Tokenization and Morphological Fidelity in Uralic NLP: A Cross-Lingual Evaluation

Nuo Xu
University of Eastern Finland
xnuo@student.uef.fi

Ahrii Kim*
AI-Bio Convergence Research Institute
Soongsil University
ahriikim@ssu.ac.kr

Abstract

Subword tokenization critically affects Natural Language Processing (NLP) performance, yet its behavior in morphologically rich and low-resource language families remains underexplored. This study systematically compares three subword paradigms—Byte Pair Encoding (BPE), Overlap BPE (OBPE), and Unigram Language Model—across six Uralic languages with varying resource availability and typological diversity.

Using part-of-speech (POS) tagging as a controlled downstream task, we show that OBPE consistently achieves stronger morphological alignment and higher tagging accuracy than conventional methods, particularly within the Latin-script group. These gains arise from reduced fragmentation in open-class categories and a better balance across the frequency spectrum. Transfer efficacy further depends on the downstream tagging architecture, interacting with both training volume and genealogical proximity.

Taken together, these findings highlight that morphology-sensitive tokenization is not merely a preprocessing choice but a decisive factor in enabling effective cross-lingual transfer for agglutinative, low-resource languages.

1 Introduction

Subword tokenization plays a key role in modern multilingual Natural Language Processing (NLP) as it determines how linguistic information is segmented and represented for downstream models. In multilingual settings, tokenizers are typically trained to balance vocabulary efficiency and coverage across many languages, often under strong data imbalance (Selvamurugan et al., 2025; Downey et al., 2024; Limisiewicz et al., 2023; Petrov et al., 2023).

This design choice poses particular challenges for morphologically rich languages, where grammatical functions are encoded through extensive inflection and agglutination. In such languages, inappropriate segmentation can obscure morpheme boundaries, inflate sequence length, and hinder cross-lingual transfer (Teklehaymanot et al., 2025; García et al., 2025; Asgari et al., 2025), affecting downstream tasks such as part-of-speech (POS) tagging (Libovický and Helcl, 2024) and machine translation (Kim and Kim, 2022).

These issues are particularly pronounced in the Uralic language family, which combines highly productive morphology with severe data imbalance across languages spanning a wide spectrum of resource availability—from relatively high-resource languages such as Finnish and Hungarian to severely low-resource languages such as Northern Sami and Komi-Zyrian (Chelombitko and Komissarov, 2024; Tereschenko et al., 2025). Recent studies show that standard multilingual tokenizers allocate disproportionately little effective vocabulary capacity to Uralic languages (Chelombitko and Komissarov, 2024). While alternative strategies such as Overlap BPE (OBPE) have been proposed that encourage shared subword units between high-resource and low-resource languages (Patil et al., 2022), existing work typically considers tokenization strategies in isolation or on limited language pairs, leaving it unclear how different approaches compare across the morphological and resource diversity of the Uralic language family.

In this work, we conduct a systematic comparison of recent and widely adopted tokenization strategies across six Uralic language pairs. We train tokenization models in the Universal Dependencies dataset and evaluate the impact of different tokenization choices on downstream task performance. Through this comparative analysis, we aim to clarify how tokenization strategies interact with morphological complexity and resource imbalance

*Corresponding author.

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 493–503

in Uralic languages. We release our codes.[1]

Our contributions are threefold:

- We conduct a controlled comparison of three tokenization paradigms—BPE, Unigram, and OBPE—across six Uralic languages spanning high-, mid-, and low-resource conditions, revealing how resource imbalance affects segmentation behavior.
- We present the first systematic evaluation of OBPE for Uralic languages, demonstrating its consistent gains in morphological alignment and cross-lingual transfer efficiency over conventional multilingual tokenizers.
- We link intrinsic segmentation quality with extrinsic POS tagging performance, showing that morphology-aware tokenization leads to measurable improvements in downstream accuracy, particularly under low-resource and typologically related conditions.

2 Related Work

This section addresses challenges arising from low-resource and morphologically rich languages in multilingual contexts (§ 2.1), and examines how these issues are amplified within the Uralic case (§ 2.2).

2.1 Challenges in Multilingual Tokenization

Modern multilingual NLP systems rely on data-driven segmentation algorithms such as BPE (Sennrich et al., 2016) and the Unigram language model (Kudo, 2018). Although BPE can be formally understood as a compression algorithm solving a combinatorial optimization problem (Zouhar et al., 2023), its linguistic consequences are far from neutral. In morphologically rich languages, statistical subword merges often fail to align with true morpheme boundaries (Bostrom and Durrett, 2020; Arnett et al., 2026), fragmenting semantic units into opaque sequences (Hofmann et al., 2021) and weakening compositional generalization on rare forms (Wolleb et al., 2023). Studies on Turkish and Finnish show that, for certain architectures, maintaining full word boundaries can outperform subword segmentation in low-resource conditions because excessive fragmentation disperses semantic information (Hu, 2025).

Beyond morphology, tokenization also encodes structural inequities across languages. Standard vocabulary generation procedures maximize global compression efficiency, which implicitly favors high-resource languages that dominate the training corpus (Foroutan et al., 2025). Consequently, text from low-resource languages is tokenized into longer sequences, amplifying computational costs and degrading model accuracy. To address this, "Parity-aware BPE" (Foroutan et al., 2025) modifies the merge selection criterion to improve compression for the worst-off languages, thereby enhancing cross-lingual fairness . Similarly, OBPE (Patil et al., 2022) adjusts the BPE objective to encourage shared tokens across related languages, trading a small loss in global compression for greater cross-lingual consistency.

2.2 Tokenization in Uralic Languages

Recent studies show that massively multilingual models struggle to adapt to Uralic languages for two main reasons. First, the "curse of multilinguality" reduces per-language capacity as more languages are added (Downey et al., 2024), and vocabulary coverage remains poor for morphologically rich forms (A Pirinen, 2024; Hämäläinen, 2019). Second, cross-lingual transfer is constrained by genealogical distance; effective transfer requires structural similarity, largely absent between Uralic and Indo-European languages (Bankula and Bankula, 2025).

Recent work has addressed these challenges through two main approaches: statistical vocabulary adaptation and linguistically informed segmentation. In the statistical paradigm, specialized monolingual tokenizers improve compression and lexical coverage for Northern Sámi and Estonian compared to multilingual baselines (Chelombitko and Komissarov, 2024), and compact, domain-adapted vocabularies often outperform large generic ones (Downey et al., 2024).

In the linguistic paradigm, morphology-aware tokenizers explicitly integrate morphological boundaries into segmentation. Asgari et al. (2025) introduced *MorphBPE*, which optimizes merges using morphological cues, while García et al. (2025) showed that morphology-aware segmentation enhances language modeling for agglutinative languages. Cross-linguistic analyses further reveal that tokenizers aligned with morphological units yield more consistent downstream performance (Arnett et al., 2026).

Overall, while statistical adaptation enhances efficiency, morphology-aware segmentation better

[1] https://github.com/farfromshallow/Uralic-language-NLP.git

preserves linguistic structure—an aspect this study systematically evaluates across Uralic languages.

3 Subword Tokenization Framework

This study evaluates three representative tokenization paradigms—Byte Pair Encoding (BPE), Unigram Language Model, and Overlap-Based BPE (OBPE)—within the context of multilingual and low-resource language modeling. All three methods aim to optimize the trade-off between vocabulary compactness and representational adequacy, yet differ fundamentally in their learning objectives, inference dynamics, and treatment of morphological variation.

Byte Pair Encoding. BPE (Sennrich et al., 2016) is an agglomerative algorithm adapted from data compression (Gage, 1994), which iteratively merges the most frequent adjacent symbol pair (A, B) until the target vocabulary size is reached. Formally, it maximizes the compression utility (Zouhar et al., 2023), reducing total encoded corpus length. By design, BPE produces deterministic segmentations, ensuring stability but often merging across morphological boundaries. While its frequency-driven merges are highly efficient for frequent tokens, this greediness leads to *morphological opacity*, where semantically distinct morphemes are concatenated simply due to frequent co-occurrence (Toraman et al., 2023). Such behavior disproportionately harms low-resource languages, where co-occurrence statistics are sparse and biased toward dominant language distributions (Chelombitko and Komissarov, 2024). Recent adaptations attempt to mitigate these effects through adaptive merge thresholds and entropy-based pre-tokenization (Hu et al., 2025), yet standard BPE remains sensitive to corpus imbalance.

Unigram Language Model. The Unigram tokenizer (Kudo, 2018) follows a probabilistic, subtractive approach designed to maximize the marginal likelihood of the training corpus. Unlike agglomerative algorithms such as BPE, Unigram begins with a large seed vocabulary $\mathcal{V}_0$ (e.g., all substrings) and iteratively prunes tokens based on their contribution to the overall likelihood:

$$\mathcal{L}(\mathcal{V}) = \sum_{X \in D} \log \sum_{\mathbf{x} \in \text{Seg}(X)} \prod_{x_i \in \mathbf{x}} P(x_i). \quad (1)$$

The optimization alternates between estimating subword usage (E-step) and re-estimating probabilities with pruning (M-step). This procedure implicitly favors morphologically coherent subwords, as meaningful morphemes tend to have higher independent probabilities (Bostrom and Durrett, 2020). Its stochastic segmentation via *subword regularization* further exposes the model to multiple valid tokenizations of the same surface form, improving robustness in morphologically productive or noisy settings (Kudo, 2018; Vemula et al., 2025). Recent studies confirm that Unigram consistently outperforms deterministic schemes such as BPE in morphologically rich and agglutinative languages (Vemula et al., 2025).

Overlap-Based BPE. OBPE (Patil et al., 2022) reformulates BPE's merge criterion to promote lexical parity in multilingual corpora. In standard multilingual BPE, tokens from high-resource languages dominate due to higher frequency counts, fragmenting low-resource languages into longer sequences. OBPE introduces a dual optimization objective that jointly maximizes compression and cross-lingual overlap by aggregating token frequencies using a generalized mean:

$$\text{Overlap}(L_i, L_h, S) = \sum_{k \in S} \left(\frac{f_{ki}^p + f_{kh}^p}{2} \right)^{1/p}, \quad p \leq 1. \quad (2)$$

When $p \to -\infty$, the criterion prioritizes merges that maximize the minimum shared token frequency across languages—favoring lexical forms that exist in both corpora. This ensures that vocabulary growth benefits both high- and low-resource languages, preventing representational dominance. Recent work shows that such overlap-aware tokenization reduces sequence length disparity and improves cross-lingual generalization in morphologically rich, low-resource families (Karthika et al., 2025; Limisiewicz et al., 2023).

4 Experiment Setup

4.1 Dataset

We use Uralic treebanks from the Universal Dependencies (UD) v2 collection (Nivre et al., 2020), a standardized framework that provides linguistically consistent annotations for multilingual NLP tasks, including part-of-speech and dependency relations.

The selected datasets vary widely in size, ranging from 6,475 down to a mere 397 sentences for Komi-Zyrian. In terms of textual domain, the majority of the corpora consist of newswire text. How-

Language	Code	Resource	Train	Dev	Test
● **Latin script languages**					
Finnish	fin	HIGH	6,475	769	809
Estonian	est	HIGH	5,444	833	913
Hungarian	hun	MID	910	441	449
North Sámi	sme	LOW	1,873	624	625
● **Cyrillic script languages**					
Russian	rus	HIGH	2,080	589	813
Komi-Zyrian	kpv	UNDER	397	133	133

Table 1: Dataset statistics for all languages (unit: sentences). Resource availability is indicated by categorical tags; UNDER denotes severely under-resourced languages. Hungarian is treated as a mid-resource control due to its relatively small dataset size.

ever, significant variations exist: the Komi-Zyrian dataset is derived exclusively from fiction, whereas the high-resource anchor languages (Finnish and Estonian) exhibit broader genre coverage, including blogs, wikis, and social media content.

This severe sparsity, particularly for Komi-Zyrian, imposes strict constraints on model complexity. To account for this disparity, we adopted a dynamic data partitioning strategy. For languages with sufficient resources, we applied the conventional 8:1:1 split for training, development, and testing. For extremely low-resource treebanks, we used a 6:2:2 ratio, following Sheyanova and Tyers (2017), to ensure stable evaluation and mitigate lexical overfitting. Although this reduces the amount of training data, it guaranties adequate test coverage and statistical reliability across all resource levels.

4.2 Language Pairs

To evaluate cross-lingual adaptation under different tokenization schemes, we selected high-resource *source* and low-resource *target* languages with varying degrees of linguistic similarity. In all experiments, models were first trained on the source language and subsequently finetuned on the target language. Languages were additionally grouped by script—Latin and Cyrillic—to eliminate confounding factors from mixed writing systems (Tufa et al., 2024). All selected languages are morphologically rich and predominantly agglutinative, where grammatical information is expressed through extensive suffixation. This property makes them particularly sensitive to subword segmentation and therefore well-suited for analyzing tokenizer effects.

For the Latin-script group (●), Finnish and Estonian serve as source languages. Both are Uralic languages with highly productive case systems and complex verbal morphology. They are paired with typologically distinct targets to explore different transfer scenarios under the same adaptation protocol. North Sámi, a closely related Uralic language, enables evaluation of transfer between structurally similar but resource-imbalanced languages. Hungarian, by contrast, represents a more distant Ugric branch with mixed agglutinative–fusional morphology and a rich inflectional system. This setup allows us to test whether structural similarity or morphological typology better predicts adaptation effectiveness when lexical overlap is limited.

For the Cyrillic-script group (●), we use Russian as the source and Komi-Zyrian as the target. Although genealogically unrelated—Russian being Indo-European and Komi-Zyrian Uralic—they share the Cyrillic alphabet and a history of geographic contact. Russian exhibits fusional morphology, while Komi-Zyrian retains agglutinative Uralic structure with extensive case marking. This pairing isolates the effect of shared script from genealogical similarity, revealing whether orthographic commonality alone facilitates adaptation.

4.3 Preprocessing

To ensure consistent alignment between subword tokenization and gold-standard annotations, we adopt a three-stage preprocessing pipeline widely used in subword-level sequence labelling.

Gold-Standard Extraction. We extracted the FORM column from the UD CoNLL-U files to preserve the original token boundaries while omitting metadata. This approach maintains morphological integrity for complex tokens, such as hyphenated compounds (e.g., `100-aastased`) and abbreviations (e.g., `4:sta`), ensuring strict consistency with the gold segmentation defined in the treebank (Chiarcos and Schenk, 2019).

Greedy Alignment. We then aligned tokenizer outputs to the gold tokens through character-level greedy matching after Unicode normalization (NFKC). Implementation-specific boundary markers, such as the SentencePiece _ (U+2581) and OBPE `</w>` symbols, were removed prior to alignment, following common alignment strategies used in multilingual parsing pipelines (Rosa and Mareček, 2018; Che et al., 2018).

First-Subword Tagging. Finally, gold token-level labels (e.g., POS tags) were projected to

subword sequences using a first-subword tagging scheme (Devlin et al., 2019; Pires et al., 2019), where each token label is assigned to its first subword while subsequent subwords are padded. This choice may lead to information loss, particularly in Uralic languages, which often express grammatical categories through suffixes. Nonetheless, we consider this normalization necessary to obtain comparable metrics across tokenizers with different levels of granularity. We further assume that, despite this limitation, the effect is alleviated by the bidirectional nature of the BiLSTM encoder.

4.4 Training Setup

We trained the standard BPE and Unigram LM tokenizers using the SentencePiece toolkit (Kudo and Richardson, 2018). Each tokenizer was trained on monolingual data from its corresponding language pair to avoid cross-lingual contamination.

For OBPE, we adopted the configuration proposed by Patil et al. (2022) to promote lexical equity across languages. We assigned equal weights ($\alpha = 0.5$) to the compression and overlap objectives, and set the generalized mean exponent to $p = -\infty$ to prioritize merges that maximize the minimum shared token frequency between paired languages. To maintain comparable capacity across all conditions, the vocabulary size was fixed at $|\mathcal{V}| = 5{,}000$ subword units for all models. Although this limit is smaller than standard industrial vocabularies used for high-resource agglutinative languages, it was required by the severe data sparsity of our target datasets. In a pilot study, we observed that scaling the vocabulary to 8,000 operations caused rapid lexical overfitting in the lowest-resource Komi-Zyrian. Consequently, we adopted the 5,000 limit as a strategic compromise to prioritize model generalizability across the diverse resource tiers in our evaluation.

Cross-lingual training setup. To evaluate how tokenizer design affects cross-lingual adaptation, we trained all models on a high-resource source language and subsequently finetuned them on each low-resource target language. This setup allows us to measure how effectively representations learned from the source language transfer to the target language under different tokenization schemes.

4.5 Evaluation on Downstream POS Task

We evaluate tokenizer performance through POS tagging, a controlled extrinsic task that reflects how well subword segmentation supports morphosyntactic learning in agglutinative Uralic languages. To ensure robustness across architectures, we employ two complementary sequence labeling models: a BiLSTM–CRF (Lample et al., 2016; Ma and Hovy, 2016) and Flair (Akbik et al., 2018, 2019). The BiLSTM–CRF relies solely on local contextual representations learned from subword embeddings, providing a clean testbed for segmentation sensitivity, while Flair uses character-level contextual string embeddings that capture long-range dependencies and subword-internal regularities, which is particularly advantageous for morphologically rich languages.

We report Accuracy and Macro-F1 as evaluation metrics. Accuracy reflects overall tagging correctness, whereas Macro-F1 assigns equal weight to all POS categories, preventing frequent tags from dominating the evaluation. Together, these metrics allow us to assess how tokenizer design influences both general and category-sensitive POS performance.

5 Results

The general performance. Table 2 reports tokenizer performance on POS tagging. OBPE generally surpasses both BPE and Unigram baselines in terms of accuracy and F1 scores in most settings. The main exception is the Cyrillic group, where Unigram consistently performs better. In addition, Flair achieves stronger results with BPE on Hungarian. These observations will be examined in more detail later.

The long-tail issue. A marked discrepancy between Accuracy and Macro-$F1$ in the Hungarian results highlights a structural limitation of frequency-based segmentation. Although standard BPE attains high Accuracy (0.8096), its Macro-$F1$ (0.7013) lags substantially behind OBPE (0.7412). This disparity illustrates the *long-tail* challenge in agglutinative morphology: a small set of central categories dominates the corpus, while numerous complex inflected forms occur only rarely. Because standard BPE seeks to minimize the corpus description length, it preferentially segments the frequent stems of the dominant classes, thus achieving strong frequency-weighted accuracy (Gutierrez-Vasques et al., 2023).

Since Macro-$F1$ weights all classes equally, independent of how frequent they are, it penalizes models that fail to generalize to rare morphologi-

Source	Target	Sim.	Tokenizer	BiLSTM–CRF		Flair	
				Acc	Mac F1	Acc	Mac F1
● est	hun	✗	BPE	0.8096	0.7013	0.9509	0.7930
			Unigram	0.7840	0.6663	0.9408	0.7651
			OBPE	0.8496	0.7398	0.9614	0.7902
● est	sme	✓	BPE	0.7749	0.7573	0.9075	0.8050
			Unigram	0.7830	0.7573	0.9078	0.7885
			OBPE	0.8152	0.7850	0.9373	0.8390
● fin	hun	✗	BPE	0.8096	0.7013	0.9509	0.7930
			Unigram	0.7840	0.6663	0.9408	0.7651
			OBPE	0.8514	0.7412	0.9581	0.7907
● fin	sme	✓	BPE	0.7749	0.7573	0.9075	0.8050
			Unigram	0.7830	0.7573	0.9078	0.7885
			OBPE	0.8036	0.7914	0.9264	0.8164
● rus	kpv	✗	BPE	0.6744	0.4742	0.3941	0.4409
			Unigram	0.7401	0.5209	0.9101	0.6367
			OBPE	0.7207	0.5022	0.8930	0.5693

Table 2: POS tagging performance across source–target pairs. Target pairs with North Sámi (✓) exhibit high linguistic similarity.

cal patterns. By contrast, OBPE uses cross-lingual anchors to better maintain these infrequent affix boundaries, resulting in a higher Macro-$F1$ that reflects strong performance across the entire range of morphological detail.

Cyrillic performance gap. We attribute the performance gap in the Cyrillic group primarily to the substantial linguistic distance between the source (Russian) and target (Komi-Zyrian) languages. While centuries of contact have introduced Russian loanwords into Komi-Zyrian, the core grammatical inventory remains distinct (Leinonen, 2006). There is structural discordance between the fusional morphology of Russian and the agglutinative typology of Komi-Zyrian. Russian suffixes often encode case, number, and gender into a single fused unit; for instance, the ending *-am* in *stol-**am*** [2](tables) simultaneously marks dative case and plural number. In contrast, Komi-Zyrian concatenates distinct suffix chains for each category in the form of Root-Num-Case, as seen in *pezan-**jas**-**li***, where the plural (*-jas*) and dative (*-li*) markers remain distinct segments. Furthermore, Russian relies heavily on free-standing prepositions to express spatial relations, whereas Komi-Zyrian encodes these relations exclusively through bound case suffixes like most Uralic languages. This asymmetry creates a bottleneck for OBPE: the overlap objective cannot find the shared morphological anchors despite the shared script. Thus, the shared Cyrillic script acts as a "false friend".

Category level effects We report detailed per-tag metric scores in Table 3. The results indicate that tokenizer-related discrepancies occur mainly in open-class, morphologically productive tag categories.

In North Sámi, switching from BPE to OBPE leads to the largest improvements in morphologically complex open-class categories. For ADJ, BPE attains an F1 of just 0.5173, while OBPE (sme-fi) increases this to 0.6440. NOUN tagging similarly benefits, rising from 0.7200 (BPE) to 0.7681 (OBPE). These gains reflect the differing vocabulary construction goals of the two methods.

Functional categories like PUNCT and CCONJ remain highly stable across tokenizers ($F1 > 0.98$ for North Sámi). **This indicates that tokenization choice mainly affects morphologically rich categories, not fixed-vocabulary function classes.** Since both BPE and Unigram are designed to keep high-frequency items intact in the vocabulary (Sennrich et al., 2016; Kudo, 2018), these closed-class tokens are almost always encoded as single units in every model, leading to consistently similar performance.

The limitations of BPE in extremely low-resource conditions are clearly illustrated by the

[2]Cyrillic examples are transliterated using the transliteration toolset provided by the COPIUS initiative: https://www.copius.eu/ortho.php.

POS	sme								kpv						hun							
	BiLSTM				Flair				BiLSTM			Flair			BiLSTM				Flair			
	BPE	Uni	OFin	OEst	BPE	Uni	OFin	OEst	BPE	Uni	OBPE	BPE	Uni	OBPE	BPE	Uni	OFin	OEst	BPE	Uni	OFin	OEst
ADJ	0.517	0.583	**0.644**	0.637	0.638	0.594	0.718	0.667	0.412	0.409	0.350	0.449	0.427	0.464	0.627	0.789	**0.732**	0.716	0.835	0.789	0.830	0.834
ADP	0.750	0.717	0.764	0.767	0.829	0.728	0.749	0.772	0.425	0.595	0.587	0.600	0.675	0.590	0.856	0.907	0.888	0.860	0.920	0.907	0.897	0.905
ADV	0.753	0.753	0.732	0.748	0.822	0.782	0.789	0.784	0.597	0.654	0.639	0.696	0.731	**0.722**	0.774	0.754	0.822	**0.806**	0.866	0.833	0.868	0.850
AUX	0.785	0.789	0.791	0.810	0.807	0.811	0.811	0.825	0.737	0.768	0.721	0.775	0.777	0.807	0.750	0.792	0.816	0.779	0.812	0.750	0.765	0.808
CCONJ	0.987	0.968	0.990	0.984	0.990	0.958	0.967	0.977	0.798	0.732	0.834	0.798	0.850	0.859	0.939	0.919	0.949	0.947	0.944	0.939	0.947	0.927
INTJ	0.857	0.857	0.857	0.667	0.400	0.667	0.667	0.857	0.000	0.000	0.000	0.000	0.000	0.000	–	–	–	–	–	–	–	–
NOUN	0.720	0.739	**0.768**	0.784	0.809	0.769	0.810	0.831	0.587	0.679	0.632	0.298	0.766	**0.734**	0.844	0.736	0.812	0.811	0.887	0.844	0.876	0.882
NUM	0.630	0.563	0.594	0.646	0.706	0.671	0.667	0.740	0.455	0.556	0.556	0.400	0.476	0.133	0.895	0.832	0.736	0.809	0.895	0.833	0.810	0.860
PART	0.699	0.750	0.805	0.778	0.838	0.773	0.813	0.803	–	–	–	–	–	–	0.857	0.490	0.867	0.903	0.857	0.903	0.968	0.933
PRON	0.922	0.894	0.912	0.931	0.948	0.933	0.947	0.943	0.758	0.802	0.757	0.629	0.893	0.809	0.681	0.673	0.713	0.682	0.781	0.732	0.806	0.776
PROPN	0.589	0.504	0.665	0.680	0.677	0.612	0.690	0.741	0.000	0.000	0.000	0.000	0.000	0.000	0.781	0.670	0.816	0.833	0.880	0.811	0.895	0.904
PUNCT	0.999	0.997	0.995	0.998	0.999	0.994	0.993	0.999	0.992	0.985	0.994	0.997	0.997	0.985	0.998	0.992	0.997	0.997	0.998	0.997	0.991	0.996
SCONJ	0.742	0.798	0.859	0.840	0.857	0.808	0.861	0.868	0.829	0.739	0.800	0.571	0.857	0.737	0.953	0.911	0.949	0.928	0.953	0.902	0.941	0.922
VERB	0.653	0.693	0.704	0.721	0.761	0.731	0.767	0.780	0.523	0.695	**0.665**	0.375	0.772	**0.780**	0.891	0.708	0.799	0.804	0.891	0.805	0.888	0.873
Macro Avg	0.757	0.757	0.791	0.785	0.805	0.789	0.816	0.839	0.508	0.558	0.538	0.441	0.6367	0.5693	0.701	0.666	0.741	0.740	0.793	0.765	0.790	0.790

Table 3: POS-wise F1 across three cross-lingual settings (sme, kpv, hun). OFin/OEst denote OBPE trained on Finnish/Estonian.

Komi-Zyrian results in Table 3. The BPE model performs disastrously on VERB ($F1 = 0.375$), while the OBPE model captures substantially more verbal structure ($F1 = 0.780$). This aligns with Bostrom and Durrett (2020) showing that BPE's deterministic merge operations are ill-suited to settings where the available data are too sparse to derive reliable frequency statistics for complex morphologies.

6 Further Study

Although dataset size sets a basic upper bound on model performance, our analysis shows that Tag Diversity plays an important compensatory role. To operationalize this notion, we computed POS Entropy (H) from the tag distribution in the training data (see Table 4 for summary statistics). Specifically, for a tag set T and tag probabilities $P(t)$:

$$H(T) = -\sum_{t \in T} P(t) \log_2 P(t) \qquad (3)$$

Higher H values correspond to a more even spread of grammatical categories, thereby providing wider morphological coverage.

Hungarian illustrates this relationship clearly. Despite its dataset being roughly half the size of the North Sámi corpus (910 vs. 1,873 sentences), it exhibits higher syntactic entropy ($H \approx 2.275$ vs. 2.195). As a result, the performance degradation is relatively modest given the data reduction: OBPE attains a Macro-$F1$ of 0.741 on Hungarian, remaining close to the North Sámi score of 0.815 even with 50% fewer training instances. In contrast, Komi (kph), which is affected by both severe data sparsity (397 sentences) and the lowest entropy ($H \approx 2.018$), shows a pronounced performance loss (Unigram $F1$: 0.521). These findings suggest that languages with richer and more evenly distributed tag inventories enable tokenizers to generalize better under limited supervision, while low-entropy distributions intensify the impact of data scarcity.

Tokenizer	sme	hun	kph
Train size / POS Ent.	*1873 / 2.195*	*910 / 2.275*	*397 / 2.018*
BPE	**0.757**	0.701	0.674
Unigram	**0.783**	0.666	0.521
OBPE (fi)	**0.804**	0.741	–
OBPE (et)	**0.815**	0.7398	–
OBPE (rus)	–	–	0.5022

Table 4: BiLSTM Macro-$F1$ across languages. The top section details training size and POS entropy for each language.

7 Qualitative Analysis

We present three categories of morphologically challenging examples from Uralic languages and show how OBPE addresses them effectively compared to other approaches.

- **Agglutination:** a process where complex words are formed by concatenating distinct suffixes to a single root stem. As illustrated in Table 5, a single Uralic token often conveys information that would require an entire phrase in English. For instance, the Hungarian token *szervezeteire* exhibits a massive accumulation of suffixes: it combines the root *szervezet* (organization) with the possessive marker *-e*, the plural marker *-i*, and the case marker *-re* This morphological density results in an explosion of word forms. This poses a severe challenge for frequency-based tokenizers like BPE: the specific combination may appear only once in the corpus, preventing the model from learning a stable representation for the frequent

Phenomenon	Segmentation	Gloss
Agglutination	*szervezet-e-i-re* (ORG+POSS+PL+CASE)	"to its organizations"
Vowel Harmony	*globalizáció* + *-ra* *rendőrség* + *-re*	"to globalization" "to the police"
Consonant Gradation	*veahkki* (lemma) *veahki*	"help"

Table 5: Typological challenges extracted from the train corpora: agglutination, vowel harmony, and consonant gradation.

constituent parts.

- **Vowel Harmony:** Suffixes must match the harmonic class (front/back) of the root. This forces the tokenizer to learn multiple allomorphs (e.g., *-ban* vs. *-ben*). Crucially, this affects tokenization by decreasing the statistical signal of grammatical markers.
- **Consonant Gradation:** In contrast to harmony, North Sami exhibits consonant gradation, where the stem's internal consonants change length or quality based on the grammatical context. For instance, the lemma *veahkki* ("help") appears as *veahki* in the accusative case (see Table 5). This stem alternation disrupts the static subword patterns that BPE relies on for consistent root recognition.

As in Table 5, the Hungarian token *szervezeteire* encapsulates a root plus three distinct morphemes. This structure creates a explosion of word forms, resulting in a "long tail" where rare suffix combinations evade the frequency thresholds of standard tokenizers like BPE, leading to over-segmentation.

8 Conclusion

This study presented a controlled evaluation of subword tokenization strategies across six Uralic languages, clarifying how morphology and resource imbalance shape tokenization behavior. Our findings reveal that tokenization is not a neutral preprocessing step but **a decisive factor in low-resource and morphologically rich contexts**.

Three major insights emerge. First, genealogical proximity amplifies cross-lingual transfer: Overlap BPE (OBPE) achieves substantial gains when low-resource languages share a structural base with their source languages (e.g., Finnish–North Sámi), but not when such alignment is absent (e.g., Russian–Komi-Zyrian). Second, in isolated low-resource settings, the Unigram model provides more faithful morphological segmentation than BPE, owing to its probabilistic pruning and subword regularization. Third, these improvements are concentrated in open-class categories—nouns and verbs—where morphological complexity most challenges standard segmentation.

Taken together, our findings indicate that standard BPE is often ill-suited for Uralic and other highly agglutinative language families, a limitation that is likely to become more pronounced in real-world multilingual applications. As a practical guideline, we recommend OBPE in genealogically grounded multilingual settings, and Unigram in isolated low-resource scenarios. This provides a simple yet effective strategy for morphology-sensitive tokenization in multilingual NLP.

Limitations

Our research concentrates on a particular subset of the Uralic language family, which constrains the direct applicability of our results to other, typologically different language groups. In addition, the very limited amount of labeled data—especially for Komi-Zyrian (397 sentences)—reduces the statistical reliability of our experiments and makes extensive cross-validation infeasible.

To address the severe data sparsity in Komi-Zyrian, we fixed the tokenizer vocabulary size to 5,000 subword units. We recognize that this is substantially smaller than the typical vocabulary sizes (30k or more) used in contemporary Large Language Models (LLMs). Accordingly, it remains unresolved whether the morphological benefits of OBPE would hold when scaled up to the much larger settings required for open-domain text generation.

Our evaluation is limited to POS tagging, a rel-

atively "shallow" syntactic task that primarily relies on local contextual signals. Future research must determine whether our conclusions also extend to more "deep" semantic tasks such as NLI or Machine Translation, where subword segmentation has a stronger impact on meaning-bearing elements.

Lastly, we highlight a possible domain bias in our cross-lingual transfer configuration. The high-resource source models are trained on newswire text, whereas the low-resource Komi-Zyrian corpus contains only fictional prose. This domain mismatch acts as a confound, since the models must simultaneously adapt to a new language and a different textual genre.

Discussion

While we attribute the primary performance gap to typological distance, we acknowledge that the distinct scripts introduce a confounding variable that is difficult to isolate in the current setup. A rigorous disentanglement of orthographic opacity from morphological incompatibility would require projecting the Cyrillic data into a shared phonemic space (e.g., via UPA or IPA transliteration). Such a control would isolate the morphological alignment process from script-specific artifacts, thereby clarifying whether the limitations of OBPE in this setting are intrinsic to the linguistic mismatch or exacerbated by orthographic divergence.

Besides, we must contextualize our results within the landscape of practical NLP. While OBPE outperforms baselines, the absolute scores for low-resource targets remains far below "production-ready" standards. For comparison, the top-performing system for English in the CoNLL 2018 Shared Task achieved a UPOS $F1$ of 0.959 (Zeman et al., 2018). This indicates that achieving reliable utility for downstream tasks in extreme low-resource settings will likely require complementary transfer techniques, such as adapters or syntactic projection, to bridge the gap between statistical significance and industrial usability.

Acknowledgment

This research was supported by G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317).

References

Flammie A Pirinen. 2024. Keeping up appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 123–131. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Catherine Arnett, Marisa Hudspeth, and Brendan O'Connor. 2026. Evaluating morphological alignment of tokenizers in 70 languages. In *Tokenization Workshop*.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *Preprint*, arXiv:2502.00894.

Ajitesh Bankula and Praney Bankula. 2025. Cross-linguistic transfer in multilingual nlp: The role of language families and morphology. *Preprint*, arXiv:2505.13908.

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. Association for Computational Linguistics.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64. Association for Computational Linguistics.

Iaroslav Chelombitko and Aleksey Komissarov. 2024. Specialized monolingual BPE tokenizers for Uralic languages representation in large language models. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 89–95. Association for Computational Linguistics.

Christian Chiarcos and Niko Schenk. 2019. CoNLL-Merge: Efficient Harmonization of Concurrent Tokenization and Textual Variation. In *2nd Conference*

on Language, Data and Knowledge (LDK 2019), volume 70 of *Open Access Series in Informatics (OASIcs)*, pages 7:1–7:14, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

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. Association for Computational Linguistics.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. Targeted multilingual adaptation for low-resource language families. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15647–15663. Association for Computational Linguistics.

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut, and Rico Sennrich. 2025. Parity-aware byte-pair encoding: Improving cross-lingual fairness in tokenization. *Preprint*, arXiv:2508.04796.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Alba Táboas García, Piotr Przybyła, and Leo Wanner. 2025. Exploring morphology-aware tokenization: A case study on Spanish language modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30493–30506, Suzhou, China. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages through the looking glass of BPE compression. *Computational Linguistics*, 49(4):943–1001.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608. Association for Computational Linguistics.

Jinfan Frank Hu. 2025. Tokenization strategies for low-resource agglutinative languages in word2vec: Case study on turkish and finnish. In *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, page 1–6. IEEE.

Yifan Hu, Frank Liang, Dachuan Zhao, Jonathan Geuter, Varshini Reddy, Craig W. Schmidt, and Chris Tanner. 2025. Entropy-driven pre-tokenization for byte-pair encoding. *Preprint*, arXiv:2506.15889.

Mika Hämäläinen. 2019. Uralicnlp: An nlp library for uralic languages. *Journal of Open Source Software*, 4(37):1345.

N J Karthika, Maharaj Brahma, Rohit Saluja, Ganesh Ramakrishnan, and Maunendra Sankar Desarkar. 2025. Multilingual tokenization through the lens of indian languages: Challenges and insights. *Preprint*, arXiv:2506.17789.

Ahrii Kim and Jinhyeon Kim. 2022. Vacillating human correlation of SacreBLEU in unprotected languages. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 1–15. Association for Computational Linguistics.

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.

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. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Marja Leinonen. 2006. *The russification of Komi*, pages 234–245. Number 27 in Slavica Helsingiensia. University of Helsinki, Finland.

Jindřich Libovický and Jindřich Helcl. 2024. Lexically grounded subword segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7420. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Rudolf Rosa and David Mareček. 2018. CUNI x-ling: Parsing under-resourced languages in CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 187–196. Association for Computational Linguistics.

Aishwarya Selvamurugan, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. From bias to balance: How multilingual dataset composition affects tokenizer performance across languages. In *Second Workshop on Language Models for Underserved Communities (LM4UC)*.

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. Association for Computational Linguistics.

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in north sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*, pages 66–75.

Hailay Kidu Teklehaymanot, Dren Fazlija, and Wolfgang Nejdl. 2025. MoVoC: Morphology-aware subword construction for Ge'ez script languages. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13131–13144, Suzhou, China. Association for Computational Linguistics.

Yehor Tereschenko, Mika Hämäläinen, and Svitlana Myroniuk. 2025. Evaluating OpenAI GPT models for translation of endangered UralicLanguages: A comparison of reasoning and non-reasoning architectures. In *Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages*, pages 131–139, Joensuu, Finland. Association for Computational Linguistics.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Wondimagegnhue Tufa, Ilia Markov, and Piek Vossen. 2024. Unknown script: Impact of script on cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 124–129. Association for Computational Linguistics.

Saketh Reddy Vemula, Sandipan Dandapat, Dipti Sharma, and Parameswari Krishnamurthy. 2025. Rethinking tokenization for rich morphology: The dominance of unigram over BPE and morphological alignment. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 232–252, Mumbai, India. Association for Computational Linguistics.

Benoist Wolleb, Romain Silvestri, Georgios Vernikos, Ljiljana Dolamic, and Andrei Popescu-Belis. 2023. Assessing the importance of frequency versus compositionality for subword-based tokenization in NMT. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 137–146. European Association for Machine Translation.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21. Association for Computational Linguistics.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. 2023. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614. Association for Computational Linguistics.