Typology-Guided Adaptation in Multilingual Models

Ndapa Nakashole[1,2]
[1]University of California, San Diego
[2]Okalai AI
nnakashole@ucsd.edu

Abstract

Multilingual models often treat language diversity as a problem of data imbalance, overlooking structural variation. We introduce the *Morphological Index* (MoI), a typologically grounded metric that quantifies how strongly a language relies on surface morphology for noun classification. Building on MoI, we propose *MoI-MoE*, a Mixture of Experts model that routes inputs based on morphological structure. Evaluated on 10 Bantu languages—a large, morphologically rich and underrepresented family—MoI-MoE outperforms strong baselines, improving Swahili accuracy by 14 points on noun class recognition while maintaining performance on morphology-rich languages like Zulu. These findings highlight typological structure as a practical and interpretable signal for multilingual model adaptation.

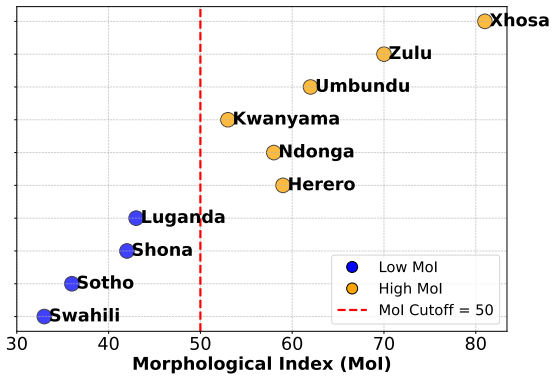

Figure 1: **Morphological variation in Bantu languages.** The Morphological Index (MoI) quantifies noun class prefix richness. Languages above the cutoff (orange) show stronger morphological marking; those below (blue) rely more on semantics. This distinction informs our typology-driven model adaptation.

1 Introduction

Multilingual models have largely progressed through monolithic transformers such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021), which share parameters across typologically diverse languages. While effective on high-resource languages, these models often underperform on African and other low-resource languages—not solely due to data scarcity, but due to *capacity dilution*: the inclusion of typologically distant languages forces representational resources to be spread too thin (Conneau et al., 2020).

Recent models like AfriBERTa (Ogueji et al., 2021) and BantuBERTa (Ogunremi et al., 2023) improve performance by grouping African languages genealogically, favoring lineage-based specialization over one-size-fits-all multilingualism. Yet genetic ancestry alone is often too coarse to reflect key functional differences. For example, Swahili and Zulu are both Bantu languages, but they diverge in their surface morphological structure. This raises a question: *What if multilingual models were grouped not by lineage, but by linguistic structure?* Linguistic typology—the systematic study of structural variation across the world's languages (Comrie, 1989; Croft, 2003)—offers a more precise lens for modeling this diversity. Typologically informed modeling has shown promise in several NLP tasks (Bender, 2016; O'Horan et al., 2016; Cotterell and Eisner, 2018; Ponti et al., 2019; Üstün et al., 2022).

Bantu languages, comprising over 300 languages spoken across sub-Saharan Africa (Maho, 1999), provide a rich testbed due to their typological diversity. Bantu grammar is characterized by a *noun class system*, where nouns are grouped into 10–20 classes, each marked by a distinct prefix (Guthrie, 1967). While reminiscent of gender systems, Bantu noun classes encode a broader array of semantic features, such as animacy, shape, and size.

Critically, the degree to which noun classes are

marked morphologically varies across Bantu languages. Some, like Zulu, use long, distinct prefixes that strongly signal class membership. Others, like Swahili, use shorter or fewer forms, requiring more semantic context. This variation affects model behavior: transformer models rely more on morphology in high-morphology languages, and more on semantics in lower-morphology languages (§6).

To capture this structural signal, we introduce the *Morphological Index (MoI)*, a continuous, interpretable metric that quantifies how much a language relies on morphology for noun classification (§2). We then use MoI to guide model adaptation. Specifically, we design **MoI-MoE**, a Mixture of Experts (MoE) model (Jacobs et al., 1991; Shazeer et al., 2017) in which each expert specializes in a different morphological regime. A lightweight, non-learned router uses MoI to select the appropriate expert, enabling typology-driven inference.

Contributions

1. We introduce the *Morphological Index (MoI)*, a typologically grounded metric that quantifies a language's morphological salience, enabling structure-aware adaptation (§2).
2. We define *Noun Class Recognition (NCR)* as a core task for evaluating structural variation, and release two datasets covering 65K nouns across 10 Bantu languages, including five with limited prior NLP coverage (§3, §4). Resources are available at `https://github.com/okalai-ai/moimoe`.
3. We develop *MoI-MoE*, a typology-aware MoE model that routes inputs by morphological structure, and show that MoI-MoE amplifies semantic features in low-morphology languages while maintaining accuracy in high-morphology ones. (§5).
4. We validate MoI via three independent analyses: (1) correlation with rule-based accuracy ($r = 0.75$), (2) interaction with semantic vs. morphological features ($r = -0.736$), and (3) performance of MoI-based expert routing (§6).

MoI complements modular approaches such as adapters and LoRA (Pfeiffer et al., 2021; Hu et al., 2022) by offering a typologically grounded criterion for parameter sharing, promoting transfer among structurally aligned languages and reducing cross-linguistic interference.

Most prior work treats typology as metadata, via language vectors or coarse typological features (Littell et al., 2017; Üstün et al., 2022); we treat it as an architectural design principle.

English Word	Bantu Cognates
snake	Sotho: *noha*, Ndonga: *eyoka*, Swahili: *nyoka*, Zulu: *inyoka*, Umbundu: *onyoha*
finger	Herero: *ominwe*, Ndonga: *omunwe*, Shona: *munwe*
leopard	Sotho: *nkwe*, Xhosa: *ingwe*, Zulu: *ingwe*, Kwanyama: *ongwe*
kidney	Sotho: *phio*, Kwanyama: *ofiyo*, Swahili: *figo*, Shona: *itsvo*

Table 1: **Lexical similarity, morphological divergence.** Cognates for basic concepts across Bantu languages often share phonological roots but differ in noun class morphology, highlighting the need for structure-aware modeling via MoI.

2 Morphological Index (MoI)

Phonological similarity does not equal structural similarity. As Table 1 shows, many Bantu languages share cognates for basic concepts like *snake* or *finger*, yet realize them with different morphological patterns, especially in their noun class prefixes. These surface overlaps mask variation in how structure is encoded. This variation is not just linguistic, it affects how models generalize, and what features they rely on to make predictions. We ask: in multilingual models, can linguistic structure guide model design?

Noun Class Structure and Variation Bantu nouns consist of a *prefix* and a *stem*, as illustrated in Figure 2. The prefix is a short morpheme that signals the *noun class*—a grammatical category grouping nouns by features such as animacy, size, or shape. For example, humans typically fall into Class 1/2 (singular/plural), animals into Class 9/10, and liquids into Class 5/6. Yet cultural nuance can override these tendencies: in Umbundu and Zulu, the words for *criminal* [1] are placed in Class 9, normally reserved for animals, reflecting a view of criminals as "lacking humanity" (Buthelezi, 2008).

While the overall structure of noun classes is shared across Bantu languages, the number of active

[1] Criminal is *ondingavivi* in Umbundu , *inswelaboya* in Zulu.

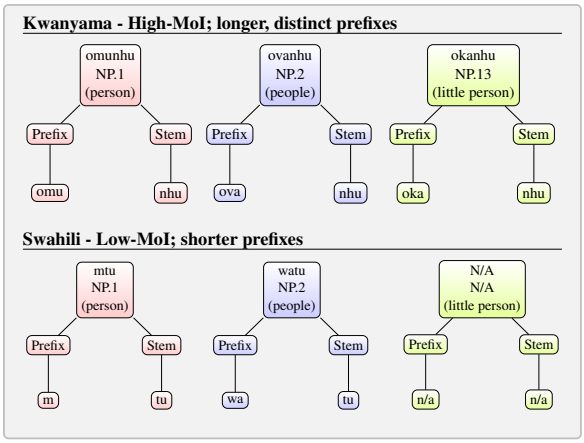

Figure 2: **Morphological Index (MoI) in Bantu noun structure.** High-MoI languages like Kwanyama use longer, distinct prefixes that strongly signal noun class. Low-MoI languages like Swahili use shorter, reused prefixes, making classification more reliant on semantics. Some classes are missing entirely in Low-MoI languages (e.g., Swahili lacks Class 13). Across all languages, prefixes carry no meaning by themselves, they derive semantics only in combination with stems.

classes and the distinctiveness of their prefixes vary considerably. The Bleek-Meinhof system (Bleek, 1851, 1869; Meinhof, 1899, 1932) remains the most widely used framework, assigning numbered classes (1–23) to enable cross-linguistic comparison. Most modern Bantu languages use only a subset (e.g., Swahili: 18, Shona: 20).

Despite this variation, alignment across languages remains strong. For example, Class 7 in Swahili (*ki-*) corresponds to Class 7 in Zulu (*isi-*), Shona (*chi-*), and Kwanyama (*oshi-*): a pattern reflected even in the languages' names: Kiswahili, IsiZulu, ChiShona, Oshikwanyama.

Quantifying Morphological Dependence The Morphological Index (MoI) quantifies how explicitly noun class distinctions are marked in a language's surface morphology. High-MoI languages use longer, more distinctive prefixes that reliably signal class; Low-MoI languages use shorter or overlapping prefixes, shifting more of the classification burden to semantics. MoI captures this gradient as a continuous score—not a binary split—offering a fine-grained measure of structural reliance on morphology. Specifically, MoI is computed as the total number of characters used across a language's noun class prefixes:

$$\text{MoI}_l = \sum_{c \in \mathcal{C}_l} |p_c|$$

Class	Swahili	Xhosa
1	m-	um-
2	wa-	aba-, abe-
3	m-	um-
7	ki-	isi-
10	N-	iiN-, iziN-

Table 2: **Sample noun class prefixes for Swahili and Xhosa.** These prefixes determine agreement across verbs, adjectives, and more, shaping much of the morphosyntactic structure.

where $\mathcal{C}_l$ is the set of Bleek-Meinhof noun classes in language l, and $|p_c|$ is the character length of the prefix for class c.

An alternative formulation is the average prefix length:

$$\text{MoI}_l = \frac{1}{|\mathcal{C}_l|} \sum_{c \in \mathcal{C}_l} |p_c|$$

While more interpretable, the average form is less effective at separating High- vs. Low-MoI languages in practice. We use the total length as our main metric, as it better captures the cumulative structural signal, a finding supported by empirical trends in Figure 1 and later analyses (§6).

MoI requires only a list of noun class prefixes, which are well-documented in grammar books and comparative studies of Bantu languages (Guthrie, 1967; Maho, 1999). Table 2 presents an abridged set of prefixes for two languages, illustrating the basis for computing MoI. Full prefix tables for all languages in our study are provided in Table 8 in Appendix A.

Why MoI Matters for NLP Prior typology-aware NLP methods often ask: *What kind of language is this, in the abstract?*—using coarse, static descriptors like SVO order, agglutinativity, or noun class presence (Dryer and Haspelmath, 2013; Moran et al., 2014; Hammarström et al., 2015; Lewis et al., 2015; Littell et al., 2017). But such features are too sparse and high-level to capture the fine-grained structural variation found even within language families like Bantu.

Unlike coarse typological metadata in databases such as WALS (Dryer and Haspelmath, 2013), MoI derives a language's structural profile from surface morphological patterns. Figure 3 shows WALS' global noun class coverage, with notable concentration in Africa. However, its limited granularity and sparse coverage for Bantu languages reduce its practical utility. MoI fills this gap by leveraging re-

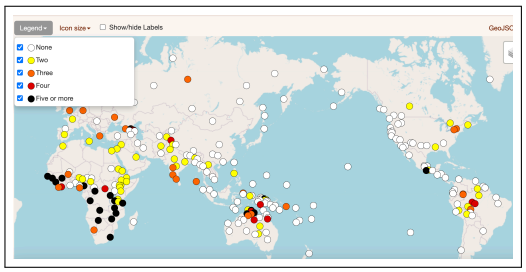

Figure 3: **Global distribution of noun class systems.** Derived from WALS (Dryer and Haspelmath, 2013), darker regions indicate more noun classes. While these systems are common in Africa, Asia, and the Americas, WALS provides limited coverage of Bantu languages, motivating alternatives like MoI.

	DictionaryNCR			WikiNCR
	Train	Val.	Test	
Swahili (swa)	20,603	626	1,500	1,297
Ndonga (ndo)	14,705	362	1,500	–
Xhosa (xho)	8,737	204	1,000	898
Zulu (zul)	3,303	123	700	502
Luganda (lug)	3,067	117	700	–
Shona (sna)	2,457	130	700	–
Kwanyama (kua)	958	82	300	–
Umbundu (umb)	616	54	300	–
Total	**54,446**	**1,698**	**6,700**	**2,697**
Herero (her)	–	–	403	–
Sotho (sot)	–	–	284	–

Table 3: **DictionaryNCR and WikiNCR dataset statistics.** Each language name is followed by its ISO 639-3 code. DictionaryNCR is used for model training and development; WikiNCR derived from Wiktionary, is used for out-of-domain evaluation.

sources like grammars and dictionaries to produce a continuous, morphology-grounded metric.

3 Noun Class Recognition (NCR)

In English agreement is mostly governed by number (e.g., *cat is* vs. *cats are*). Bantu languages encode agreement via noun classes. Thus, predicting the correct class is critical to grammaticality, making Noun class recognition (NCR) a natural testbed for structure-sensitive modeling in Bantu languages.

We define NCR as a classification task:

$$f(n) \rightarrow c, \quad c \in C$$

where f maps a noun n to its correct class c, drawn from a language-specific inventory.

NCR is not just a linguistic curiosity; it has concrete value for NLP tasks such as POS tagging, dependency parsing, and text generation. To demonstrate its syntactic relevance, we integrate NCR into Swahili POS tagging by replacing the generic "NP" tag with class-specific labels (e.g., "NP.1", "NP.2") using the MasakhaPOS dataset (Dione et al., 2023), in our experiments. This enables models to learn agreement patterns that hinge on noun class identity.[2]

To illustrate why agreement matters, consider Kwanyama, a Bantu language spoken in Namibia and Angola. In this language, the verb "eat" (*lya*) must agree with the noun class of its subject. The prefix governs both the noun form and the appropriate verb conjugation:

- *omunhu (NP.1) okwa lya* — the person ate
- *ovanhu (NP.2) ova lya* — the people ate
- *oshikombo (NP.7) osha lya* — the goat ate

These patterns are not idiosyncratic—they are core to Bantu grammar. By modeling NCR as a prediction task, we enable models to capture such structure-sensitive dependencies, improving syntactic accuracy in downstream tasks.

Finally, during dataset construction, we observed that some Bantu dictionaries lack noun class labels. NCR models offer a scalable way to fill this gap, assisting both NLP applications and lexicographic annotation efforts.

4 Datasets

Existing computational resources for Bantu noun class systems are sparse. Prior work has operated with fewer than 100 annotated examples per language (Byamugisha et al., 2018). To support large-scale evaluation and typologically informed modeling, we construct two datasets for noun class recognition (NCR). Each entry includes a Bantu noun, its class label, an English gloss, and a GPT-4–expanded definition. **DictionaryNCR** (63K nouns, 10 languages) is our primary dataset, **WikiNCR** (2.6K nouns, 3 languages) serves as an out-of-domain test set for evaluating generalization. Dataset statistics are summarized in Table 3.

[2]Like POS tags in English dictionaries, noun classes can be assigned in isolation (NCR without sentence context), and many Bantu dictionaries include such labels. However, just as with POS tagging, context often disambiguates class membership. In this paper, we focus on the context-free setting, while demonstrating a contextualized variant via Swahili POS tagging.

DictionaryNCR This dataset is built from bilingual dictionaries and grammar books obtained via Google Books,[3] the Library of Congress,[4] and university archives. Most sources were scanned or in PDF form, requiring OCR (via MathPix) followed by parsing and structuring via GPT-4. Manual inspection of 200 samples across two languages revealed no hallucinated content.

We distinguish two source types (see Figure 4):

- **Type 1 (Explicit class labels)**: Dictionaries for Swahili, Kwanyama, Shona, Luganda, and Sotho explicitly annotate noun classes. These entries (31,524 of 62,844) required only extraction and formatting.
- **Type 2 (Inferred class labels)**: Type 2 sources, provide nouns labeled with prefixes, but lack explicit class labels. This was the case for Ndonga, Umbundu, Xhosa, Zulu, and Herero. For these languages, we inferred class labels from prefixes using grammar-derived mappings. Ambiguous cases were resolved heuristically via prefix patterns and glosses.

WikiNCR This dataset is sourced from Wiktionary,[5] where certain languages provide explicit noun class labels (e.g., `https://en.wiktionary.org/wiki/Category:Swahili_nouns_by_class`). These HTML entries were parsed using BeautifulSoup. While Wiktionary coverage is limited, this dataset offers a complementary testbed for out-of-domain evaluation, a key challenge in robust NLP (Jia and Liang, 2017; Ribeiro et al., 2020).

Noun class distributions are shown in Table 9 for DictionaryNCR, and in Table 10 for WikiNCR, Appendix A.

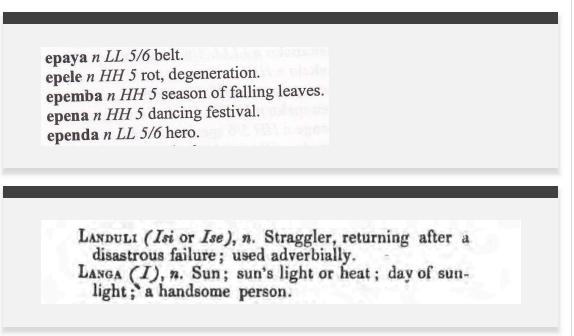

Figure 4: **Sources of noun class annotations used to create the DictionaryNCR dataset**. (**Top**) Type 1 sources, such as this Kwanyama grammar book, explicitly list noun classes. (**Bottom**) Type 2 sources, like this Zulu dictionary sourced from Google Books, provide nouns labeled with prefixes, but lack explicit class labels.

5 Methodology

We build a multilingual model for Noun Class Recognition (NCR) that puts MoI to work—not just as a typological insight, but as a design principle. Our architecture draws on two key signals that shape Bantu noun systems: morphology (prefix patterns) and semantics (conceptual meaning).

Morphological Features: Subword Tokens Morphology offers powerful features for noun class prediction, so much so that no Bantu grammar book is complete without a table of noun class prefixes. These prefixes act as morphosyntactic markers of class membership, yet their surface forms differ widely across languages (see Table 8 in Appendix A). Instead of hand-coding prefix rules, we rely on SentencePiece tokenization (Kudo and Richardson, 2018), enabling the model to discover subword patterns that correlate with noun classes, no explicit supervision needed.

Semantic Features: English Glosses Bantu noun classes exhibit broad semantic regularities: humans often fall into Class 1/2, plants into Class 3/4, and large animals into Class 9/10 (see Table 4). While these patterns are well-documented, their predictive utility has rarely been quantified at scale. But how do we put meaning to work in languages where the model has limited—or no—pretraining exposure? Large language models do encode semantic knowledge, but their coverage of Bantu languages is sparse. Feeding Bantu nouns directly into such models yields poor representations. To bridge this gap, we pair each Bantu noun with its English gloss during training, allowing the model to ground predictions in a shared semantic space through pretrained English embeddings. This not only enhances generalization, but also allows us to empirically test when and where semantic features truly matter.

Supervised Fine-Tuning To train our model to predict noun classes using both morphological and semantic features, we fine-tune `mT5-XL` (3.7B parameters) on paired Bantu nouns and their English

[3] `https://books.google.com/`
[4] `https://www.loc.gov/`
[5] `https://www.wiktionary.org/`

Class	Generalized Semantics
1/2	Humans
3/4	Plants, trees
5/6	Paired objects, liquids, masses
7/8	Inanimates, diminutives
9/10	Animals, some inanimates
14	Abstracts, mass nouns
15	Infinitives
16–18	Locatives

Table 4: **Semantics in Bantu noun classification.** Many noun classes map to broad semantic categories. This mapping supports the use of English glosses as a semantic proxy in our model. See full list in Table 7 (Appendix A).

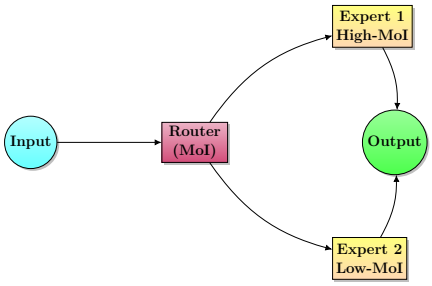

Figure 5: **MoI-MoE: Typology-Aware Expert Routing for NCR.** MoI routes inputs to specialized experts: morphology-driven or semantics-driven.

glosses. The prompt takes the form: *Given the Bantu noun [X] whose English meaning is [Y], predict its noun class*, where [X] is the noun and [Y] is its English gloss with an expanded definition. This setup allows the model to align surface form with meaning in a controlled way.

Interestingly, we found that specifying the Bantu language in the prompt hurt performance. Possibly because it causes the model to overfit to surface features, rather than generalizing across typologically similar languages. The model is trained to predict a noun class label (e.g., "1", "15") using a standard sequence-to-sequence loss:

$$\mathcal{L}_{\text{SFT}}(f(x), y; \theta) = -\sum_{k=1}^{|y|} \log P(y_k \mid y_{<k}, f(x); \theta),$$

where θ are model parameters and y_k denotes the k-th output token.

To prevent overfitting to high-resource languages, we apply balanced language sampling during training (see Appendix B).

MoI-MoE: Typology-Driven Adaptation via Experts MoI serves not only as a diagnostic signal, but as a structural guide for model design. We introduce **MoI-MoE**, a Mixture of Experts (MoE) model (Jacobs et al., 1991; Shazeer et al., 2017), in which each expert is specialized for a distinct morphological regime. A lightweight, non-parametric router uses MoI scores to assign each input to the most appropriate expert, enabling dynamic, structure-aware adaptation at inference time (Figure 5).

We define two experts:

- **High-MoI Expert:** For languages where subword token patterns, approximating prefixes and stems, provide strong morphological features.
- **Low-MoI Expert:** For languages where prefix forms are less informative and semantic context plays a greater role.

Unlike traditional language-specific routing, MoI-based routing generalizes across structurally similar languages, even when they differ in vocabulary and share few subword tokens.

6 Experiments

Our experiments evaluate MoI as both a predictive measure of typological variation, and a structural signal for guiding model adaptation. We fine-tune `mT5-XL` (3.7B parameters) using a learning rate of 5×10^{-5}, batch size of 16, and mixed-precision (FP16) training. All experiments are run on four NVIDIA A40 GPUs and complete within 5 hours.

To separate High- and Low-MoI languages, we apply a MoI cutoff of 50.0 based on validation data.

We compare a range of models: a rule-based classifier (**RULES**) using fixed prefix-to-class mappings; a no-subword model (**VOCAB**) trained on full tokens; a **PMASK** variant where prefixes are masked to probe reliance on morphology; a semantics-only model (**SEM**) using English glosses as input but no Bantu words; a morphology-only model (**MORPH**) using Bantu words without English glosses; and the full model (**NCR-θ_0**) trained with both morphology and semantics. In addition, we evaluate two MoE architectures: **TO-MoE**, with routing based on token overlap between languages, and our proposed **MoI-MoE**, which uses MoI as a typologically informed routing signal.

		Languages								Avg
		Kwanyama	Luganda	Ndonga	Shona	Swahili	Umbundu	Xhosa	Zulu	
Rule-Based	RULES	86.33	45.13	90.33	54.29	48.73	87.50	75.22	80.00	71.20
	VOCAB	29.00	23.12	24.11	28.86	24.68	21.50	22.56	19.25	24.13
	PMASK	33.33	54.10	20.12	21.32	31.29	22.18	50.00	44.51	34.61
	SEM	39.33	36.77	62.82	38.76	63.34	30.33	42.34	48.07	45.22
	MORPH	94.67	77.70	96.62	88.84	80.07	84.98	88.25	**92.78**	87.99
	NCR-θ_0	91.33	80.29	95.40	95.14	93.27	**88.00**	87.32	88.57	89.91
	Δ SEM	-3.34	2.59	-1.38	6.30	13.20	3.02	-0.93	-4.21	1.92
	TO-MoE (t.r.)	**95.00**	46.91	90.58	45.49	94.13	87.37	86.80	92.48	79.8
	MoI-MoE	**95.00**	**84.57**	**97.37**	**95.57**	**94.80**	87.37	**88.67**	92.48	**92.00**

Table 5: **Accuracy on Bantu Noun Class Recognition (NCR)** across eight languages. MoI-MoE outperforms all baselines. ΔSEM shows the performance gap between a morphology-only model (MORPH) and the full model (NCR-θ_0), capturing the added value of semantic features. TO-MoE uses a token-overlap-based router; MoI-MoE routes by MoI.

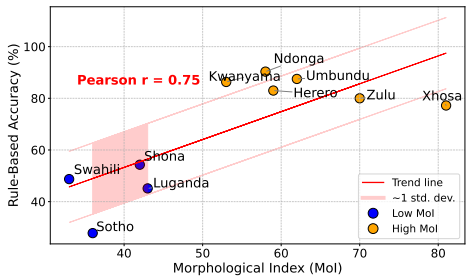

Figure 6: Correlation between MoI and rule-based accuracy. High-MoI languages (blue dots) show stronger rule-based NCR (orange dots). Pearson $r = 0.75$ indicates a strong positive correlation.

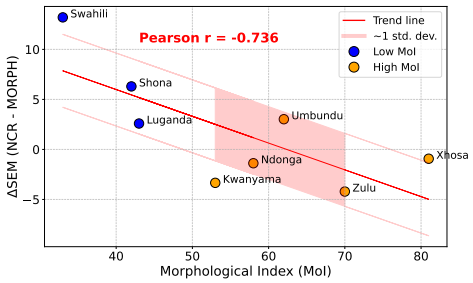

Figure 7: **Correlation between MoI and ΔSEM.** High-MoI languages rely less on semantic features (lower ΔSEM). Blue dots indicate Low-MoI languages, orange dots indicate High-MoI. Pearson $r = -0.74$ reflects a strong negative correlation, showing that as morphological richness increases, the benefit of adding semantics decreases.

6.1 Main Results

Table 5 reports accuracy on the NCR task across eight Bantu languages. With 22 noun classes, random accuracy is just 4.5%. MoI-MoE achieves the highest overall performance (92.0%), outperforming both the token-overlap MoE baseline (TO-MoE: 79.8%) and a monolithic multilingual model (89.91%). Crucially, MoI-based routing enables targeted integration of semantic features in low-MoI languages. On Swahili, a structurally light language, MoI-MoE improves over the morphology-only model by +14.7 points (94.80% vs. 80.07%). On Zulu, a high-MoI language, the models perform nearly identically (92.48% vs. 92.78%), suggesting that rich morphological features alone are sufficient.

In what follows, we analyze the MoI signal in detail, addressing three key questions: (i) Does MoI predict the effectiveness of rule-based classifiers? (ii) Does it capture where semantic features supplement morphological ones? (iii) Can it guide expert routing in a mixture-of-experts model?

6.1.1 Analysis #1: Rule-Based Accuracy Tracks MoI

If MoI captures morphological salience, then rule-based classifiers, relying purely on prefixes, should perform better in High-MoI languages. Figure 6 confirms this: rule-based accuracy correlates strongly with MoI ($r = 0.75$), suggesting that in languages with richer morphology, morphology alone is often sufficient for accurate classification.

6.1.2 Analysis #2: Semantic Gain Tracks Morphological Weakness

If morphology is weak, semantics must carry some of the load. To test this, we measure ΔSEM—the accuracy gain from adding semantic features to a morphology-only model:

$$\Delta\text{SEM} = \text{Acc}_{\text{NCR-}\theta_0} - \text{Acc}_{\text{MORPH}}$$

As shown in Figure 7, ΔSEM is negatively correlated with MoI ($r = -0.736$): High-MoI languages (e.g., Ndonga, Zulu) show little to no benefit from semantics, while Low-MoI languages (e.g., Swahili) rely on them heavily. This pattern strengthens the case for MoI as a structural signal that captures where meaning must supplement form.

6.1.3 Analysis #3: Typological-Guided Adaptation

If MoI captures typologically meaningful distinctions, it should support effective expert selection in Mixture-of-Experts (MoE) models. **MoI-MoE** achieves the highest overall accuracy (92%), significantly outperforming **TO-MoE** (80%), which relies on token overlap for routing. TO-MoE misroutes languages with shared morphology but low lexical similarity. In contrast, MoI enables High-MoI languages to benefit from morphology-driven processing, while Low-MoI languages are routed to experts that integrate semantic features.

6.1.4 Main Results Summary

Together, the three core analyses: (i) rule-based accuracy, (ii) morphology vs. semantics sensitivity, and (iii) MoI-guided expert routing, offer converging evidence that MoI is a robust, linguistically grounded measure of structural variation in Bantu languages, with clear utility for model adaptation.

6.2 Ablating Morphology

MoI measures structural reliance on morphology, but which morphological signals are actually used by the model? We probe this by ablating key components: subword tokenization and noun class prefixes. Without subword tokenization, (**VOCAB**), the model is forced to operate on full words. Performance drops across the board (average 24.13% vs 87.99%), confirming that segmenting words into morpheme-like units is essential for learning morphological distinctions. Masking noun class prefixes (**PMASK**) also causes substantial degradation (34.61% vs 87.99%). This confirms that prefixes are not only essential to model predictions but also justify their centrality in MoI computation.

6.3 Probing the MoI Signal

While MoI is a continuous, data-driven metric, one might ask whether it merely recovers familiar linguistic features, such as, the presence of augments. Augments are vowel-like morphemes (Blois, 1970; Van der Wal and Lusekelo, 2022) that precede the noun class prefix in some Bantu languages (e.g., Zulu's *umu-*, where *u-* is the augment and *mu-* the prefix). They are common in High-MoI languages and largely absent in Low-MoI ones, making them a plausible driver of MoI values.

To test this, we recomputed the correlations in Figure 6 and Figure 7 after removing augment segments from all prefixes. While the correlations weakened, they did not collapse—indicating that MoI captures structural variation beyond just augment presence. Model performance using this "augmentless" MoI also remained high (91.11%), only slightly below the original 92%. This suggests that while augments are a key component, MoI encodes a broader typological signal.

6.4 Error Analysis

MoI-MoE performs strongly, with over 90% accuracy on several languages, but where do the remaining errors come from? Table 6 shows they are not random: they track structural ambiguity.

In High-MoI languages, a leading cause of the remaining errors (58%) is overlapping prefixes. Zulu's *um(u)-* marks both Class 1 (humans) and Class 3 (trees). These overlaps are less disruptive in Low-MoI languages where the semantics-aware expert can disambiguate. Yet, reintroducing semantic features into High-MoI languages tends to degrade performance, a future approach may require selective integration of semantics. In Low-MoI languages, errors are more semantically driven. For example, in Swahili, a common confusion is between Class 9 (animals) and Class 5 (miscellaneous), due to the ambiguous nasal (zero) prefix in Class 9. In both settings, the nature of the error reflects the underlying structure, further providing interpretability to the MoI-guided approach.

6.5 Generalization and Broader Utility

To understand the robustness and applicability of MoI-guided modeling, we evaluate its performance

Lang.	True → Pred.	Err.	Cause
Zulu	1 → 3	58.1%	Shared prefix
Ndonga	9 → 7	26.8%	Mistokenization
Swahili	9 → 5	14.3%	Nasal (zero) prefix

Table 6: Representative NCR errors, grouped by likely structural cause.

under domain shift, in zero-shot settings, and in a downstream syntactic task.

Cross-Dataset Evaluation First, cross-dataset evaluation shows that models trained on DictionaryNCR experience a notable drop when tested on WikiNCR: Swahili falls from 94.8% to 75.1%, Xhosa from 88.7% to 78.3%, and Zulu from 92.5% to 76.1%. This reflects overfitting to dictionary-style input. Unlike DictionaryNCR, WikiNCR includes more everyday terms, proper names, and locatives (Classes 16–18), which are underrepresented in grammar books but common in natural usage. These results mirror broader trends in NLP, where cross-domain generalization remains a persistent challenge.

Zero-Shot Generalization Second, MoI-based routing improves zero-shot generalization to unseen languages in comparison to the monolithic NCR model. Herero achieves 74.1% accuracy from MoI-MoE (vs. 71.9% monolithic NCR) potentially due to high subword overlap with training languages(see Figure 9), while Sotho has 60.6% (vs. 52.11% monolithic NCR) despite limited lexical similarity (Figure 9). Full results are provided in Appendix D.

Downstream Tasks Finally, noun class distinctions extend beyond NCR. We integrate NCR into Swahili POS tagging. Fine-tuning AfroXLMR-large(Alabi et al., 2022), which holds the best performance on MasakhaPOS(Dione et al., 2023), on an NCR expanded tagset results in only a 1.0% drop in accuracy (93.5% → 91.7%). Given the small dataset size (800 sentences), this performance drop is expected, as models struggle to generalize over a more granular tagset with limited supervision. See Appendix E for details.

7 Related Work

Bantu noun class systems are well-studied in linguistics (Guthrie, 1967; Maho, 1999), but computational approaches remain limited. Prior work focuses on rule-based pluralization (Byamugisha et al., 2018) and heuristics (Reid et al., 2021), often lacking scale or reusability.

Typological Features in Multilingual Model Several typology-aware approaches rely on databases like WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran et al., 2014), Glottolog (Hammarström et al., 2015), and Ethnologue (Lewis et al., 2015), which provide language-level features such as word order, phoneme inventories, or case marking. Tools like lang2vec (Littell et al., 2017) package these into language vectors that can be used to guide parameter sharing, adapter conditioning, or zero-shot inference. However, these features are often too coarse, and sparsely populated to distinguish typologically similar languages such as Bantu languages.

Mixture of Experts: From Scale to Structure Mixture of Experts (MoE) models scale transformers via sparse activation, routing each input to a subset of experts (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022). While effective for efficiency, these models rarely incorporate linguistic structure; routing is typically learned via token-level gating without typological grounding. Our MoI-MoE model introduces a typologically interpretable routing mechanism based on MoI, bridging structural variation and expert selection to improve multilingual generalization.

8 Conclusion

We introduce the Morphological Index (MoI), a typologically grounded metric that captures structural variation in Bantu languages. By linking linguistic structure to transformer model behavior, MoI enables interpretable, morphology-aware adaptation. Our MoI-MoE model outperforms monolithic and token overlap-based baselines on noun class recognition, demonstrating that structural typology can serve as an effective basis for model design. While our focus is on Bantu, this work points toward a broader approach: one where multilingual NLP is informed not only by token overlap or genetic lineage, but by the structural logic that shapes human language.

Limitations

Our Mixture of Experts (MoE) model improves performance by adapting to linguistic typology, but it relies on hard routing based on MoI, which may not fully account for intermediate cases. A soft routing strategy could enhance flexibility. Additionally, our datasets, while large, cover only 10 Bantu languages out of over 500. Expanding to more languages remains crucial but challenging due to limited digital resources and scarce structured annotations. Addressing this gap will require community-driven data collection, language archives, and corpus-based approaches beyond dictionary-derived data.

While MoI provides a useful typological signal, it simplifies linguistic variation by focusing solely on noun class prefix statistics. Some Bantu languages exhibit irregular noun class behavior due to phonological changes, loanwords, or diachronic shifts that MoI does not capture.

Beyond noun class recognition, can MoI guide other NLP tasks such as language modeling, parsing, or machine translation? These open questions were not explored in this work but present promising directions for future research.

However, we focused on Transformer-based models due to their current popularity and effectiveness in NLP tasks. Future work should explore the applicability of MoI to other architectures and tasks.

Potential Risks

Our work carries risks related to linguistic representation. We analyze structural variation across 10 Bantu languages, but there are over 500, and our findings may not generalize universally. Although the Morphological Index (MoI) improves model adaptability, rigid typological categorizations could oversimplify linguistic diversity particularly in languages with irregular noun class behavior or diachronic shifts.

Acknowledgements

The author expresses gratitude to David Nakashole, who helped shape the path that made this work possible.

The author also thanks Sharon Rose, and the anonymous reviewers for their constructive feedback on an earlier draft of this paper.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 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.

Emily M Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Wilhelm Heinrich Immanuel Bleek. 1851. *De nominum generibus linguarum Africae Australis, Copticae, Semiticarum aliarumque sexualium*. Dissertation, Universität zu Bonn.

Wilhelm Heinrich Immanuel Bleek. 1869. *A Comparative Grammar of South African Languages. Part II: The Concord; Section I: The Noun*. J. C. Juta and Trübner & Co., Cape Town and London.

Kornelis Frans de Blois. 1970. The augment in the bantu languages. *Africana linguistica*, 4(1):85–165.

Thabisile M Buthelezi. 2008. Exploring the role of conceptual blending in developing the extension of terminology in isizulu language. *Alternation*, 15(2):181–200.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing nouns across agglutinating Bantu languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2633–2643, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Ryan Cotterell and Jason Eisner. 2018. A deep generative model of vowel formant typology. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 37–46, New Orleans, Louisiana. Association for Computational Linguistics.

William Croft. 2003. *Typology and universals*. Cambridge University Press.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Matthew S Dryer and Martin Haspelmath. 2013. The world atlas of language structures online. leipzig: Max planck institute for evolutionary anthropology. *Online: http://wals.info*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Malcolm Guthrie. 1967. Comparative bantu. an introduction to comparative linguistics and the prehistory of the bantu languages.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. Glottolog 2.6. `https://glottolog.org`. Max Planck Institute for the Science of Human History, Jena.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the World*, eighteenth edition. SIL International, Dallas, Texas.

Patrick Littell, David R. 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.

Jouni Filip Maho. 1999. *A Comparative Study of Bantu Noun Classes*, volume 13 of *Orientalia et Africana Gothoburgensia*. Acta Universitatis Gothoburgensis, Göteborg. Doctoral dissertation, University of Gothenburg, November 1999.

Carl Meinhof. 1899. *Grundriss einer Lautlehre der Bantusprachen*. F. A. Brockhaus, Leipzig.

Carl Meinhof. 1932. *Introduction to the Phonology of the Bantu Languages*, revised, enlarged, and translated from german (grundriß einer lautlehre der bantusprachen, second edition, 1910) edition. Dietrich Reimer Verlag, Berlin. Reprinted 1984.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. PHOIBLE Online. `https://phoible.org`. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

Tolulope Ogunremi, Dan Jurafsky, and Christopher Manning. 2023. Mini but mighty: Efficient multilingual pretraining with linguistically-informed data selection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1251–1266, Dubrovnik, Croatia. Association for Computational Linguistics.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the*

European Chapter of the Association for Computational Linguistics: Main Volume, pages 487–503, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*. OpenReview.net.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022. UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592.

Jenneke Van der Wal and Amani Lusekelo. 2022. The v and cv augment and exhaustivity in kinyakyusa. *Studies in African Linguistics*, 51(2):323–345.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Noun Classes and Datasets - Additional Details

Class	Generalized Semantics
1	Singular for humans
1a	Kinship terms, personified beings
2	Plural of Class 1, honorific forms
2x	Plural of Class 1a, polite forms
3	Trees, plants, certain inanimates
4	Plural of Class 3
5	Paired objects, augmentatives, miscellaneous
6	Liquids, masses, collectives, plural for Classes 5, 9, 11, 14, and 15
7	Inanimates, styles, diminutives, augmentatives
8	Plural of Class 7
9	Animals, certain inanimate objects
10	Plural of Classes 9 and 11
11	Long/thin objects, abstracts
12	Diminutives
13	Plural of Class 12
14	Abstracts, mass nouns
15	Infinitives (verbal nouns)
16	Locatives (near, specific places)
17	Locatives (general, distant)
18	Locatives (interior, enclosed)
19	Diminutives, small objects
20	Augmentatives, diminutives
21	Pejoratives, augmentatives
22	Plural of Class 20
23	Locatives (unspecified category)

Table 7: Generalized semantic roles of Bantu noun classes. Most languages use a subset of 11 to 20 noun classes. From (Maho, 1999).

Noun Class	Languages									
	Sotho	**Herero**	**Umbundu**	**Zulu**	**Kwanyama**	**Swahili**	**Shona**	**Xhosa**	**Ndonga**	**Luganda**
1	mo-	omu-	omu-, u-, o-	um(u)-	omu-	m-	mu	um-	omu-	mu-
1a	∅	∅	-	u-	∅	N/A	∅	u-	∅	N/A
2	ba-	ova-	oma-, ova-, a-	aba-, abe-	ova-	wa-	va	aba-, abe-	aa-	ba-
2a	bo-	oo-	-	o-	oo-	N/A	va-,vana-,a-	oo-	oo-	N/A
3	mo-	omu-	u-	um(u)-	omu-	m-	mu	um-	omu-	mu-
4	me-	omi-	ovi-	imi-	omi-	mi-	mi	imi-	omi-	mi-
5	le-	e-	e-	i-, ili-	e-	ji-	ri	i-, ili-	e-	li-
6	ma-	oma-	a-, ova-	ama-	oma-	ma-	ma	ama-, ame-	oma-	ma-
7	se-	otji-	oci-	isi-	oshi-	ki-	chi	is(i)-	oshi-	ki-
8	di-	ovi-	ovi-	izi-	oi-	vi-	zvi	iz-, iz(i)-	ii-	bi-
9	∅	o-	o-, ∅	in-	o-	N-	i	iN-	o-	n-
10	di-	ozo(N)-	olo(N)	izin-	ee-	N-	dzi	iiN-, iziN-	oo-	n-
11	N/A	oru-	olu-	u-, ulu-	olu-	u-	ru	u-, ulu-, ulw-, ul-	olu-	lu-
12	N/A	oka-	oka-	N/A	oka-	N/A	ka	N/A	oka-	ka-
13	N/A	otu-	otu-	N/A	N/A	N/A	tu	N/A	omalu-	tu-
14	bo-	ou-	-	ubu-	ou-	u-	u	ubu-, ub-, uty-	uu-	bu-
15	ho-	oku-	oku-	uku-	oku-	ku-	ku	uku-	oku-	ku-
16	∅	pu-	pa-	N/A	po-	-ni / pa-	pa	N/A	po-, pu-	wa-
17	ho-	ku-	ko-	uku-	ko-	-ni / ku-	ku	N/A	ko-, ku-	ku-
18	mo-	mu-	vu-	N/A	mo-	-ni / m-	mu	N/A	mo-, mu-	mu-
19	N/A	N/A	N/A	N/A	N/A	N/A	svi	N/A	N/A	N/A
20	N/A	ku-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	gu-
21	N/A	N/A	N/A	N/A	N/A	N/A	zi	N/A	N/A	N/A
22	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	ga-
23	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	e-

Table 8: Noun class prefixes for the 10 Bantu languages in our study. While prefixes facilitate noun classification, they are not directly given to the model, which must learn their patterns via subword tokenization. We compute the Morphological Index (MoI) using Classes 1–15, represent core grammatical distinctions; higher classes (e.g., 16–18) are typically locative and less consistent across languages. Prefix tables were derived from grammars and dictionaries.

Class	Herero	Kwanyama	Luganda	Ndonga	Shona	Sotho	Swahili	Umbundu	Xhosa	Zulu
1	48	156	858	1045	454	58	1228	27	691	696
2	5	6	22	1216	0	0	1150	0	27	5
3	35	54	293	419	1024	30	1063	88	1177	52
4	2	0	13	426	6	0	1128	13	21	24
5	45	408	238	2527	0	84	2779	244	2834	822
6	34	31	52	2431	531	16	3289	0	174	88
7	40	163	660	1153	1254	60	1298	245	1164	866
8	9	32	109	1327	0	2	1263	0	38	32
9	72	264	927	2296	23	302	4090	201	845	666
10	0	24	34	1961	2	10	3977	15	3	20
11	27	0	50	175	0	0	532	96	1184	13
12	17	91	265	524	109	0	1	32	0	0
13	0	0	2	72	0	0	0	7	0	0
14	50	111	232	982	0	6	901	0	502	220
15	19	0	104	13	2	0	14	2	969	622
16	0	0	16	0	0	0	2	0	0	0
17	0	0	1	0	5	0	14	0	0	0
18	0	0	1	0	5	0	0	0	0	0
19	0	0	0	0	1	0	0	0	0	0
20	0	0	1	0	0	0	0	0	0	0
22	0	0	3	0	0	0	0	0	0	0
23	0	0	2	0	0	0	0	0	0	0

Table 9: Noun class distribution across 10 Bantu languages in the DictionaryNCR dataset. Higher-numbered classes are not present in all languages and, even when present, they tend to be rare.

Noun Class	Swahili	Xhosa	Zulu
1	205	105	95
3	198	147	84
4	1	2	0
5	124	137	79
6	194	5	7
7	197	156	77
8	11	0	0
9	133	158	70
10	8	0	0
11	223	109	41
14	0	63	29
15	0	15	16
16	3	0	0
17	0	1	4

Table 10: Noun class distribution in the WikiNCR dataset.

B Methodology - Additional Details

Balanced Language Sampling The languages in our datasets differ widely in data availability, risking overfitting to high-resource languages like Swahili while underperforming on low-resource ones. To mitigate data imbalance, we applied a balanced sampling strategy following prior work on multilingual training (Conneau and Lample, 2019; Xue et al., 2021), adjusting language contributions during training:
$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha}$, where $p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$ represents raw data proportions. Lower α values boost low-resource languages. We set $\alpha = 0.7$, which provided a good balance across languages. Figure 8 in Appendix A visualizes the effect of different α values on the DictionaryNCR dataset.

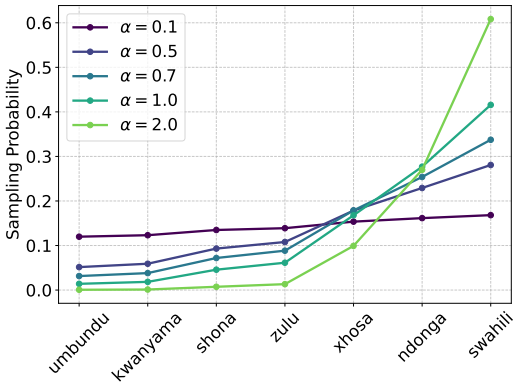

Figure 8: the behavior of the sampling strategy for different values of α on our DictionaryNCR dataset.

C Experiments: Token Overlap Matrix

Figure 9 shows the subword and full-token overlap matrices across the 10 Bantu languages in our study, as well as English. The subword matrix was used to inform the design of the TO-MoI baseline, which groups languages based on lexical similarity. The subword overlap reveals considerable but uneven token sharing among Bantu languages, while English remains highly divergent. The full-token overlap matrix shows substantially lower overlap overall, which shows that the languages generally do not share many full tokens.

D Experiments: Zero-Shot Inference

We further evaluate generalization to languages unseen during training. Table 11 shows that MoI-MoE outperforms the monolithic NCR model in zero-shot settings, correctly routing Low-MoI languages (e.g., Sotho) and High-MoI ones (e.g., Herero) to the appropriate expert.

While absolute performance is modest, MoI-based routing improves accuracy in both cases. Herero benefits from substantial subword overlap with training languages, while Sotho's lower overlap (Figure 9) limits transfer. This highlights a limitation of token-overlap approaches: structurally similar but lexically divergent languages are often misrouted. MoI, by contrast, captures structural similarity directly—providing a principled way to enable generalization even when surface forms differ.

Methods	NCR	MoE
Sotho	52.11	60.56
Herero	71.90	74.12

Table 11: Zero-Shot NCR results

E Experiments: A Preliminary Study on POS Tagging

Noun class distinctions are not limited to noun categorization—they also drive morphosyntactic agreement across verbs, adjectives, and pronouns. To explore this, we adapt the Swahili POS tagging dataset from Dione et al. (2023) by replacing the generic "NP" tag with class-specific labels (e.g., *NP.1*, *NP.2*, etc.).

Table 12 shows that fine-tuning AfroXLMR-large on this expanded tagset yields only a modest performance drop (from 93.5% to 91.7%), despite the increased label granularity and limited training data (800 sentences). This suggests that fine-grained noun class information can be integrated into downstream syntactic tasks without substantially sacrificing performance—opening the door to richer agreement modeling in African NLP.

Model	UD	UD+NC
CRF	89.3	-
mBERT	92.0	-
XLM-R-large	93.2	-
AfroXLMR-large	93.2	-
Ours	**93.5**	**91.7**

Table 12: Swahili POS tagging accuracy with and without noun class (NC) tags. UD: standard tagset. UD+NC: extended tagset with fine-grained noun classes.

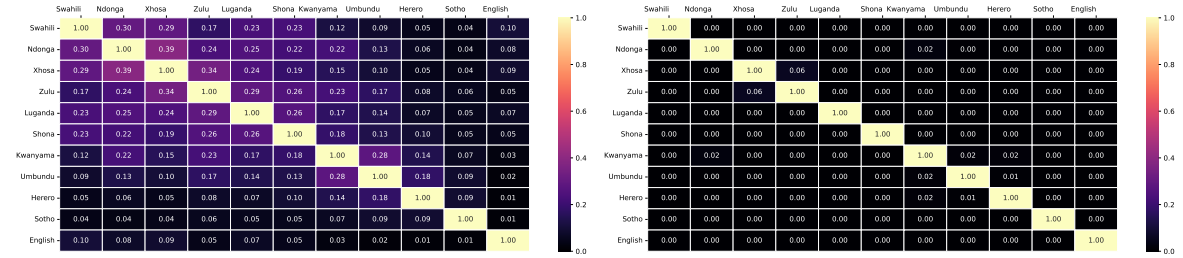

Figure 9: **Left**: Subword token overlap across Bantu languages and English, where lighter colors indicate greater overlap and darker colors highlight linguistic divergence. Bantu languages exhibit significant but uneven subword overlap, while English shows minimal overlap with Bantu languages. **Right**: Full-token overlap matrix, showing reduced overlap which, as shown in our experiments, results in significant performance drops in noun class prediction.

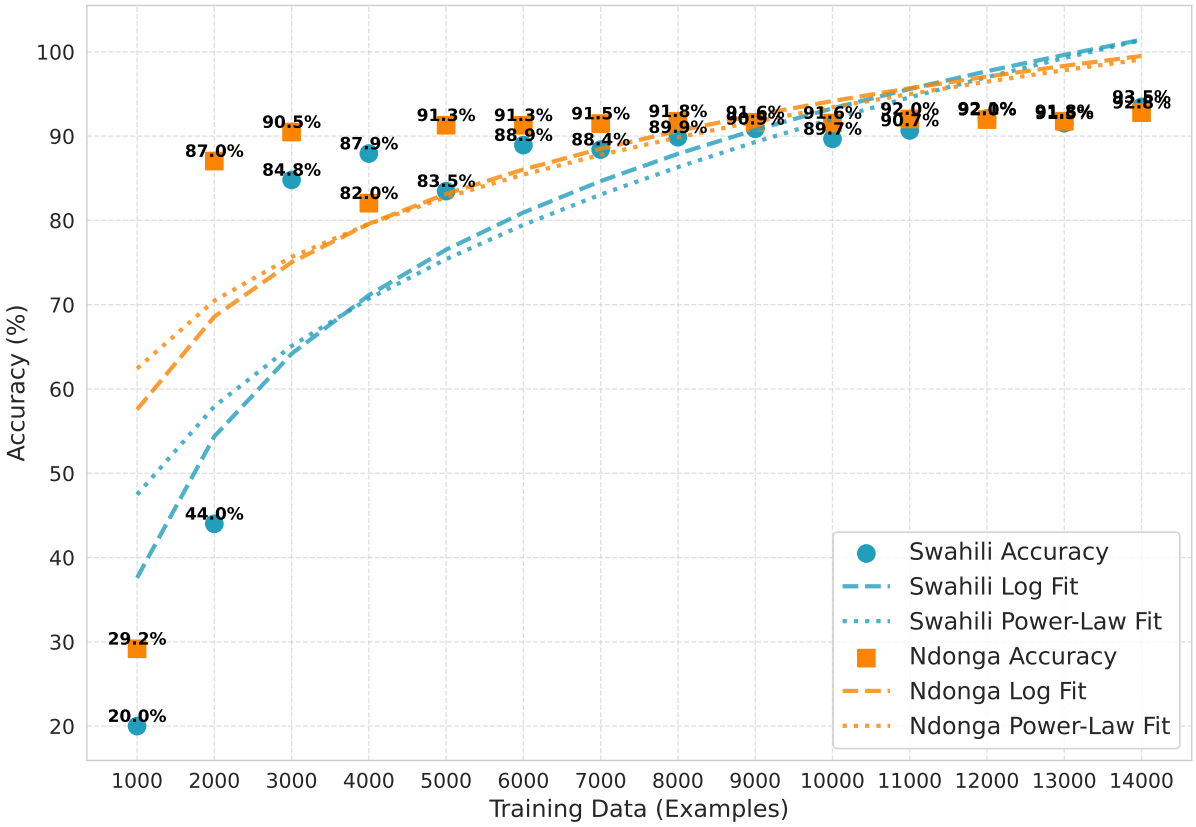

Figure 10: Data scaling trends for Swahili and Ndonga.

F Experiments: Scaling Data and Model Size

Data Scaling We analyze NCR scaling trends by evaluating accuracy growth in Swahili (Low-MoI) and Ndonga (High-MoI), the two most well-represented languages in our dataset. Figure 10 shows distinct learning curves: Ndonga reaches 90% accuracy with approximately 4K examples, while Swahili requires over 10K. Both follow log and power-law trends, demonstrating diminishing returns as data increases.

Final accuracy converges at 92–93%, but multilingual models achieve higher scores, suggesting that cross-linguistic transfer further enhances NCR. However, additional scaling experiments (Figure 11) are constrained by limited data.

Model Scaling To assess the impact of model size, we compare performance across different `mT5` variants. Accuracy improves from 62.2% with `mT5-Small` to 92.0% with `mT5-XL`, demonstrating clear benefits from larger model capacity. However, gains diminish beyond `mT5-Base`, with only a 3% increase from `mT5-Large` to `mT5-XL`, suggesting diminishing returns at larger scales.

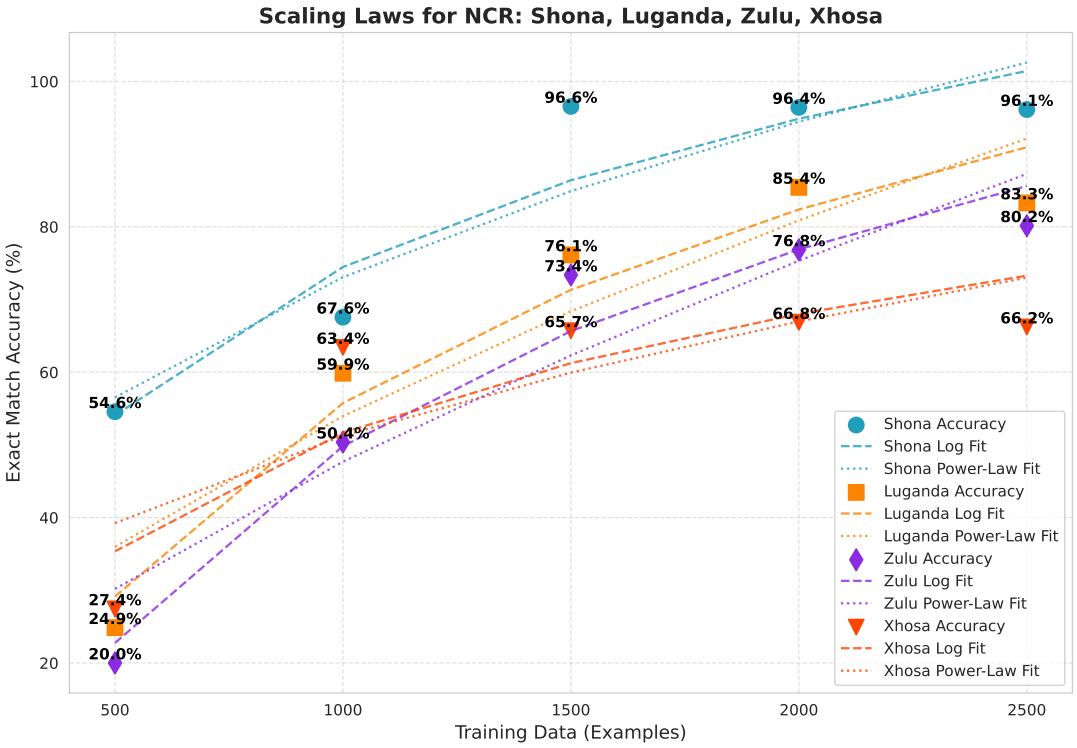

Figure 11: Data scaling analysis for the DictionaryNCR dataset across Bantu languages

While further scaling could improve performance, our results highlight that structural priors, such as the Morphological Index (MoI) in MoI-MoE, offer an alternative path to efficiency. Instead of relying solely on larger models, integrating linguistic knowledge allows NCR to generalize effectively even at moderate scales (3.7B parameters).