with learned ranking approaches like LangRank (Lin et al., 2019) incorporating genetic features alongside dataset properties. **Geographic distance** has been proposed as an alternative that captures language contact and regional borrowing (Nasir and Mchechesi, 2022; Winata et al., 2022). Nasir and Mchechesi (2022) demonstrated that geographic proximity can predict optimal source languages for African language translation, while Winata et al. (2022) found that geographically similar languages improve cross-lingual adaptation. **Embedding similarity**, computed directly from multilingual model representations, has emerged as a model-based alternative. Lin et al. (2023) showed that similarity induced from pretrained models outperforms linguistic features by 1–2% on zero-shot transfer, and Ebrahimi et al. (2025) demonstrated that representation-based ranking beats feature-based baselines by 35.56 points in Normalized Discounted Cumulative Gain (NDCG), a ranking quality metric that measures how well relevant items are placed near the top of a ranked list (Järvelin and Kekäläinen, 2002).

For African languages specifically, Idris et al. (2026b) evaluated embedding similarity metrics for predicting cross-lingual transfer across NER, POS tagging, and sentiment analysis, the same tasks examined in this work, finding that cosine gap and retrieval-based metrics moderately predict transfer success (ρ = 0.4–0.6).

2.2. Multi-Source Cross-Lingual Transfer

Recent work demonstrates benefits from combining multiple source languages. Lim et al. (2024) showed that multi-source transfer consistently outperforms single-source approaches and investigated optimal numbers of source languages, finding that combining diverse sources leads to increased mingling of embedding spaces across languages. Ansell et al. (2021) developed MAD-G, which generates language-specific adapters by combining information from multiple sources weighted by typological similarity. However, Lim et al. focused on European and Asian languages, while MAD-G employed adapter-based methods rather than direct fine-tuning. Neither study systematically compared selection strategies for African languages or investigated whether findings about optimal source counts generalize to typologically distinct language families.

2.3. African Language NLP

The African NLP community has developed dedicated resources to address language underrepresentation (Joshi et al., 2020). Benchmarks such as MasakhaNER (Adelani et al., 2022), Masakha-POS (Dione et al., 2023), and AfriSenti (Muhammad et al., 2023) enable systematic cross-lingual transfer evaluation. Multilingual models like AfriBERTa (Ogueji et al., 2021), AfroXLMR (Alabi et al., 2022), and Serengeti (Adebara et al., 2023) have been developed specifically for African languages. Studies using these resources show that source language choice significantly impacts performance,

with Adelani et al. (2022) finding 14 F1 point improvements over English for NER. However, these studies examined source selection post-hoc rather than systematically comparing selection strategies. Our work addresses this gap by providing the first controlled comparison of genetic, geographic, and embedding-based selection strategies for African languages across multiple tasks, target languages, and model architectures, while also investigating how the number of source languages affects transfer performance.

3. Methodology

3.1. Source Language Selection Strategies

We compare four strategies for selecting source languages in multi-source cross-lingual transfer:

Random Selection. We randomly sample K languages from the available source pool, serving as an uninformed baseline. We use a fixed seed to ensure reproducibility, making this baseline deterministic across all experiments.

Genetic Distance. We select the K languages with smallest genetic distance to the target, computed using the URIEL typological database (Littell et al., 2017) via `lang2vec`. Genetic distance captures language family relationships, with languages from the same family having smaller distances.

Geographic Distance. We select the K languages with smallest geographic distance to the target, also computed via URIEL. Geographic distance captures spatial proximity between speaker populations, which may reflect contact-induced similarities.

Embedding Similarity. We select the K languages with highest embedding similarity to the target. We extract mean-pooled sentence embeddings from each model's final layer using 2,000 parallel sentences from FLORES-200 (NLLB Team, 2024). We then compute the cosine gap score, defined as the difference between the average cosine similarity of correct translation pairs and that of incorrect pairs. This metric addresses the anisotropy problem in multilingual embeddings, where raw cosine similarity fails to discriminate between languages due to embeddings clustering in narrow cones. Raw cosine similarity between multilingual embeddings tends to produce uniformly high scores across all language pairs due to this clustering, making it difficult to distinguish genuinely aligned languages from superficially similar ones. Cosine gap addresses this by measuring whether a model can distinguish correct translation pairs from incorrect ones: for a well-aligned language pair, correct translations should score noticeably higher than random cross-lingual pairings, producing a large gap. A small gap indicates that the model treats correct and incorrect pairings similarly, suggesting weak functional alignment regardless of the raw cosine score. We select the K sources with highest cosine gap scores relative to the target. The specific values of K tested are detailed in Section 3.5.

3.2. Tasks and Datasets

We evaluate on three sequence labeling and classification tasks from the Masakhane project:

Named Entity Recognition (NER). We use MasakhaNER 2.0 (Adelani et al., 2022), which provides manually annotated NER data for 20 African languages with four entity types (PER, ORG, LOC, DATE). We report entity-level F1 scores following standard practice.

Part-of-Speech Tagging (POS). We use MasakhaPOS (Dione et al., 2023), which provides POS annotations for 20 African languages using the Universal Dependencies tagset. We report token-level accuracy following Universal Dependencies conventions.

Sentiment Analysis. We use AfriSenti (Muhammad et al., 2023), which provides sentiment-annotated tweets for 14 African languages with three classes (positive, negative, neutral). We report weighted F1 score to account for class imbalance in social media data.

Table 1 shows training set sizes per language and task. Dataset sizes vary substantially for NER (3,384 to 7,825 sentences) and sentiment (1,810 to 14,172 sentences), while POS datasets are more uniform (693 to 893 sentences). For NER and POS, the source pool contains 19 languages per target. For sentiment analysis, only 8 languages have available data in AfriSenti, resulting in 7 candidate source languages per target.

3.3. Models

We evaluate three multilingual models designed for African languages:

AfroXLMR (Alabi et al., 2022): An XLM-R model adapted through continued pretraining on African language data.

AfriBERTa (Ogueji et al., 2021): A transformer model pretrained from scratch on 11 African languages.

Serengeti (Adebara et al., 2023): A multilingual model covering 517 African languages and language varieties.

3.4. Languages

We select five typologically diverse target languages: Hausa (hau), Yoruba (yor), Swahili (swa), Igbo (ibo), and Kinyarwanda (kin). These span two major language families: Afro-Asiatic (Hausa) and

Language	NER	POS	Sentiment
amh	–	–	5,985*
bam	4,462	775	–
bbj	3,384	750	–
ewe	3,505	728	–
fon	4,343	810	–
hau†	5,716	753	14,172
ibo†	7,634	803	10,192
kin†	7,825	757	3,302
lug	4,942	733	–
luo	5,161	758	–
nya	6,250	728	–
pcm	5,646	752	5,121
sna	6,207	747	–
swa†	6,593	693	1,810
tsn	3,489	754	–
twi	4,240	785	3,481
wol	4,593	782	–
xho	5,718	752	–
yor†	6,876	893	8,522
zul	5,848	753	–

Table 1: Training set sizes (sentences) per language and task. †Target languages. *Source only (not used as target). Data sources: MasakhaNER 2.0 (NER), MasakhaPOS (POS), AfriSenti (Sentiment).

Niger-Congo, with the latter including both Bantu (Swahili, Kinyarwanda) and Volta-Niger (Yoruba, Igbo) branches. For each target, the source pool consists of all other languages available in the respective datasets, yielding 19 candidate sources for NER and POS, and 7 for sentiment analysis.

3.5. Experimental Design

Phase 1: Strategy Comparison. We fix $K = 3$ source languages following Lim et al. (2024), who found this to be effective for multi-source transfer. For each combination of task, model, and target language, we select sources using each of the four strategies, fine-tune on the combined source data, and evaluate on the target test set. All five target languages appear in all three datasets, yielding 3 tasks $\times$ 3 models $\times$ 5 targets $\times$ 4 strategies = 180 experimental configurations.

Phase 2: Optimal Source Count. Using embedding-based selection (which achieves the highest overall average across Phase 1), we vary the number of source languages $K \in \{1, 2, 3, 5\}$ to investigate whether the choice of $K = 3$ generalizes to African languages. We test all combinations of task, model, and target language for each value of K, yielding 3 tasks $\times$ 3 models $\times$ 5 targets $\times$ 4 values of $K = 180$ additional configurations.

Strategy	NER	POS	Sent.	Avg
Geographic	**.673**	**.711**	.427	.604
Genetic	.645	.677	.489	.604
Embedding	.644	.694	**.505**	**.614**
Random	.638	.632	.451	.574

Table 2: Phase 1 results: Average performance by selection strategy (K=3). Best results per task in **bold**. NER and sentiment report F1; POS reports accuracy.

3.6. Training Details

We fine-tune all models for up to 10 epochs using the AdamW optimizer with learning rate 2×10^{-5}, batch size 16, weight decay 0.01, and warmup ratio 0.1. Maximum sequence length is set to 256 tokens for NER and POS, and 128 tokens for sentiment. We use epoch-level evaluation and select the best checkpoint based on source language validation F1 (or accuracy for POS), maintaining zero-shot evaluation on target languages. For multi-source training, we concatenate training data from all selected source languages without resampling, allowing natural dataset size variation. We use the standard train/dev/test splits provided by each benchmark.

4. Results

4.1. Phase 1: Strategy Comparison

Table 2 presents the performance of each source language selection strategy across all three tasks, averaged over 3 models and 5 target languages per task.

The results reveal that no single selection strategy dominates across all tasks. Geographic distance achieves the best performance on both NER (.673) and POS (.711), while embedding-based selection leads on sentiment (.505). All three informed strategies outperform random selection, with gains of 3–8 percentage points in overall average.

Several task-specific patterns emerge. For the two sequence labeling tasks (NER and POS), geographic distance provides the strongest transfer signal. This may reflect the fact that geographically proximate African languages share structural features relevant to token-level prediction, such as similar morphological patterns or shared borrowings, even when they belong to different language families. For sentiment analysis, embedding-based selection provides the best performance, while geographic distance performs worst (.427), even below random selection (.451). This divergence likely reflects domain mismatch: geographic proximity captures structural similarities relevant to sequence labeling, but sentiment expression patterns in social media may depend more on the cross-lingual

Model	Strategy	NER	POS	Sent.
AfriBERTa	Geographic	**.629**	**.682**	.411
	Genetic	.542	.627	.491
	Embedding	.614	.663	**.502**
	Random	.570	.544	.457
AfroXLMR	Geographic	**.693**	**.726**	.486
	Genetic	.707	.711	.490
	Embedding	.667	.723	**.502**
	Random	.674	.669	.450
Serengeti	Geographic	**.696**	**.725**	.444
	Genetic	.686	.694	.487
	Embedding	.650	.697	**.511**
	Random	.672	.683	.446

Table 3: Phase 1 results per model, averaged across five target languages (K=3). Best results per model and task in **bold**.

alignment captured by embedding similarity.

Across all tasks, the three informed strategies substantially outperform random selection, confirming that principled source language selection provides meaningful benefits for cross-lingual transfer to African languages.

Per-Model Performance. Table 3 reports average performance for each model individually. The strategy rankings are largely consistent across models: geographic distance leads on NER for all three models and on POS for AfroXLMR and Serengeti, while embedding similarity leads on sentiment for all three. However, individual models show notable divergences. AfriBERTa exhibits substantially lower scores under genetic distance for NER (average .542 vs. .707 for AfroXLMR), particularly for Hausa (.382) where the selected sources (Bambara, Ghomala, Ewe) share neither family nor pretraining data with the target. AfroXLMR achieves the highest scores on NER and POS, while Serengeti leads slightly on sentiment.

Per-Language Performance. Table 4 shows NER performance broken down by target language, averaged across three models. Geographic distance achieves the highest scores for three of five targets (Hausa, Igbo, Kinyarwanda). For Swahili, genetic distance performs best, correctly identifying Bantu relatives (Luganda, Chichewa, Kinyarwanda). For Yoruba, genetic distance also leads by identifying the closely related Igbo. These results demonstrate that the effectiveness of each strategy depends on the typological relationships available in the source pool.

Source Selection Patterns. Table 5 shows the source languages selected by each strategy across all five targets. The strategies exhibit distinct selection behaviors reflecting their underlying assumptions.

Strategy	hau	yor	swa	ibo	kin
Embedding	.662	.504	.698	.673	.681
Genetic	.535	**.546**	**.777**	.688	.681
Geographic	**.728**	.430	.773	**.740**	**.692**
Random	.681	.482	.716	.668	.645

Table 4: NER F1 scores by target language, averaged across three models (K=3). Best per-language results in **bold**.

Genetic distance selects based on language family relationships. For Bantu targets (Swahili, Kinyarwanda), it correctly identifies Bantu relatives: Luganda, Chichewa, and the other Bantu target (Kinyarwanda for Swahili, Swahili for Kinyarwanda). For Yoruba, it identifies Igbo (both Volta-Niger). However, for Hausa (Afro-Asiatic), no close relatives exist in the source pool, and the strategy defaults to distantly related Niger-Congo languages.

Geographic distance selects based on speaker proximity. East African targets (Swahili, Kinyarwanda) receive East African sources (Kinyarwanda/Swahili, Luganda, Luo), while West African targets (Hausa, Yoruba, Igbo) receive West African sources including Nigerian Pidgin, Ghomala, Yoruba, and Fon. Notably, geographic selection places Yoruba as a source for both Hausa and Igbo, capturing the regional clustering of Nigerian languages.

Embedding-based selection shows a distinct pattern: it consistently selects Hausa, Swahili, and Southern Bantu languages (Zulu, Xhosa) across targets, regardless of genetic or geographic proximity. Notably, all three models select identical sources despite very different pretraining compositions: AfriBERTa was pretrained on 11 languages (not including Zulu or Xhosa), AfroXLMR was adapted on 17 languages (including Zulu and Xhosa), and Serengeti covers 517 languages. The fact that AfriBERTa selects Zulu and Xhosa despite never being pretrained on them suggests that the Bantu languages form a tight typological cluster in embedding space: once a model learns representations for some Bantu languages (e.g., Swahili, Kinyarwanda), it develops representations that align well with other Bantu languages even without direct exposure. This likely reflects the deep structural similarity among Bantu languages, including shared noun class systems, verb morphology, and extensive cognate vocabulary. While pretraining data volume may reinforce this clustering for well-resourced languages, the AfriBERTa evidence indicates that typological similarity is a primary driver.

Target	Strategy	Selected Sources
Hausa	Embedding	swa, zul, xho
	Genetic	bam, bbj, ewe
	Geographic	pcm, bbj, yor
Yoruba	Embedding	hau, kin, swa
	Genetic	ibo, bbj, nya
	Geographic	fon, pcm, ewe
Swahili	Embedding	hau, zul, xho
	Genetic	lug, nya, kin
	Geographic	kin, lug, luo
Igbo	Embedding	hau, zul, swa
	Genetic	bbj, nya, lug
	Geographic	bbj, yor, pcm
Kinyarwanda	Embedding	swa, zul, hau
	Genetic	lug, swa, nya
	Geographic	swa, lug, luo

Table 5: Source languages selected by each strategy for NER (K=3).

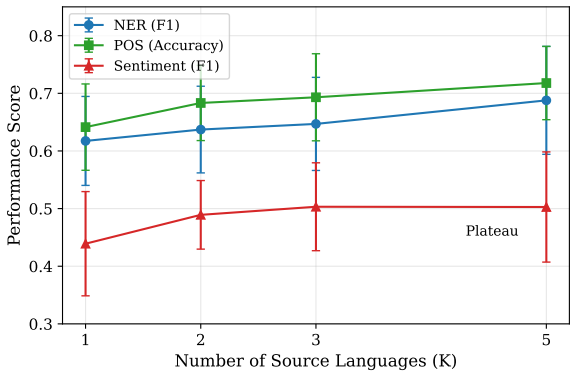

Figure 1: Transfer performance vs. number of source languages (K) using embedding-based selection. NER and POS improve monotonically up to K=5, while sentiment plateaus at K=3. Points show mean performance across three models and five target languages; error bars indicate standard deviation across these 15 configurations.

4.2. Phase 2: Optimal Number of Sources

Using embedding-based selection (which achieves the highest overall average in Phase 1), we vary the number of source languages to determine the optimal K. Figure 1 presents results averaged across three models and five targets.

The results reveal that more source languages generally improve transfer performance. For NER and POS, performance increases monotonically from K=1 to K=5, with K=5 outperforming K=3 by 4.1 and 2.5 percentage points respectively. This contrasts with prior findings on European languages, where Lim et al. (2024) observed diminishing returns beyond K=3. The continued improvement with K=5 for African languages may reflect greater typological diversity in our source pool, where even the fifth-best source provides complementary information not captured by the top three.

Sentiment analysis exhibits different behavior, with performance plateauing at K=3. This may reflect the smaller source pool for sentiment (8 languages vs. 19 for NER/POS) or domain mismatch between the Twitter-based sentiment data and the news-domain FLORES-200 sentences used to compute embedding similarity.

4.3. Analysis

Why does geographic distance perform well on sequence labeling? The strong performance of geographic distance on NER and POS suggests that regional proximity captures linguistically relevant features for token-level tasks. Geographically proximate African languages often share lexical borrowings, similar morphological strategies, and overlapping named entity conventions due to shared cultural and political contexts. For instance, geographic selection places Nigerian Pidgin, Ghomala, and Yoruba as sources for Hausa, all of which share the West African linguistic area where extensive language contact has produced convergent features in vocabulary, phonology, and morphosyntax. For East African targets, it identifies languages from the Great Lakes and East African coastal regions, where Bantu languages have undergone contact-induced convergence distinct from their Southern Bantu relatives. This also helps explain why geographic distance outperforms embedding similarity on sequence labeling despite using less total training data (Table 6): geographic selection produces structurally diverse regional sources, whereas embedding similarity concentrates on the Bantu cluster, providing redundant structural signal from typologically similar languages.

Why does embedding similarity lead on sentiment? Embedding-based selection outperforms all other strategies on sentiment analysis while performing comparably on sequence labeling tasks. For sentiment, the pooled sequence representation used for classification may benefit more from aligned embedding spaces than from token-level structural similarity. Embedding similarity, computed from FLORES-200 parallel sentences, captures how well two languages share a common representation space in a given model, which directly relates to how well sentiment-bearing features transfer. In contrast, geographic and genetic proximity may select structurally similar languages that nonetheless express sentiment differently in the social media domain.

When do linguistic distances help? Genetic distance is most effective when the target has close relatives in the source pool. For Swahili, genetic distance selects fellow Bantu languages (Luganda, Chichewa, Kinyarwanda) and achieves the highest NER F1 (.777). For Yoruba, selecting Igbo (both Volta-Niger) also proves effective. However, for Hausa (Afro-Asiatic), no close relatives exist in the source pool, and genetic distance defaults to distantly related Niger-Congo languages, yielding the lowest NER F1 (.535) among all strategy-target combinations.

Task differences. The three tasks show different strategy rankings, which we attribute to the distinction between token-level and sequence-level prediction. NER and POS are sequence labeling tasks where each token's representation directly determines its label. These tasks benefit from structural similarities captured by geographic proximity. Sentiment analysis relies on a pooled sequence representation for classification, making it more sensitive to overall embedding space alignment than to token-level structural features. This distinction suggests practitioners should consider the task type when choosing a selection strategy.

Training data quantity analysis. Since source languages have varying dataset sizes (Table 1), different strategies yield different total training data. To assess whether performance differences simply reflect data quantity, Table 6 reports the average total training sentences alongside performance for each strategy and task. Because all three models select identical sources for each strategy (genetic and geographic distances are model-independent, and embedding similarity produces the same rankings across models), these training data totals apply equally to all models; per-model performance can be cross-referenced in Table 3. The data quantity rankings do not align with performance rankings on any of the three tasks. For NER, geographic distance achieves the highest F1 (.673) while using the *least* average training data (15,986 sentences), compared to 18,378 for embedding similarity. At the per-target level, the strategy with the most training data is the best performer in only 1 of 5 cases. For POS, all strategies use nearly identical totals (2,211 to 2,302 sentences) due to the uniform dataset sizes in MasakhaPOS, effectively providing a natural control for data quantity; geographic distance still leads. For sentiment, geographic distance uses the most data (23,320 sentences) yet performs worst (.427), while embedding similarity achieves the best performance (.505) with a smaller total (20,062). These patterns indicate that the observed strategy differences reflect genuine language selection effects rather than data quantity artifacts.

Task	Strategy	Data	Score
NER	Embedding	18,378	.644
	Genetic	15,999	.645
	Geographic	15,986	**.673**
POS	Embedding	2,211	.694
	Genetic	2,223	.677
	Geographic	2,302	**.711**
Sent.	Embedding	20,062	**.505**
	Genetic	18,191	.489
	Geographic	23,320	.427

Table 6: Average total training sentences (Data) and performance (Score) per strategy and task. The best-performing strategy (bold) does not use the most data for any task.

5. Conclusion

We presented a systematic comparison of source language selection strategies for multi-source cross-lingual transfer to African languages. Our experiments across three tasks, five target languages, and three multilingual models yield several key findings.

First, no single strategy dominates across all tasks. Geographic distance achieves the best performance on both sequence labeling tasks (NER and POS), while embedding similarity leads on sentiment analysis. This task dependence has not been previously documented for African languages and contrasts with prior work on European languages suggesting embedding-based methods consistently outperform linguistic distance measures (Lin et al., 2023; Ebrahimi et al., 2025).

Second, all informed selection strategies outperform random selection, with gains of 3–8 percentage points overall. This confirms that principled source language selection provides meaningful benefits regardless of which strategy is chosen, and that the common practice of transferring from English by default leaves substantial performance on the table.

Third, more source languages generally improve transfer performance. Unlike prior findings on European languages showing diminishing returns beyond K=3 (Lim et al., 2024), we observe continued improvement up to K=5 for NER and POS, suggesting that typologically diverse source pools benefit from additional languages.

These findings have practical implications for practitioners building NLP systems for low-resource African languages. For token-level tasks like NER and POS tagging, geographic distance provides a simple and effective selection criterion. For sequence-level classification tasks like sentiment

analysis, practitioners should compute embedding similarity (e.g., using cosine gap on FLORES-200 parallel sentences) from the target multilingual model. When resources permit, including more source languages (K=5) yields better results than the commonly used K=3.

6. Limitations

Our study has several limitations. First, our strategy comparison does not control for total training data quantity. Since source languages have varying dataset sizes (Table 1), different strategies yield different total training examples. Our analysis in Table 6 shows that data quantity rankings do not align with performance rankings for any task, indicating that the observed differences are not simply data quantity artifacts. Nevertheless, a fully controlled comparison holding total training data constant would more rigorously isolate strategy effects from data quantity effects. In concurrent work, we address this directly through a budget-constrained framework that holds total training data constant across strategies, finding that once data quantity is controlled, multi-source transfer remains strongly beneficial while differences among specific allocation strategies are modest (Idris et al., 2026a).

Second, we report results from a single random seed for the genetic and geographic strategy comparisons due to GPU compute constraints. While the consistent patterns across 15 configurations per task (3 models $\times$ 5 targets) and the per-model analysis in Table 3 provide reasonable evidence, multi-seed experiments would strengthen confidence. Our experiments varying the number of source languages (Section 3.5) and embedding-based selection both include multiple seeds and show consistent trends.

Third, we evaluate only three models; while these represent diverse pretraining strategies, results may differ for other architectures. Fourth, our embedding similarity metric relies on FLORES-200, which may not capture domain-specific similarity for tasks like sentiment analysis based on social media text.

7. Ethics Statement

This work uses publicly available datasets and pretrained models. We do not foresee negative societal impacts from this research. All datasets used (MasakhaNER 2.0, MasakhaPOS, AfriSenti) were created with appropriate consent and annotation practices as described in their respective publications.

8. Acknowledgements

This publication was developed as part of the Center for Inclusive Digital Transformation of Africa (CIDTA), and, the Afretec Network which is managed by Carnegie Mellon University Africa and receives financial support from the Mastercard Foundation. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Carnegie Mellon University or the Mastercard Foundation.

9. Bibliographical References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2023. Serengeti: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 75–94, Toronto, Canada. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Oluwadara Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing K. Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin P. Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius M Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Brasil. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8615–8631, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Leandro Rodrigues de Souza, Thiago Almeida, Roberto A. Lotufo, and Rodrigo Nogueira. 2024. Measuring cross-lingual transfer in bytes. *arXiv preprint arXiv:2404.08191*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7231–7246, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Kathleen Siminyu, Andiswa Bukula, Rooweither Mabuya, Happy Buzaaba, Godson Kalipe, Jonathan Mukiibi, Victoire Auguste Memdjokam Koagne, Blessing K. Sibanda, Tatiana Moteu Ngoli, Tosin Adewumi, Fatoumata Kabore, Chris Chinenye Emezue, Catherine Gitau, Edwin Munkoh-Buabeng, Oreen Yousuf, Tajuddeen Gwadabe, Shamsuddeen Hassan Siminyu, Vukosi Marivate, and Dietrich Klakow. 2023. Masakhapos: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11504–11522, Singapore. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the world.

Abteen Ebrahimi, Adam Wiemerslage, and Katharina von der Wense. 2025. Model-based ranking of source languages for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2510.03202*.

Tewodros Kederalah Idris, Roald Eiselen, and Prasenjit Mitra. 2026a. Budget-xfer: Budget-constrained source language selection for cross-lingual transfer to african languages. *arXiv preprint arXiv:2603.27651*.

Tewodros Kederalah Idris, Prasenjit Mitra, and Roald Eiselen. 2026b. Can embedding similarity predict cross-lingual transfer? a systematic study on african languages.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293.

Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, St. Julian's, Malta. Association for Computational Linguistics.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André F. T. Martins, and Hinrich Schütze. 2023. mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Proceedings of ACL*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Belber, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alípio Jorge, Felermino Ali, Chester Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Yamusi, Hailu Bekele, Emran Gebremichael, Nathnaiel Yohannes, and Aster Setotaw. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5319–5336, Singapore. Association for Computational Linguistics.

Muhammad Umair Nasir and Innocent Amos Mchechesi. 2022. Geographical distance is the new hyperparameter: A case study of finding the optimal pre-trained language for English-isiZulu machine translation. In *Proceedings of EMNLP*.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Odunayo Ogundepo, David Ifeoluwa Adelani, Akintunde Oladipo, Dietrich Klakow, and Jimmy Lin. 2023. AfriQA: Cross-lingual open-retrieval question answering for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11867–11882, Singapore. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data and model architecture on ranking transfer languages for cross-lingual pos tagging. *arXiv preprint arXiv:2503.19979*.

Harish Thangaraj, Ananya Chenat, Jivat Walia, and Vukosi Marivate. 2024. Cross-lingual transfer of multilingual models on low resource African languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Genta Indra Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of AACL-IJCNLP*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.