Neural Cross-Lingual Tokenization: A Morphologically-Aware Approach to Multilingual Subword Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Current multilingual language models rely on frequency-based tokenization methods like Byte Pair Encoding (BPE) and SentencePiece, which treat languages as mere symbol sequences without leveraging linguistic structure. While these approaches have achieved remarkable success through scale, they demonstrate significant limitations for morphologically complex languages, leading to over-segmentation, poor cross-lingual transfer, and suboptimal compression ratios. We propose a neural multilingual tokenization framework that learns morphologically-aware subword representations through cross-lingual consistency constraints and dynamic programming-based segmentation. Our approach addresses fundamental limitations of existing tokenizers by learning semantic equivalences between morphemes across languages, rather than relying solely on frequency statistics. We want to demonstrate that this method achieves superior compression ratios and cross-lingual transfer performance, particularly for agglutinative and morphologically rich languages that constitute over 40% of the world's linguistic diversity.

Keywords: multilingual NLP, tokenization, morphology, cross-lingual transfer, subword segmentation

1 Introduction

Tokenization serves as the fundamental preprocessing step in modern natural language processing pipelines, directly influencing model efficiency, linguistic fidelity, and cross-lingual transfer capabilities. The dominant paradigm in multilingual language models employs frequency-based subword tokenization methods such as Byte Pair Encoding (BPE) Sennrich et al. [2015] and SentencePiece Kudo and Richardson [2018], which have enabled the development of powerful multilingual models like mBERT Devlin et al. [2019] and XLM-R Conneau et al. [2019].

However, these statistical approaches treat natural language as mere symbol sequences, ignoring the rich morphological and semantic structure that characterizes human language. This limitation becomes particularly pronounced for morphologically complex languages, where a single root can generate hundreds of inflected forms through agglutination, fusion, or other morphological processes. Recent empirical evidence suggests that current multilingual models exhibit systematic performance gaps between morphologically simple languages (e.g., English) and complex ones (e.g., Turkish, Finnish, Arabic) Jabbar [2023].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

This paper introduces a neural multilingual tokenization framework that addresses these fundamental limitations through three key innovations:

1. Morphologically-aware representation learning that captures cross-lingual morphological equivalences
2. Convex optimization-based segmentation that optimizes for both compression and linguistic coherence
3. Typological clustering that enables zero-shot support for rare languages through linguistic prototype transfer

Our approach moves beyond frequency-based heuristics to learn principled subword segmentations that respect morphological boundaries while maximizing cross-lingual transfer potential.

2 Related Work and Current Limitations

2.1 Current State-of-the-Art

The current multilingual NLP landscape is dominated by large-scale transformer models trained with frequency-based tokenization:

- AYA (Cohere AI) Abagyan et al. [2025]: Massive multilingual tokenization based on byte-pair-encoding.[1]
- XLM-R Conneau et al. [2019]: Trained on 2.5TB of CommonCrawl data across 100 languages, achieving significant improvements over mBERT on cross-lingual benchmarks (+14.6% on XNLI, +13% on MLQA)
- mT5 Xue et al. [2020]: A multilingual text-to-text transformer covering 101 languages
- PaLM Chowdhery et al. [2023]: 540B parameter model demonstrating emergent multilingual capabilities

dynamic tokenizers

theoretical findings

2.2 Multi-lingual datasets

FLORES-200, mC4, AYA, HPLT

2.3 Multi-lingual evaluation datasets

MasakhaNER 2.0, XTREME, XTREME-UP, afaik AYA has some instruct-tuning datasets too, COSMMIC, HERCULE, Global MMLU (key dataset), MaXIFE.

2.4 Quantitative Evidence of Current Limitations

2.4.1 Morphological Segmentation Quality

Recent analysis reveals systematic failures in how current tokenizers handle morphologically complex languages:

Turkish Analysis Kaya and Tantuğ [2024]: XLM-R SentencePiece creates morphologically meaningless subwords. For example, Turkish "-bandela" is segmented into "ba", "#ndel", "#a", which lacks morphological coherence. Over-segmentation leads to 2-3× longer sequences compared to morphologically-aware alternatives.

[1] Cohere has many of the most recent papers on multilingual tokenization: https://cohere.com/research/papers

Arabic Challenges Tawfik et al. [2019]: Root-pattern morphology is poorly captured by frequency-based methods, with dialectal variations creating inconsistent tokenization within the same semantic space.

Japanese

2.4.2 Performance Gaps Across Language Families

Empirical evidence demonstrates systematic performance degradation for morphologically complex languages:

Language Type	XNLI Accuracy	Relative Performance
English (Fusional)	91.3%	Baseline
Turkish (Agglutinative)	79.1%	-13.4%
Finnish (Agglutinative)	81.2%	-11.1%
Arabic (Root-pattern)	77.8%	-14.8%

Table 1: XLM-R performance across language types showing consistent degradation for morphologically complex languages.

Key findings Jabbar [2023]: Consistent 10-15% performance degradation for agglutinative languages, with the gap persisting even with increased training data. Morphological alignment shows potential for 2-8% improvements.

2.4.3 Compression Efficiency Limitations

Current tokenizers demonstrate suboptimal compression for morphologically rich languages:

Language	Tokens/Word	Overhead
English	1.3	Baseline
Turkish	2.1	+61%
Finnish	1.9	+46%
Arabic	2.3	+77%

Table 2: Compression inefficiency directly translates to 30-70% more computational cost.

2.5 Fundamental Limitations of Frequency-Based Approaches

2.5.1 Lack of Semantic Awareness

BPE and SentencePiece optimize for:

$$\max \sum frequency(subword_i) \tag{1}$$

This objective ignores:

- Morphological boundaries
- Cross-lingual semantic equivalences
- Compositional meaning structure
- Corpora distribution (often towards English)

2.5.2 Language-Specific Bias

Training on concatenated multilingual corpora introduces systematic biases where high-resource languages dominate vocabulary, morphologically simple languages receive disproportionate representation, and cross-lingual patterns emerge accidentally rather than by design.

2.5.3 Extra papers

1. https://arxiv.org/pdf/2301.10472
2. https://arxiv.org/pdf/2403.10691
3. https://arxiv.org/pdf/2402.06619
4. https://cohere.com/research/papers
5. https://oldi.org/
6. https://arxiv.org/pdf/2305.11938
7. https://arxiv.org/abs/2506.15372
8. https://arxiv.org/abs/2405.07883
9. https://arxiv.org/pdf/2411.18553
10. https://arxiv.org/pdf/2412.15210
11. https://arxiv.org/pdf/2505.19599
12. https://arxiv.org/pdf/2410.13394
13. https://arxiv.org/pdf/2412.03304
14. https://arxiv.org/pdf/2503.10267
15. https://arxiv.org/pdf/2506.01776
16. https://arxiv.org/pdf/2003.11080

References

Diana Abagyan, Alejandro R Salamanca, Andres Felipe Cruz-Salinas, Kris Cao, Hangyu Lin, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. One tokenizer to rule them all: Emergent language plasticity via multilingual tokenizers. arXiv preprint arXiv:2506.10766, 2025.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113, 2023.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

Haris Jabbar. Morphpiece: A linguistic tokenizer for large language models. arXiv preprint arXiv:2307.07262, 2023.

Yiğit Bekir Kaya and A Cüneyd Tantuğ. Effect of tokenization granularity for turkish large language models. Intelligent Systems with Applications, 21:200335, 2024.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226, 2018.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

Ahmed Tawfik, Mahitab Emam, Khaled Essam, Robert Nabil, and Hany Hassan Awadalla. Morphology-aware word-segmentation in dialectal arabic adaptation of neural machine translation. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 11–17, 2019.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.