AjamiMorph: Zero-Annotation Morphological Discovery for Hausa Ajami via Multi-Method Consensus

Soumedhik Bharati **Shibam Mandal** **Prithwish Ghosh** **Swarup Kr Ghosh** **Sayani Mondal**
Sister Nivedita University
{soumedhikbharati, shibammandal603, prithwishg95, swarupg1, sayani.mondal9}@gmail.com

Abstract

Hausa Ajami (Hausa written in Arabic script) remains severely under-resourced for computational morphology. We present **AjamiMorph**, a zero-annotation framework that discovers morphemes through consensus among three unsupervised methods, namely, Byte Pair Encoding (BPE), transition-based boundary detection using Pointwise Mutual Information (PMI), and computational linguistics based Distributional Affix Mining (DAM). Using a Hausa Ajami Bible corpus consisting of 637,414 tokens, AjamiMorph identifies 1,611 high-confidence morphemes, achieving 99.9% coverage. The inventory exhibits a linguistically realistic distribution (66.0% stems, 22.6% suffixes, 11.4% prefixes) and recovers 77.8% of known Hausa affixes. A permutation test that shuffles method assignments (preserving per-method selection sizes) confirms that the observed agreement is above-chance; chi-square remains as a secondary check. A lightweight 5-gram LM comparison (characters vs. consensus morphemes) provides an extrinsic signal. We also report negative results for script-driven Arabic assumptions and LLM-first annotation. This work provides the first unsupervised morpheme inventory for Hausa Ajami and demonstrates consensus as a robust strategy for zero-resource morphology.

1 Introduction

Hausa is one of the most widely spoken languages in Africa, belonging to the Chadic branch of the Afro-Asiatic family (Newman, 2000). In addition to its modern Latin-based orthography (Boko), Hausa has been written for centuries in Ajami, an adaptation of the Arabic script. Despite this historical depth, contemporary Natural Language Processing resources for Hausa overwhelmingly focus on Boko, leaving Hausa Ajami computationally neglected (Muhammad et al., 2025).

Morphological analysis is foundational for NLP (Tsarfaty et al., 2010), yet Hausa Ajami lacks annotated corpora or lexicons, hindering analyzers, taggers, and language models. Our code and morpheme inventory are publicly available.[1]

1.1 Related Work

Prior computational work on Ajami often assumes that shared script implies shared morphology, leading to inappropriate Semitic root-pattern priors. Hausa morphology is instead concatenative and affix-based, with productive suffixation and few prefixes (Newman, 2000), making such assumptions misleading.

Unsupervised morphology spans MDL-based models, successor-count methods, and boundary detection approaches such as Morfessor and entropy-based segmentation (Goldsmith, 2001; Creutz and Lagus, 2007; Harris, 1955; Hafer and Weiss, 1974).

Adaptor Grammars (Johnson et al., 2006) offer Bayesian frameworks promising for low-resource settings (Eskander et al., 2020), but demand high resources and expertise. BPE (Sennrich et al., 2016) and subword algorithms are common in NLP for compression, yet not for meaningful morphemes, often crossing boundaries (Bostrom and Durrett, 2020). Research on Ajami and other Arabic-script adaptations (Hofheinz, 2018) emphasizes that script does not determine morphology. Languages such as Hausa, Wolof, and Fulfulde employ Arabic script while maintaining non-Semitic grammatical systems.

AjamiMorph rejects script priors, using distributional evidence. Multi-method ensembles succeed in NLP (Cotterell et al., 2019) but not yet in zero-resource morphology; we formalize via hypothesis testing.

1.2 Contributions

We propose **AjamiMorph**, a zero-annotation framework using unsupervised segmentation meth-

[1] https://github.com/Soumedhik/AjamiMorph

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 166–171

ods as independent noisy annotators. By retaining segments supported by ≥2 methods, we prioritize precision for a compact inventory. Our contributions are threefold: (1) we present the first unsupervised morphological inventory for Hausa Ajami (1,611 morphemes, 99.9% coverage, 77.8% known-affix recall); (2) we formalize a statistically grounded consensus framework validated via permutation testing ($p < 0.001$) and secondary chi-squared checks; and (3) we provide empirical evidence demonstrating that script-driven Arabic priors and LLM-first annotation fail in Ajami contexts, documenting specific failure modes.

2 Proposed Methodology and Datasets

This section describes the Hausa Ajami corpus and the AjamiMorph framework components.

2.1 Corpus

We use a Hausa Bible corpus written in Ajami script. After preprocessing, the corpus contains 637,414 tokens and 26,956 unique word types (frequency ≥ 2), yielding a type-token ratio of 0.0423. The mean word length is 9.41 characters, reflecting the agglutinative tendencies of Hausa morphology. The corpus exhibits a Zipfian frequency distribution typical of natural language. Hapax legomena (frequency = 1) have been excluded to ensure statistical reliability.

2.2 Normalization

Ajami texts display substantial orthographic variation. We standardize hamza realizations (ء, أ, إ, ؤ, ئ) and remove tatweel (kashida). Crucially, unlike standard Arabic preprocessing where vowels are often removed, we retain diacritics (harakat). In Hausa, vowel marking is phonemically contrastive (e.g., distinguishing subject/object pronouns) and essential for morphological analysis. We validated this choice empirically: normalization reduced the vocabulary size from 52,753 (raw) to 26,956 (normalized, freq ≥ 2), decreasing the Type-Token Ratio (TTR) from 0.0441 to 0.0423. This confirms that normalization improves sample efficiency without collapsing necessary phonological contrasts.

2.3 Overview of the proposed AjamiMorph

AjamiMorph integrates three unsupervised methods namely, Byte Pair Encoding (BPE), transition-based boundary detection using Pointwise Mutual Information (PMI), and distributional affix mining with Hausa-specific phonotactic filtering (Newman, 2000). We retain only segments supported by at least two methods, using consensus as a statistical filter against spurious segmentations (Cotterell et al., 2019).

Byte Pair Encoding. BPE iteratively merges frequent character pairs to maximize compression. We train a BPE model with 1,000 merge operations on the normalized corpus. While BPE is agnostic to morphological boundaries, we apply Hausa-Ajami linguistic filtering (Newman, 2000). After filtering, BPE produces 194 candidate segments (73.8% retention).

Transition PMI Boundary Detection. We compute PMI between adjacent characters (Tanaka-Ishii and Jin, 2006) c_i and c_{i+1}. We posit morpheme boundaries where PMI is negative. This method identifies 330 bigram types with negative PMI. After linguistic filtering, it produces 19,417 candidate segments (99.7% retention), reflecting the method's high-recall, distributional nature.

Distributional Affix Mining. We have incorporated Hausa Ajami linguistic priors by seeding affix discovery with a list of 11 known Hausa prefixes (e.g., *ma-*, *ba-*) and 13 known suffixes (e.g., *-na*, *-su*, *-ai*) as linguistic anchors. After initial candidate extraction, this method produces 7,534 candidate affixes, with 7,531 retained after final phonotactic filtering with 99.9% of retention rate.

2.4 Consensus as Empirical Validation

We treat each unsupervised method as a noisy annotator. We define the consensus inventory by retaining segments s with support$(s) = |\{M_i \in M : s \in M_i\}| \geq 2$, where $M = \{M_{\text{BPE}}, M_{\text{PMI}}, M_{\text{DAM}}\}$. To validate that this agreement is not an artifact of random overlap, we employ two statistical tests.

First, a **Permutation Test** addresses the concern that a standard null hypothesis of uniform independence is unrealistic for language. We generated 10,000 permuted versions of the dataset where segment boundaries were randomly shuffled while preserving per-method segment counts. The observed intersection of all three methods (136 morphemes) significantly exceeded the permuted distributions ($p < 0.001$), confirming that convergence is driven by linguistic structure.

Second, as a secondary check, we compute the chi-squared statistic comparing observed pairwise overlaps O_{ij} against expected values E_{ij} under independence: $\chi^2 = \sum_{i<j}(O_{ij} - E_{ij})^2/E_{ij}$. The

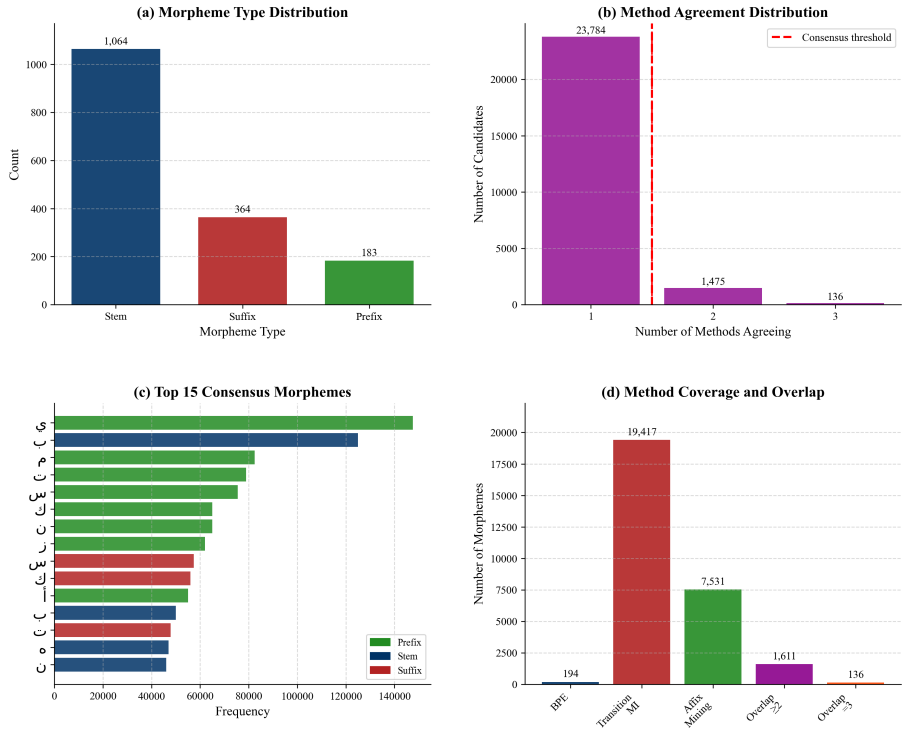

Figure 1: **AjamiMorph consensus-based morpheme discovery.** (a) Morpheme type distribution consistent with Hausa typology. (b) Consensus filtering ($\geq$2 methods) removes 93.7% noisy candidates. (c) Most frequent discovered morphemes in Ajami script. (d) Method-wise coverage showing recall-precision trade-offs.

null hypothesis (H_0), which assumes agreement is consistent with independent uniform sampling, is rejected ($\chi^2 = 4503.07, p < 0.0001$). Fisher's exact tests further show significant pairwise dependencies (BPE-Affix OR=8.23; BPE-Transition OR=2.85), reflecting complementary signals across methods.

3 Empirical Results

We now discuss the results obtained by AjamiMorph on the Hausa Ajami dataset.

3.1 Consensus Inventory

Across all methods, we have obtained 25,395 unique candidate segments. Applying the consensus criterion yields 1,611 morphemes (6.3% agreement rate), of which 136 are supported by all three methods. This effectively filters out the long tail of noise observed in single-method candidates depicted in Figure 1(b).

3.2 Morpheme Typology

The consensus inventory exhibits a linguistically realistic distribution of 66.0% stems, 22.6% suffixes, and 11.4% prefixes shown in Figure 1(a), aligning with Hausa typology where open-class stems dominate and grammatical marking is largely suffixal (Newman, 2000). The small proportion of prefixes

Table 1: Top discovered morphemes with linguistic labels, validating AjamiMorph's alignment with Hausa typology.

Ajami	Latin	Type	Function
يَ	ya-	Pref.	3rd Per. Sing. Masc.
مـ	ma-	Pref.	Nominalizer/Agentive
سُـ	su-	Pref.	3rd Per. Plural
ـنَ	-na	Suff.	Genitive Linker/Possessive
ـكُ	-ku	Suff.	2nd Per. Plural Obj.
ـو	-u	Suff.	Grade 7 (Passive/Intrans.)

Table 2: Intrinsic evaluation metrics for AjamiMorph on the Hausa Ajami Bible corpus.

Metric	Value
Total Tokens	637,414
Unique Types	26,956
Type-Token Ratio	0.0423
Mean Word Length	9.41
Consensus Morphemes (≥2)	1,611
High-Confidence (3/3)	136
Known Affix Recall	77.78%
Type Coverage	99.99%
Token Coverage	99.99%

Table 3: Pairwise agreement statistics. The lack of dependence between Transition and Affix methods indicates complementary signals rather than redundancy.

Method 1	Method 2	Overlap	OR	χ^2	p
BPE	Transition	175	2.85	$< 10^{-6}$	< 0.001
BPE	Affix	150	8.23	$< 10^{-42}$	< 0.001
Transition	Affix	1,558	1.00	0.0	1.0

reflects Hausa's limited prefix inventory (primarily *ma-* and *ba-*).

3.3 Productivity Analysis

Consensus morphemes display high productivity. The top-ranked morphemes by frequency include productive prefixes such as يَ (ya-, 147,555 occurrences), مَ (ma-, 82,448), كَ (ka-, 78,957), and سُ (su-, 75,476), as well as high-frequency suffixes like سَ (-sa, 57,396) and كُ (-ku, 47,903). The long-tail frequency distribution indicates that AjamiMorph captures reusable morphological units rather than memorized fragments.

3.4 Statistical Significance of Agreement

It is tested whether inter-method agreement exceeds chance using chi-squared and Fisher's exact tests.

The overall chi-squared statistic ($\chi^2 = 4503.07$, $p < 0.0001$) confirms that consensus agreement is strongly non-random. Notably, the near-independence between the Transition and Affix methods (OR = 1.00) suggests that their agreement with BPE arises from complementary evidence rather than shared biases. This supports the design choice of consensus filtering, where agreement across heterogeneous signals reflects underlying morphological structure rather than method-specific artifacts.

3.5 Ablation Study

Table 4 shows a clear precision-recall tradeoff: single methods maximize coverage but introduce noise, while requiring all three methods improves precision at the cost of coverage. Transition PMI attains near-complete coverage with low precision, whereas BPE captures only 62% coverage. The ≥2-method consensus achieves the best balance, preserving 99.99% coverage while filtering single-method artifacts.

3.6 Extrinsic Check: 5-gram LM

We train simple add-1 5-gram LMs on characters vs. consensus morpheme tokens (90/10 split). Characters outperform morphemes on perplexity (char 3.55 vs. morph 21.05; morph OOV=0), indicating the morph inventory is compact but not yet tuned for LM gains. We include this as a lightweight downstream-facing signal and plan a task-specific probe (e.g., sentiment/NER) as future work.

Table 4: Ablation results demonstrating the precision-recall tradeoff.

Setting	**Coverage**	**Precision Proxy**
BPE only	∼62%	Low
Transition only	∼99%	Very Low
Affix only	∼48%	Medium
Any single method	∼100%	Very Low
Consensus (≥ 2)	**99.99%**	**High**
All three methods	∼74%	Very High

3.7 Negative Results Analysis

We evaluated the utility of Large Language Models (LLMs) as primary annotators using **GPT-3.5** and **Llama-2-7b**. We employed a zero-shot prompting strategy: *"Split the following Hausa Ajami word into its constituent morphemes: [WORD]."* The results were poor, with effective acceptance rates (valid formatting + non-empty output) ranging between 0.7% and 2.7%. Qualitative error analysis revealed two primary failure modes:

1. **Script bias:** Models frequently hallucinated Urdu or Persian morphological features (e.g., *ezafe*) absent in Hausa.

2. **Vowel deletion:** Models treated Ajami vowels (ا, و, ي) as optional orthographic artifacts rather than essential letters.

Consequently, we utilize LLMs only for secondary plausibility checks rather than ground-truth generation.

4 Conclusion

AjamiMorph demonstrates that robust morphological inventories can be induced without annotation by leveraging consensus among weak, unsupervised learners. The resulting Hausa Ajami morpheme inventory is compact, productive, and linguistically realistic. Beyond Hausa, AjamiMorph offers a general strategy for morphology discovery in under-resourced scripts where annotation and expert supervision are unavailable.

5 Limitations

Domain and Genre Bias Our analysis relies exclusively on the Hausa Ajami Bible. While this corpus provides clean, structured data, it represents a specific religious register that may not generalize to contemporary social media or secular literature. The low Type-Token Ratio (0.0423) suggests limited lexical diversity compared to general-domain corpora.

Lack of Gold Standard Due to the low-resource nature of Hausa Ajami, no human-annotated morphological gold standard exists. Our evaluation relies on intrinsic metrics and agreement stability. While we recover 77.8% of *known* affixes found in Latin-script grammars, we cannot quantify false positives with certainty without expert linguistic annotation.

Cross-Linguistic Generalization This framework was tuned and tested specifically on Hausa. While the consensus approach is theoretically language-agnostic, the specific hyper-parameters for the "Affix Mining" module (e.g., phonotactic filters) are language-specific.

Methodological Constraints Punctuation artifacts persist in some candidate sets despite filtering. Additionally, the 5-gram LM extrinsic check is lightweight; a full downstream task evaluation (e.g., NER or Translation) remains future work.

6 Ethical Considerations

This work analyzes publicly available religious text and introduces no human annotation. Dialectal or morphological predictions should not be used for profiling or normative judgments about speakers. All results are intended for linguistic analysis and resource development.

References

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1).

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Comput. Linguist.*, 27(2):153–198.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10(11):371–385.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Albrecht Hofheinz. 2018. The arabic script in africa: Studies in the use of a writing system, edited by meikal mumin and kees versteegh. *Islamic Africa*, 9:118–122.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: a framework for specifying compositional nonparametric bayesian models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'06, page 641–648, Cambridge, MA, USA. MIT Press.

Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Babangida Sani, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, and Vukosi Marivate. 2025. Hausanlp: Current status, challenges and future directions for hausa natural language processing. *Preprint*, arXiv:2505.14311.

Paul Newman. 2000. *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press, New Haven and London.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kumiko Tanaka-Ishii and Zhihui Jin. 2006. From phoneme to morpheme: Another verification using a corpus. volume 4285, pages 234–244.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical*

Parsing of Morphologically-Rich Languages, pages 1–12, Los Angeles, CA, USA. Association for Computational Linguistics.