MEIV: Representation-Level Normalization for Robust Tokenization in Non-Concatenative Languages

Anonymous ACL submission

Abstract

Subword tokenization methods such as BPE and WordPiece perform well in concatenative languages but struggle in non-concatenative languages such as Tigrigna, Arabic, and Hebrew, where lexical meaning is distributed across consonantal roots and inflectional vowel patterns. Consequently, morphologically related words often exhibit limited surface overlap, leading to fragmented representations, vocabulary sparsity, and degraded downstream performance. Existing approaches improve segmentation but remain grounded in unstable surface forms, leaving intra-morphemic variation unresolved. These observations suggest that representation instability may play a more central role than segmentation alone in non-concatenative NLP.

We propose Morphological Explicitization of Inflectional Vowels (MEIV), a representation-level normalization framework that stabilizes lexical representations by explicitly separating consonantal structure (lexical identity) from inflectional vowel patterns (grammatical variation). We instantiate MEIV for Tigrigna and evaluate it through intrinsic analysis together with sparse lexical and dense semantic retrieval benchmarks. MEIV substantially reduces vocabulary growth and out-of-vocabulary rates while consistently improving retrieval effectiveness across both sparse and dense retrieval settings. The results suggest that representation-level normalization may provide a more effective strategy than segmentation alone for processing morphologically rich non-concatenative languages.

1 Introduction

Subword tokenization methods such as BPE (Sennrich et al., 2016) and WordPiece (Schuster and Nakajima, 2012) have enabled strong performance in concatenative languages, where words are formed through linear combinations of prefixes, stems, and suffixes (Schmidt et al., 2024; Gazit et al., 2025; Teklehaymanot et al., 2025). These methods assume that meaningful linguistic units are preserved in contiguous surface substrings, enabling efficient vocabulary compression and generalization.

This assumption breaks in non-concatenative languages such as Tigrigna, Arabic, and Hebrew (Schmidt et al., 2024; Gazit et al., 2025), where lexical meaning is encoded through consonantal roots and grammatical variation is expressed through templatic vowel patterns (Habash, 2010; Pam, 1973). As a result, morphologically related forms often exhibit limited surface overlap despite sharing a common semantic core, leading to fragmented representations, inefficient vocabulary growth, inflated out-of-vocabulary (OOV) rates, and degraded downstream performance.

More fundamentally, existing approaches attempt to improve segmentation while remaining grounded in unstable surface forms. Because subword tokenizers rely on contiguous substrings, they fail to resolve intra-morphemic variation such as internal vowel alternations in non-concatenative morphology. These observations suggest that the primary bottleneck in non-concatenative NLP may lie in representation rather than segmentation alone (Schmidt et al., 2024; Gazit et al., 2025).

This limitation is particularly evident, for example, in Tigrigna, which is written in the Ge'ez script, an abugida in which each character represents a consonant-vowel syllable (Keleta et al., 2016; Pam, 1973; Bulakh, 2019). In this writing system, morphological variation often alters consonants and vowels simultaneously, while morpheme boundaries are not explicitly marked and may follow non-linear patterns (Fisseha and Gao, 2016; Gasser, 2011). For example, as illustrated in Figure 1, forms derived from the root ሰተየ ("to drink") such as ሰትዮም, ምስሰተዩ, and ሰቲናዮ share a common lexical base but exhibit limited surface

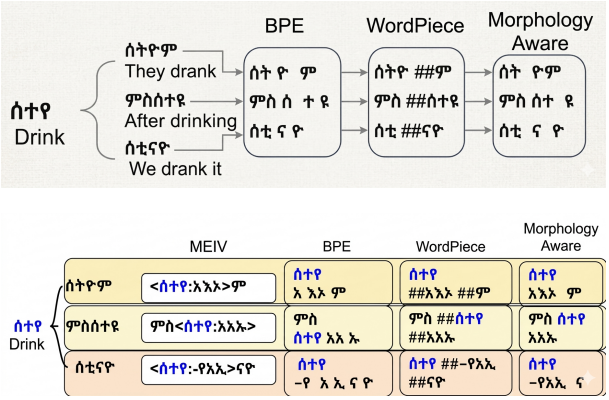


Figure 1: Surface-form divergence across morphologically related Tigrigna forms derived from shared lexical roots, illustrating representation instability under non-concatenative morphology.

overlap. Standard tokenizers segment these forms inconsistently, leading to weak lexical alignment across morphological variants. Additional examples are provided in Appendix Table 10.

This creates two major challenges for subword tokenization: (1) *surface fragmentation*, where related forms exhibit limited token overlap despite semantic relatedness, and (2) *representation instability*, where internal vowel alternations produce inconsistent stem-level representations across derivational and inflectional variants.

To address this limitation, we propose **Morphological Explicitization of Inflectional Vowels (MEIV)**, a representation framework that transforms non-concatenative surface forms into structured, tokenization-compatible representations by separating consonantal structure (lexical identity) from inflectional vowel patterns (grammatical variation), as illustrated in Figure 1. This yields stable lexical representations across morphological variants while preserving reversibility.

Unlike SPLINTER (Gazit et al., 2025), which primarily restructures surface forms for tokenizer compatibility, MEIV directly stabilizes lexical representation through explicit separation of consonantal structure and inflectional vowel variation.

We instantiate MEIV for Tigrigna and evaluate it in information retrieval, a task highly sensitive to lexical mismatch. MEIV reduces vocabulary size by over 50%, substantially lowers out-of-vocabulary (OOV) rates, and improves retrieval effectiveness, increasing Precision@10 by more than 2×. These results suggest that representation-level normalization may provide a more effective strategy than segmentation alone under non-concatenative morphology. If MEIV improves lexical stability rather than only sparse lexical overlap, its benefits should generalize beyond retrieval to neural sequence modeling tasks such as machine translation.

Contributions.

- We show that unstable lexical representation, rather than segmentation alone, is a primary bottleneck for tokenization in non-concatenative languages, and propose MEIV to address this limitation.
- We construct an expert-annotated dataset of 5,300 surface-form–MEIV pairs and curate a Tigrigna retrieval benchmark with 5,985 documents and 1,125 relevance judgments.
- We show that MEIV consistently improves both sparse lexical and dense semantic retrieval on Tigrigna and Amharic benchmarks.

2 Related Work

Subword tokenization methods such as BPE (Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), and unigram language modeling (Kudo, 2018) assume that lexical structure can be approximated through contiguous surface substrings. While effective for concatenative languages, these methods often produce fragmented and inconsistent segmentations in morphologically rich non-concatenative languages (Schmidt et al., 2024; Gazit et al., 2025; Amrhein and Sennrich, 2021). In Semitic languages such as Arabic, Hebrew, and Tigrigna, lexical meaning is distributed across discontinuous consonantal roots and vowel patterns (Pam, 1973; Habash, 2010; Hatch and Richardson, 2025), causing morphologically related forms to exhibit low surface overlap despite shared semantics (Boudelaa and Marslen-Wilson, 2015; Hatch and Richardson, 2025; Limisiewicz et al., 2023).

Prior work has addressed these challenges through morphology-aware segmentation, root extraction, templatic morphology modeling, and linguistic analyzers (Habash, 2010; Schmidt et al., 2024). For Ethiopic languages, systems such as HornMorpho (Gasser, 2011) and MoVoC (Teklehaymanot et al., 2025) incorporate morphological knowledge to improve segmentation and vocabulary consistency, while Arabic and Hebrew NLP research has explored morphology-aware tokenization and root-pattern modeling (Hatch and

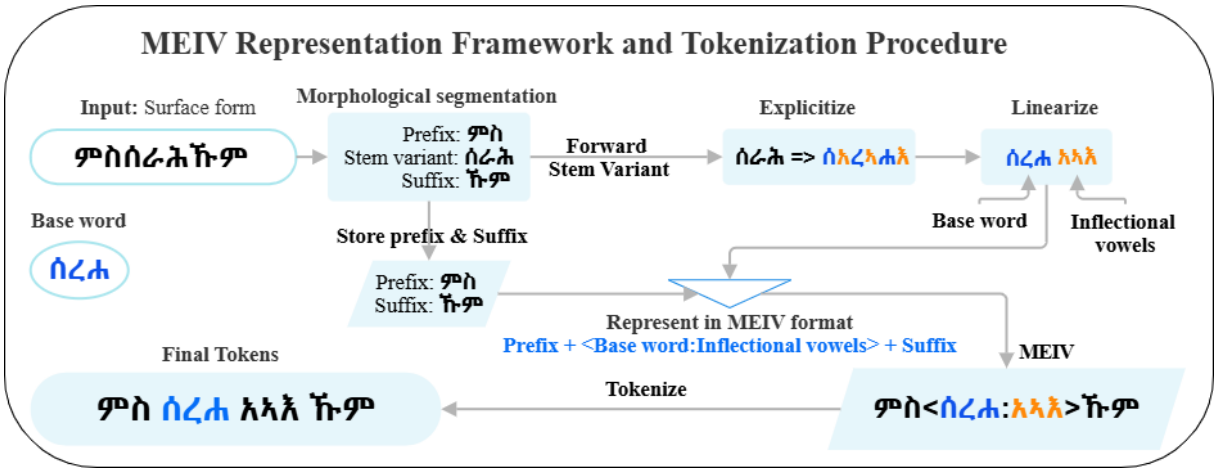


Figure 2: Overview of the MEIV transformation process. Surface forms are segmented into affixes and base variants, transformed through vowel explicitization and explicit vowel linearization, and reassembled into stable, tokenization-compatible representations that preserve lexical identity across morphological variants.

Richardson, 2025; Shmidman et al., 2018). However, these approaches remain grounded in surface-form segmentation and do not explicitly address representation-level instability caused by internal vowel alternation across morphological variants.

Recent work has emphasized representation quality in multilingual and morphologically rich NLP. Character-level and byte-level models such as CANINE (Clark et al., 2022) and ByT5 (Xue et al., 2022) reduce reliance on fixed vocabularies and subword segmentation, but often produce weaker morphological abstraction. In contrast, MEIV separates consonantal structure from vowel patterns to generate stable, reversible, and tokenization-compatible representations across morphological variants.

3 Methodology

We propose **Morphological Explicitization of Inflectional Vowels (MEIV)**, a representation framework that addresses the challenges identified in Section 2 by restructuring surface forms to align with non-concatenative morphology.

MEIV restructures surface forms through segmentation, vowel explicitization, and linearization, as illustrated in Figure 2.

3.1 Addressing Surface Fragmentation

MEIV reduces surface fragmentation by transforming Ge'ez syllabic characters into explicit consonant–vowel representations using the አ consonant family as a canonical vowel inventory corresponding to the seven Ge'ez vocalic orders (Table 7). Representative decompositions include ሰ→ሰአ, ት→ተእ, and ዮ→የኦ (Table 8).

At the word level, MEIV transforms only the stem variant while preserving concatenative prefixes and suffixes. Since fragmentation primarily arises from root–pattern variation and internal vowel alternations (Fisseha and Gao, 2016; Gasser, 2011), separating consonantal structure from inflectional variation yields stable tokenization-compatible representations (Figure 2).

3.2 Formal Representation Overview

Let Σ denote the Ge'ez syllabic alphabet. We define a decomposition $\delta : \Sigma \to C \times V$, where C and $V = \{v_1, \ldots, v_7\}$ denote consonants and vowel orders, mapping each character $\sigma \in \Sigma$ to a (consonant, vowel) pair $\delta(\sigma) = (c, v)$ (Table 8). Each word is segmented as $\text{seg}(w) = (\pi, \beta, \sigma)$ where π, β, σ denote prefix, stem, and suffix; MEIV transforms only the non-concatenative stem β (Appendix C):

$$\mathcal{M}(w) = \pi\langle BCS(\beta) : \text{IIVS}(\beta)\rangle\sigma \qquad (1)$$

where BCS (consonant spine) encodes lexical identity and $IIVS$ (vowel sequence) encodes inflectional variation. Morphologically related words sharing a lexical root yield identical BCS but distinct $IIVS$, enabling retrieval via BM25 and cosine similarity over $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{Q})$.

3.3 MEIV Converter Model

To apply MEIV at scale, we introduce a surface-to-MEIV converter based on **ByT5** (Xue et al., 2022), a byte-level sequence-to-sequence model. Byte-

level modeling avoids subword assumptions and naturally captures non-concatenative patterns.

The converter maps surface forms to structured MEIV representations (Figure 2), enabling scalable normalization and consistent tokenization with improved retrieval performance.

4 Experimental Setup

We evaluate MEIV through the following research questions:

- **RQ1:** Does it improve representation efficiency?
- **RQ2:** Does it improve downstream task performance?
- **RQ3:** Are gains due to representation or tokenization choice?

4.1 Dataset and Annotation

We construct a corpus of **5,985 documents** from Tigrigna media, institutional sources, and social media, complemented by TLMD (Gaim et al., 2021) and NLLB-200 (Team et al., 2022). From this corpus, we extract **5,300 unique surface forms**, which are manually annotated with corresponding MEIV representations.

The annotations are created manually with support from language experts at the Department of Tigrigna, Mekelle University. The resulting dataset provides an explicit and reversible representation of morphological structure not captured by existing resources.

To enable scalability beyond manual annotation, we train a surface-to-MEIV conversion model (Section 3.3)

4.2 MEIV Converter

We initialize the model from the pretrained `google/byt5-small` checkpoint (Xue et al., 2022) using the HuggingFace Transformers framework (Wolf et al., 2020). The annotated dataset contains 5,300 manually aligned surface–MEIV transformation pairs divided into training, validation, and test splits.

Training is performed for 20 epochs using AdamW with a learning rate of 3×10^{-4}, batch size 16, and 100 warmup steps. Validation and checkpointing are performed every 200 steps, and decoding uses autoregressive generation with a maximum output length of 64 tokens. Full training hyperparameters are provided in Appendix D.

4.3 Downstream Task Setup

Retrieval Setup. We construct a manually curated Tigrigna retrieval benchmark consisting of **5,985 documents**, **75 queries**, and **1,125 relevance judgments**. To evaluate cross-language generalization, we additionally construct an Amharic benchmark containing **15 queries**, **250 documents**, and **150 relevant documents** (**10 relevant documents per query**). Both benchmarks emphasize severe surface variation across morphologically related forms, a well-known challenge in Semitic and morphologically rich languages (Habash, 2010; Cotterell et al., 2018).

We evaluate retrieval under both lexical and semantic paradigms. For lexical retrieval, we use **BM25** (Robertson and Zaragoza, 2009), a sparse lexical model sensitive to tokenization quality. For semantic retrieval, we use the multilingual dense embedding model **multilingual-e5-base** (mE5) (Wang et al., 2024), to evaluate semantic alignment beyond exact lexical matching.

Unless otherwise stated, retrieval experiments use automatically generated MEIV representations produced by the trained ByT5 converter model rather than gold manual annotations.

Together, these settings enable controlled evaluation of whether MEIV improves retrieval robustness under both sparse lexical and multilingual semantic retrieval.

Machine Translation Setup. We fine-tune the pretrained `Helsinki-NLP/opus-mt-ti-en` Transformer model (Tiedemann and Thottingal, 2020) for Tigrigna–English machine translation. Training uses a maximum sequence length of 128, batch size 16, 5 epochs, and a 90/10 train–test split.

Tokenization Methods. We evaluate three tokenization methods: **WordPiece**, **BPE**, and a **morphology-aware** rule-based tokenizer. Each tokenizer is applied under two representation settings: original surface forms and MEIV-transformed text.

- **WordPiece**: A standard subword tokenizer.
- **BPE**: Byte-Pair Encoding.
- **Morphology-aware**: A rule-based tokenizer.

All tokenizers are trained on the same corpus with a shared vocabulary size of 35,000 and evaluated under identical conditions. This design iso-

lates the effect of representation from the tokenization algorithm.

Evaluation. Intrinsic metrics evaluate representation efficiency through vocabulary size and out-of-vocabulary (OOV) rates, while retrieval metrics assess lexical alignment across morphological variants.

We report **Precision@K** and **Recall@K** at multiple cutoffs ($K = 10, 100, 300$). For the surface-to-MEIV converter, we report exact match and character-level accuracy.

5 Results

We evaluate MEIV from both intrinsic and extrinsic perspectives to assess its impact on tokenization performance and downstream retrieval effectiveness. Our analysis integrates vocabulary efficiency, out-of-vocabulary (OOV) reduction, and Information Retrieval (IR) performance under the BM25 ranking framework.

5.1 Intrinsic Evaluation

5.1.1 Representation Efficiency

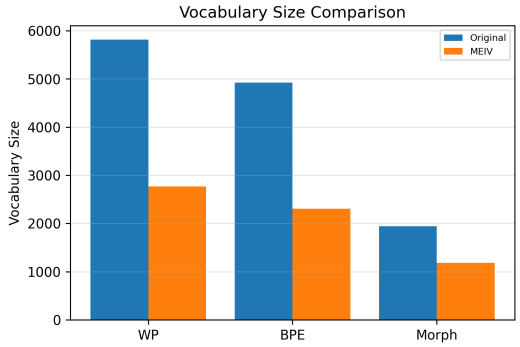
MEIV consistently improves tokenization efficiency across all methods. Vocabulary size is reduced by over 50% for WordPiece and BPE, and by 39% for the morphology-aware tokenizer, reflecting consolidation of morphologically related forms.

MEIV significantly reduces OOV rates (e.g., WordPiece: 12.89% → 1.96%), improving generalization across morphological variants.

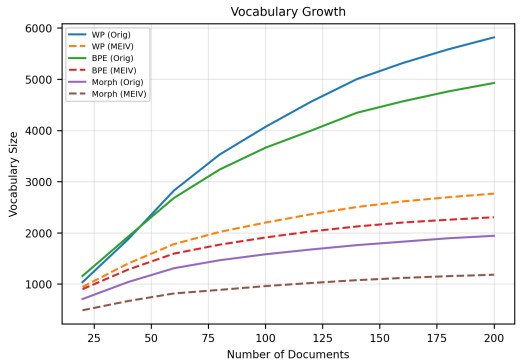
These results show that vocabulary inefficiency in non-concatenative languages is driven by morphological variation rather than true lexical diversity. By normalizing this variation at the representation-level, MEIV enables stable lexical representations and substantially reduces fragmentation.

5.1.2 MEIV Conversion Quality

Validation loss decreases from 0.179 to 0.038, indicating stable optimization with minimal overfitting. To evaluate automatic MEIV generation, we measure exact match and character-level accuracy on held-out surface–MEIV pairs. The converter achieves 87.51% exact match accuracy and 94.35% character accuracy, indicating robust transformation of surface forms into structured MEIV representations.



(a) MEIV substantially reduces vocabulary size across tokenization methods.



(b) MEIV slows vocabulary growth across tokenization methods as corpus size increases.

Figure 3: Impact of MEIV on representation efficiency. MEIV reduces vocabulary size and slows vocabulary growth across tokenization methods by normalizing morphological variation.

Representation	P@10	R@100	MRR@10
Gold MEIV	0.693	0.892	0.840
Predicted MEIV	0.599	0.852	0.730

Table 1: Retrieval comparison between gold and automatically predicted MEIV representations.

These results demonstrate that MEIV transformations can be generated reliably at scale, enabling downstream modeling without manual morphological annotation.

5.2 Gold vs. Predicted MEIV Retrieval

To evaluate robustness under automatic conversion, we compare retrieval using manually annotated (gold) and automatically predicted MEIV representations. Although predicted MEIV slightly reduces retrieval effectiveness, substantial gains remain under realistic automatic conversion settings.

Retrieval	Format	P@10	R@100	MRR@10
BM25 Retrieval				
WordPiece	Original	0.0587	0.1680	0.2224
WordPiece	MEIV	**0.5987**	**0.8524**	0.7303
BPE	Original	0.0587	0.1760	0.2253
BPE	MEIV	0.5947	0.8520	0.7314
Morph-A.	Original	0.4480	0.4462	0.6277
Morph-A.	MEIV	0.5000	0.6836	**0.8824**
Character-Level TF-IDF Retrieval				
Char-Level	Original	0.0573	0.1778	0.2377
Char-Level	MEIV	0.2040	0.6516	0.5316
Dense Retrieval (E5)				
E5	Original	0.0210	0.1227	0.0589
E5	MEIV	0.1270	0.3351	0.2871

Table 2: Retrieval performance comparison across BM25, character-level TF-IDF, and dense retrieval using E5 embeddings. Results compare original surface forms against MEIV-transformed representations. Bold values indicate the best performance across all systems for each metric.

5.3 Retrieval Performance

5.3.1 Tigrigna Retrieval Results

Table 2 presents retrieval performance on the Tigrigna benchmark across sparse lexical retrieval (BM25), character-level TF-IDF retrieval, and dense retrieval using multilingual E5 embeddings.

Across all retrieval settings, MEIV consistently improves retrieval effectiveness over original surface-form representations. Improvements over original surface-form baselines are statistically significant under paired bootstrap resampling ($p < 0.01$). Under BM25 retrieval, the largest gains are observed for WordPiece and BPE tokenization, where P@10 increases from approximately 0.059 to 0.60 and R@100 improves from below 0.18 to above 0.85. Morphology-aware tokenization also benefits from MEIV, improving P@10 from 0.4480 to 0.5000 and MRR@10 from 0.6277 to 0.8824, suggesting that representation-level normalization provides benefits beyond morphology-aware segmentation alone.

Character-level TF-IDF retrieval similarly improves under MEIV, increasing P@10 from 0.0573 to 0.2040 and R@100 from 0.1778 to 0.6516. Dense retrieval using multilingual E5 embeddings also benefits substantially, with P@10 increasing from 0.0210 to 0.1270 and R@100 from 0.1227 to 0.3351. Overall, the results indicate that MEIV improves both lexical and semantic retrieval by reducing representation instability across morphological variants.

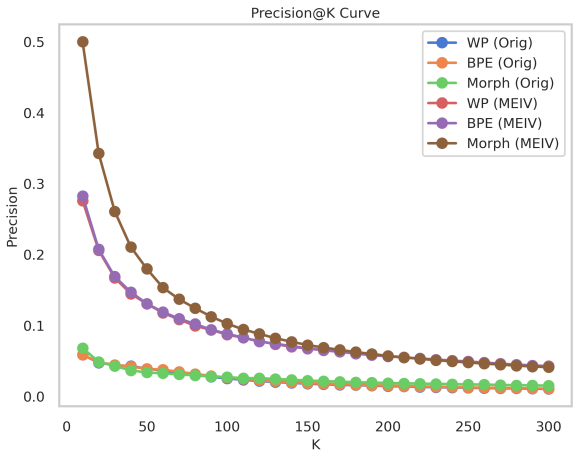


Figure 4: Precision@K comparison between original surface-form and MEIV representations across tokenization methods.

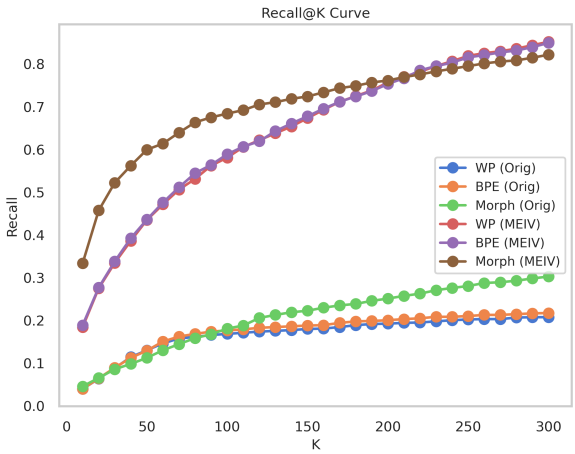


Figure 5: Recall@K comparison between original surface-form and MEIV representations across tokenization methods.

Figures 4 and 5 further show that MEIV consistently improves both early precision and retrieval coverage across different ranking depths.

5.3.2 Amharic Retrieval Results

To evaluate cross-language generalization, we additionally conduct retrieval experiments on the Amharic benchmark.

Table 3 shows that the benefits of MEIV generalize consistently across sparse lexical retrieval, character-level retrieval, and dense semantic retrieval.

MEIV consistently improves retrieval effectiveness across sparse lexical, character-level, and dense retrieval settings. Under BM25 retrieval, MEIV substantially improves ranking quality, in-

Method	Rep.	P@10	R@100	MRR@10
BM25	Original	0.1600	0.2844	0.3417
BM25	MEIV	**0.6667**	**0.6667**	**0.7711**
Char TF-IDF	Original	0.260	0.667	**0.511**
Char TF-IDF	MEIV	**0.373**	**0.667**	0.489
E5	Original	0.300	0.521	0.371
E5	MEIV	**0.400**	**0.762**	**0.467**

Table 3: Amharic retrieval performance comparing original and MEIV-transformed representations.

Rank	Score	Morphological Variants Matched
1	54.66	ሰልጢኑ, ኣቀራርባ, ሓሚማ, ሓኪምና
2	52.65	ስልጠንቲ, ቀሪቦም, ሓሚምካ, ይሓክሙ
3	52.51	ሰልጢንና, ቀሪበ, ሕሙማት, ክንሕክመኩም
4	52.10	ሰልጢነ, ቀረበ, ሓሚሙ, ሓኪመን
5	51.98	ሰልጢኖም, ቀሪብና, ሓሚሞም, ሓኪሞም

Table 4: Representative MEIV retrieval results for the query ስልጠና ሕሙማት ቀረብ ሕክምና, showing robust alignment across morphologically divergent variants.

creasing P@10 from 0.160 to 0.667 and MRR@10 from 0.342 to 0.771. Character-level TF-IDF retrieval also benefits from MEIV, improving P@10 from 0.260 to 0.373 while maintaining comparable recall performance. Similar gains are observed under dense retrieval using multilingual E5 embeddings, where P@10 increases from 0.300 to 0.400 and R@100 from 0.521 to 0.762. Overall, these results suggest that the retrieval benefits of MEIV generalize beyond Tigrigna to related Ge'ez-script languages and across both sparse and dense retrieval methods.

5.3.3 Qualitative Retrieval Analysis

To analyze retrieval behavior under extensive morphological variation, we examine the query:

ስልጠና ሕሙማት ቀረብ ሕክምና

Under the original representation, only one of the 15 relevant documents is retrieved within the top-ranked results due to substantial surface-form divergence. In contrast, MEIV successfully retrieves all 15 relevant documents, with most appearing among the highest-ranked results.

Table 4 presents representative retrieved documents together with the matched morphological variants. Despite substantial orthographic and inflectional variation, MEIV consistently aligns semantically related forms sharing the same underlying lexical structure.

These findings suggest that MEIV improves retrieval by stabilizing lexical representations across derivational and inflectional variants, thereby substantially reducing lexical mismatch under non-concatenative morphology.

Setting	P@10	R@100	MRR@10
Symmetric Representations			
MEIV → MEIV	**0.1267**	**0.3351**	**0.2871**
STEM → STEM	0.0773	0.2969	0.1465
DECOM → DECOM	0.0080	0.0462	0.0359
Cross-Representation Retrieval			
MEIV → STEM	0.0480	0.1840	0.2067
DECOM → MEIV	0.0120	0.1067	0.0287
MEIV → DECOM	0.0040	0.0373	0.0169
STEM → DECOM	0.0027	0.0658	0.0167
DECOM → STEM	0.0013	0.0338	0.0017

Table 5: Ablation analysis of MEIV representations under dense retrieval using multilingual-e5-base.

5.3.4 Ablation Analysis

To analyze the contribution of individual representation components, we conduct an ablation study under dense retrieval using multilingual-e5-base. We compare full MEIV, STEM (without explicit vowel linearization), and DECOM (character decomposition prior to linearization).

Table 5 shows that full MEIV consistently achieves the strongest retrieval performance, obtaining the best scores across all evaluation metrics. Removing explicit vowel linearization substantially reduces retrieval effectiveness, while DECOM representations perform poorly across both symmetric and cross-representation settings.

Cross-representation retrieval further reveals that representation mismatch significantly degrades retrieval quality. For example, MEIV→STEM reduces P@10 from 0.1267 to 0.0480, while DECOM-based retrieval settings exhibit near-zero performance. Overall, the results suggest that retrieval robustness depends strongly on stable and compatible lexical representations.

5.4 Machine Translation Generalization

To evaluate whether MEIV generalizes beyond retrieval, we conduct machine translation experiments using two Transformer-based translation models: (1) a baseline trained on original surface forms and (2) a model trained on MEIV-transformed inputs. Both models translate Tigrigna into English under identical training settings.

Table 6 presents the results. MEIV substantially improves translation quality across all major eval-

Metric	Surface→EN	MEIV→EN
BLEU	17.93	**32.57**
chrF	34.58	**40.91**
TER	81.06	**71.31**
ROUGE-L	40.93	**41.55**

Table 6: Machine translation performance for Tigrigna-to-English translation under original and MEIV representations.

uation metrics, increasing BLEU from 17.93 to 32.57 and chrF from 34.58 to 40.91, while reducing TER from 81.06 to 71.31. These improvements suggest that MEIV produces more stable and semantically consistent representations that better support neural sequence modeling.

5.5 Analysis of Representation Effects

The results suggest that retrieval performance in non-concatenative languages depends strongly on representation quality. Under original surface forms, all tokenizers suffer from substantial lexical fragmentation caused by derivational and inflectional variation, while morphology-aware tokenization alone remains insufficient for resolving unstable stem-level representations.

These findings align with prior work showing that substring-based tokenization struggles in morphologically rich languages due to limited surface overlap across related forms (Schmidt et al., 2024; Gazit et al., 2025; Amrhein and Sennrich, 2021; Cotterell et al., 2018). Although morphology-aware approaches improve segmentation quality (Teklehaymanot et al., 2025), they remain grounded in unstable surface forms and therefore cannot fully address intra-morphemic variation introduced by internal vowel alternations.

Across all tokenizers, MEIV consistently improves retrieval quality by stabilizing lexical representations across morphological variants. The largest gains are observed for WordPiece and BPE tokenization, indicating that frequency-based subword methods are particularly sensitive to morphological fragmentation. Dense retrieval experiments further show improved semantic alignment beyond exact lexical overlap, supporting recent observations that tokenization quality alone is insufficient for robust modeling in morphologically complex languages (Schmidt et al., 2024; Limisiewicz et al., 2023).

Cross-language evaluation on Amharic shows similar improvements under both BM25 and dense retrieval, suggesting that MEIV generalizes across Ge'ez-script Semitic languages. Overall, the results indicate that stable lexical representations are more important than segmentation granularity alone for robust retrieval in non-concatenative languages.

6 Limitations

Our evaluation is limited to Tigrigna and Amharic, and we do not validate MEIV on additional non-concatenative languages. The retrieval experiments use a moderately sized benchmark (5,985 documents and 75 queries). Although the observed gains are consistent across retrieval settings, evaluation on larger-scale multilingual retrieval benchmarks would further validate scalability and robustness. Finally, errors from the automatic MEIV converter may propagate to downstream tasks.

7 Conclusion

MEIV improves both intrinsic and extrinsic performance, reducing vocabulary growth and OOV rates while substantially improving retrieval effectiveness.

More fundamentally, the results suggest that the primary bottleneck in non-concatenative NLP lies in representation rather than segmentation alone. By stabilizing lexical identity across morphological variants, MEIV enables robust tokenization-compatible representations under extensive intra-morphemic variation.

Although evaluated on Tigrigna and Amharic, the framework naturally extends to other non-concatenative languages such as Arabic and Hebrew, highlighting representation-level normalization as a promising direction for morphologically rich NLP.

Ethical Considerations

This work uses publicly available datasets and pretrained models, including NLLB-200, TLMD, OPUS-MT, multilingual-e5, and HornMorpho, in accordance with their respective licenses and intended research use. We additionally construct manually annotated MEIV transformation datasets for Tigrigna and Amharic, intended solely for research purposes. No personally identifiable or sensitive information is included in the released data. ChatGPT was used only for limited language editing and paraphrasing assistance.

References

Christian Amrhein and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of EMNLP 2021*, pages 672–687.

Sami Boudelaa and William D. Marslen-Wilson. 2015. Structure, form, and meaning in the mental lexicon: evidence from arabic. *Language, Cognition and Neuroscience*, 30(8):955–992. PMID: 26682237.

Maria Bulakh. 2019. Tigrinya. In Semitic Languages.

Jonathan H. Clark and 1 others. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27.

Girmay Fisseha and Jian Gao. 2016. Morphological segmentation for English-to-Tigrinya statistical machine translation. *International Journal of Computer Processing of Languages*, 27(2):97–106.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. 2021. Monolingual pre-trained language models for Tigrinya. In *5th Widening NLP (WiNLP2021) Workshop, Co-located with EMNLP 2021*.

Michael Gasser. 2011. Hornmorpho: A system for morphological processing of amharic, oromo, and tigrinya. In *Proceedings of the Conference on Human Language Technology for Development*, Alexandria, Egypt.

Bar Gazit, Shaltiel Shmidman, Avi Shmidman, and Yuval Pinter. 2025. Splintering nonconcatenative languages for better tokenization. In *Findings of ACL 2025*, pages 22405–22417.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Brendan T. Hatch and Stephen D. Richardson. 2025. Semitic root encoding: Tokenization based on the templatic morphology of Semitic languages in NMT. In *Proceedings of the Third Arabic Natural Language Processing Conference*, pages 26–41. Association for Computational Linguistics.

Tedla Yemane Keleta, Kazuhide Yamamoto, and Ashuboda Marasinghe. 2016. Tigrinya part-of-speech tagging with morphological patterns and the new nagaoka tigrinya corpus. *International Journal of Computer Applications*, 146(14):33–41.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. *arXiv preprint arXiv:2305.17179*.

Martin D. Pam. 1973. *Tigrinya Phonology*. Ph.D. thesis, ProQuest Dissertations and Theses.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *IEEE ICASSP*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*, pages 1715–1725.

Avi Shmidman, Moshe Koppel, and Ely Porat. 2018. Identification of parallel passages across a large hebrew/aramaic corpus. *Journal of Data Mining & Digital Humanities*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Surendra Bharadwaj, Zhaoyu Wang, Yuki Choi, Ankur Ghodsi, Sébastien Yella, Hieu Nguyen, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hailay Kidu Teklehaymanot, Dren Fazlija, and Wolfgang Nejdl. 2025. Movoc: Morphology-aware subword construction for ge'ez script languages. In *Findings of EMNLP 2025*, pages 13131–13144.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt — building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Thomas Wolf, Lysandre Debut, Victor Sanh, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

A Ge‘ez Alphabet and Vowel System

The Ge’ez script is an abugida consisting of 32 base consonant families and seven primary vowels. Unlike the Latin script, which places consonants and vowels linearly (side-by-side), Ge’ez employs a system of *vowelized consonants(VC)*. As illustrated in Table 9, each symbol represents a unique combination of a consonant and a vowel in a fixed phonetic sequence. Consequently, each base consonant has seven distinct variations, traditionally referred to as ”orders.”

Additionally, there are five labialized consonants (rounded-lip sounds) which feature only five vowel variations instead of the standard seven. The mapping between Latin vowel equivalents and the seven Ge’ez vocalic orders is detailed in Table 7.

Order	v_1	v_2	v_3	v_4	v_5	v_6	v_7
Latin	e	u	i	a	ie	–	o
Ge’ez	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ

Table 7: The seven vocalic orders of the Ge’ez script with Latin approximations.

Representative character-level decompositions are shown in Table 8.

VC	**Base + Vowel**	**Transformation**
ሰ	ሰ + አ	ሰአ
ት	ተ + እ	ተእ
ዮ	የ + ኦ	የኦ

Table 8: Representative MEIV character decompositions.

Consonant	**e**	**u**	**i**	**a**	**ie**	**–**	**o**
S	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
T	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
Y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ

Table 9: Mapping of Latin consonant-vowel pairs to Ge’ez vowelized consonants (VC).

Root	**Noun**	**3F.Sg**	**3F.Pl**	**1Pl**
ሰረሐ (work)	ስራሕ	ስርሒ	ሰሪሐን	ሰሪሕና
ፀሓፈ (write)	ፅሑፍ	ፀሓፊ	ፅሒፈን	ፅሒፍና
በልዐ (eat)	ብልዒ	ብልዒ	በሊዐን	በሊዕና
መፀ (come)	መፃኢ	ምፂ	መፂአን	መፂና
ተሓከመ (treated)	ሕክምና	ተሓከሚ	ተሓኪመን	ንሕከም

Table 10: Morphologically related Tigrigna forms derived from the same lexical root exhibit substantial surface variation across derivational and inflectional forms.

B Detailed MEIV Representation

B.1 Word-Level Transformation

At the word level, each surface form is segmented into prefix, base variant, and suffix components. The MEIV transformation is applied only to the base variant, while prefixes and suffixes are preserved.

Surface	**Prefix**	**Base**	**Suffix**
ሰትዮም	—	ሰትዮ	ም
ምስሰተዩ	ምስ	ሰተዩ	—
ሰቲናዮም	—	ሰቲና	ዮም

Table 11: Segmentation into prefix, base variant, and suffix

The base variant is then decomposed into a sequence of base consonants and vowels:

- ሰትዮ → ሰአተእየኦ
- ሰተዩ → ሰአተአየኡ
- ሰቲ → ሰአተኢ-የ

Final Representation The final MEIV representation separates:

- **Base Consonant Sequence (BCS)**
- **Implicit Inflectional Vowel Sequence (IIVS**

These are combined using a structured format, where the transformed base is enclosed using < > and integrated with prefixes and suffixes.

Surface Form	**Prefix**	**MEIV Base MEIV Base**	**Final Representation**
ሰትዮም	—	ሰተየ:አእኦ	<ሰተየ:አእኦ>ም
ምስሰተዩ	ምስ	ሰተየ:አአኡ	ምስ<ሰተየ:አአኡ>
ሰቲናዮም	—	ሰተየ:አኢ-የ	<ሰተየ:አኢ-የ>ናዮም

Table 12: Final MEIV representation

B.2 Handling Complex Transformations

To preserve reversibility, MEIV introduces explicit markers for non-linear morphological changes:

- − deletion
- ` replacement
- + insertion
- ∗ replication

For example: ኣከያይዳና => ኣ<ኸደ:ኣ+ያ*እኣ>ና These markers allow the original surface form to be reconstructed from the MEIV representation.

C Formal Definition of MEIV

C.1 Ge'ez Character Decomposition

Let Σ denote the Ge'ez syllabic alphabet. Each character jointly encodes a consonant and a vowel order. We define a decomposition function:

$$\delta : \Sigma \to C \times V$$

where:

- C is the consonant inventory,
- $V = \{v_1, v_2, \ldots, v_7\}$ represents the seven Ge'ez vocalic orders.

For each syllabic character $\sigma \in \Sigma$,

$$\delta(\sigma) = (c, v)$$

where:

- $c \in C$ denotes the consonantal component,
- $v \in V$ denotes the associated vowel order.

Representative examples include:

$$\delta(ሰ) = (S, v_1)$$

$$\delta(ት) = (T, v_6)$$

$$\delta(ዮ) = (Y, v_7)$$

corresponding to the explicit decompositions:

ሰ → ሰኣ

ት → ተእ

ዮ → የኦ

C.2 Sequence-Level Decomposition

We extend δ to sequences using:

$$\delta^*(\sigma_1\sigma_2 \ldots \sigma_n) = c_1c_2 \ldots c_n \oplus v_1v_2 \ldots v_n$$

where:

- $c_1c_2 \ldots c_n$ represents the consonantal sequence,
- $v_1v_2 \ldots v_n$ represents the vowel sequence,
- $\oplus$ denotes structured concatenation using a separator token.

This decomposition separates lexical consonantal structure from inflectional vowel information.

C.3 Word-Level Segmentation

At the word level, each surface form w is partitioned into:

$$seg(w) = (\pi, \beta, \sigma)$$

where:
π denotes the prefix sequence,
β denotes the stem variant,
σ denotes the suffix sequence.

The key observation is that prefixes and suffixes already follow concatenative structure, while the stem β exhibits non-concatenative morphological variation. Consequently, MEIV applies transformation only to β.

C.4 Base Consonant and Vowel Sequences

Applying δ^* to the stem yields two complementary representations:

$$BCS(\beta) = c_1c_2 \ldots c_n$$

$$S(\beta) = v_1v_2 \ldots v_n$$

where:

- **BCS** (Base Consonant Sequence) encodes lexical identity,
- **IIVS** (Implicit Inflectional Vowel Sequence) encodes grammatical variation.

C.5 MEIV Transformation

The complete MEIV transformation is defined as:

$$M(w) = \pi \ \langle BCS(\beta) : \text{IIVS}(\beta) \rangle \ \sigma$$

where:

- $\langle \cdot \rangle$ denotes the structured MEIV representation,
- ":" separates lexical and inflectional components.

For example:

$$\text{ሰትዮም} \rightarrow \langle \text{ሰተየ} : \text{አእኦ} \rangle \text{ም}$$

C.6 Lexical Stability Property

The central property of MEIV is lexical stability across morphological variants.

For morphologically related forms w_i and w_j derived from the same lexical root:

$$BCS(w_i) = BCS(w_j)$$

while:

$$\text{IIVS}(w_i) \neq \text{IIVS}(w_j)$$

This enables semantically related forms to preserve stable lexical structure despite extensive inflectional and derivational variation.

C.7 Reversibility

MEIV preserves reversibility through explicit transformation markers and structured representation. Let:

$$M^{-1}$$

denote the inverse MEIV transformation. Then:

$$M^{-1}(M(w)) = w$$

for valid surface forms w.

This property ensures that MEIV preserves the original surface representation while enabling stable tokenization-compatible lexical structure.

D MEIV Converter Training Details

We initialize the model from the pretrained `google/byt5-small` checkpoint (Xue et al., 2022) using the HuggingFace Transformers framework (Wolf et al., 2020). The annotated dataset contains 5,300 manually aligned surface–MEIV transformation pairs divided into training, validation, and test splits.

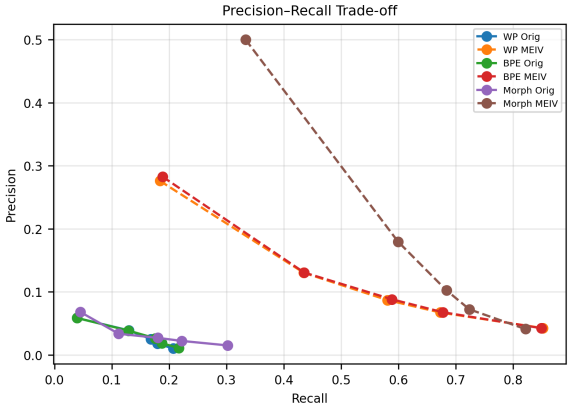


Figure 6: Precision–Recall curves for BM25 retrieval under original and MEIV representations.

Training is performed for 20 epochs using AdamW with a learning rate of 3×10^{-4}, batch size 16, and 100 warmup steps. Validation and checkpointing are performed every 200 steps, and decoding uses autoregressive generation with a maximum output length of 64 tokens.

Hyperparameter	Value
Model	ByT5-small
Epochs	20
Batch size	16
Learning rate	3×10^{-4}
Warmup steps	100
Validation interval	200 steps
Maximum generation length	64

Table 13: Training hyperparameters for the MEIV converter model.

E Precision Recall Curve

Figure 6 presents Precision–Recall curves for BM25 retrieval under original and MEIV representations. Across all tokenizers, MEIV consistently improves retrieval quality by stabilizing lexical representations across morphological variants.

F Annotation Protocol

Annotators were instructed to segment each surface form into prefix, stem, and suffix components, apply MEIV transformation rules only to the non-concatenative stem, and preserve reversibility of the original surface form. The annotation process emphasized consistent separation of lexical consonantal structure from inflectional vowel variation while maintaining valid reconstruction of the original surface representation. Annotation disagreements were resolved through discussion with Tigrigna language experts at Mekelle University.