Morphology-Aware Tokenization Improves Turkish Sentence Embeddings under Controlled Training Conditions

M. Ali Bayram
Yıldız Technical University,
Turkey
malibayram20@gmail.com

Öner Aytaş
Işık University,
Turkey
oner.aytas@
isikun.edu.tr

Tuğçe Şen
Işık University
Turkey
tugce.sen@
isikun.edu.tr

Abstract

Tokenization is often treated as a fixed implementation detail in sentence embedding pipelines, even though morphologically rich languages place unusual pressure on subword segmentation. This paper studies tokenization as an isolated experimental variable for Turkish, an agglutinative and relatively underrepresented language in NLP. We evaluate a morphology-aware tokenizer (MFT) against three Turkish BPE-based tokenizers in a controlled setup where architecture, training data, and initial weights are fixed and only the tokenizer changes. Across this setup, MFT yields the best scores on STSbTR semantic similarity and on the 26-task Turkish embedding benchmark TR-MTEB, while also producing the strongest diagnostic average on TurBLiMP. The gains are accompanied by a clear cost: MFT uses more tokens per word than the BPE baselines. We argue that, for morphologically rich languages, linguistically aligned tokenization should be treated as a measurable design variable rather than a preprocessing default.

Keywords: morphology-aware tokenization, Turkish NLP, sentence embeddings, low-resource languages, evaluation

1 Introduction

Large language models and sentence encoders have advanced quickly for high-resource languages, but their behavior remains less well understood for morphologically rich and underrepresented languages. Turkish is a particularly revealing case: productive suffixation, phonological alternations, and long derived word forms can make purely frequency-driven subword segmentation linguistically misaligned. Prior Turkish work has already shown that tokenization choices matter for downstream performance and for the linguistic integrity of representations (Toraman et al., 2023; Schmidt et al., 2024; Bayram et al., 2025).

Despite this, tokenization is still often treated as a fixed preprocessing decision rather than as an experimental variable. This is problematic for embedding models, where representation quality depends on how consistently semantically related forms are segmented before the encoder is trained or adapted. Benchmarking practice has also improved: MTEB-style evaluation has made sentence encoders easier to compare across tasks (Muennighoff et al., 2022), while Turkish now has language-specific resources such as STSbTR and TR-MTEB (Beken Fikri et al., 2021; Baysan and Gungor, 2025). What remains less common is a controlled comparison in which the tokenizer is the only factor that changes.

This paper presents such a comparison. We evaluate a morphology-aware Turkish tokenizer, MFT, against three Turkish BPE-based tokenizers in a controlled training setup where all models share the same encoder architecture, the same training data, and the same initial random weights. The results show that morphology-aware tokenization leads to stronger Turkish sentence embeddings in semantic similarity and multi-task evaluation, while also improving a diagnostic measure of sensitivity to morphologically perturbed minimal pairs. At the same time, the results make the trade-off explicit: better linguistic alignment comes with a larger token footprint.

2 Morphology-Aware Tokenization as a Measurable Variable

The tokenizer studied here, MFT, is designed for Turkish and combines root–suffix analysis, phonological normalization, and a BPE fallback path when morphological decomposition is unreliable (Bayram et al., 2025). The main motivation is to reduce semantically opaque splits and preserve morphemic structure more consistently than general subword tokenization.

Two tokenization-oriented metrics are central in this setting. The first is Turkish Ratio (TR%), which measures the share of language-appropriate tokens in the produced vocabulary footprint. The second is Purity Ratio (Pure%), which tracks how many of these units remain morphologically clean rather than being mixed or noisy segments. On TR-MMLU text, MFT reaches TR%=90.29 and Pure%=85.80 with a vocabulary of 32,768 tokens. In addition, the encode–decode cycle remains nearly lossless, with 99.48% exact word-level accuracy for `decode(encode(w))` on a 66,547-word sample (Bayram et al., 2025).

These properties do not by themselves prove downstream utility, but they make tokenization quality measurable in a linguistically interpretable way. The central question of this paper is whether this internal alignment transfers to sentence embedding quality when the rest of the training pipeline is held constant.

3 Controlled Experimental Setup

To isolate tokenization from other factors, we train four sentence encoders under identical conditions and vary only the tokenizer. All models use the same EmbeddingGemma-300M architecture (Vera et al., 2025), the same training data, and the same initial random weights. A single seed-42 initialization is sampled once and copied across all runs; the only difference is the tokenizer stream: `mft`, `mursit`, `cosmos`, or `tabi`. To avoid runtime tokenization differences during training, input IDs are prepared offline for each tokenizer, so the comparison remains focused on segmentation choices rather than implementation artifacts.

The controlled training data consists of an approximately 300K-example Turkish text subset derived from the Cosmos Turkish Corpus, with teacher embeddings precomputed before student training. It is broad web/pretraining-style Turkish text rather than task-specific STS, retrieval, or classification data, so the evaluation benchmarks are out-of-domain with respect to the training source. For fairness, each example is encoded once with all four tokenizer streams, the resulting `input_ids` columns are stored offline, and any example exceeding the shared 2048-token context limit under any tokenizer is removed from all streams. Tokenizer coverage is therefore handled before training: the BPE baselines retain their usual open-vocabulary subword behavior, while MFT uses its BPE fallback when root–suffix analysis is unreliable; remaining rare or noisy forms mainly affect segmentation quality rather than producing uncovered tokens.

Measure	MFT	Best BPE
TR%	90.29	53.48
Pure%	85.80	31.45
`decode(encode(w))`	99.48	–
Tokens / word	2.91	1.82

Table 1: Main tokenizer-side properties. The BPE side is shown as the strongest reported baseline value where relevant. MFT is more linguistically aligned, but less compact.

We evaluate the resulting models on three levels:

- **STSbTR**, a Turkish semantic textual similarity benchmark scored with Pearson and Spearman correlation (Beken Fikri et al., 2021);
- **TR-MTEB**, a 26-task Turkish embedding benchmark covering retrieval, classification, clustering, STS, NLI, and bitext mining (Muennighoff et al., 2022; Baysan and Gungor, 2025);
- **TurBLiMP**, used here as a diagnostic benchmark where sentence embeddings are tested on controlled good/bad minimal pairs via centroid-based comparison rather than language model probabilities (Başar et al., 2025).

Table 1 summarizes the main tokenizer-side properties, including the efficiency cost of MFT.

4 Results

Table 2 gives the main downstream results. MFT yields the strongest STSbTR test performance, improving over the strongest BPE baseline (Mursit) by +4.81 Pearson and +4.31 Spearman. The ranking is consistent on the STSbTR training split as well, suggesting that the effect is not limited to a single subset.

The same pattern extends to TR-MTEB. MFT achieves the best average score across 26 tasks with 39.57, ahead of Mursit (35.92), Cosmos (35.65), and Tabi (34.92). The advantage is especially visible in STS and retrieval-oriented settings, while some category-level trade-offs remain: for example, Cosmos is slightly better on clustering. This is important because it shows that morphology-aware tokenization is not a universal win on every task,

Tokenizer	P	S	TR-MTEB	TurBLiMP
MFT	**51.44**	**50.03**	**39.57**	**61.2**
Mursit	46.63	45.72	35.92	52.7
Cosmos	43.09	42.18	35.65	50.8
Tabi	43.01	42.53	34.92	51.6

Table 2: Controlled downstream results. P and S denote Pearson and Spearman on STSbTR test. TR-MTEB is the 26-task average. TurBLiMP is the average pair accuracy over 16 phenomena.

but a design choice with a clear overall benefit profile.

TurBLiMP provides an additional diagnostic signal. Using an embedding-based good/bad sentence comparison, MFT reaches the best average over 16 phenomena (61.2), again outperforming the three BPE baselines. The phenomenon-level pattern is mixed, which cautions against overgeneralization, but the average trend supports the idea that morphology-aware segmentation makes the embedding space more sensitive to controlled formal degradations in Turkish.

5 Discussion

These results support a simple claim: for Turkish sentence embeddings, tokenization should be treated as a measurable modeling variable rather than a frozen preprocessing layer. When architecture, data, and initialization are fixed, the tokenizer still changes the quality of the learned embedding space in meaningful ways.

The most important practical trade-off is efficiency. MFT is less compact than the BPE baselines and consumes more tokens per word. In other words, the gains are not "free." However, the controlled results suggest that for a morphologically rich language such as Turkish, a better-aligned segmentation policy can yield stronger semantic representations even when sequence length becomes less favorable.

The results also clarify the scope of the claim. We do not argue that morphology-aware tokenization dominates on every task or that it replaces architectural and data-centric improvements. Rather, the evidence suggests that linguistically informed segmentation can provide a robust advantage in Turkish STS, retrieval-oriented embedding tasks, and morphology-sensitive diagnostics, while still leaving room for task-specific trade-offs. The TR-MTEB category breakdown is consistent with this narrower interpretation: the strongest gains are in STS and retrieval, whereas Cosmos is slightly better on clustering, so the main evidence is an overall average advantage rather than uniform task-level dominance.

A limitation of this study is that all controlled runs use a single shared seed; future work should repeat the comparison across multiple seeds. Models and evaluation artifacts will be made available through the Magibu Hugging Face organization at `https://huggingface.co/magibu`. The MFT tokenizer code is available at `https://github.com/malibayram/turkish-tokenizer`.

6 Conclusion

This paper presented a controlled comparison of four Turkish tokenizer streams for sentence embeddings and showed that a morphology-aware tokenizer leads to the strongest overall results on STSbTR, TR-MTEB, and TurBLiMP when all other training factors are fixed. The main contribution is methodological as much as empirical: tokenization can be isolated, measured, and compared as a first-class design variable in low-resource and morphologically rich language technology.

For future work, the main next step is to test whether this pattern transfers to other morphologically rich languages and to multilingual settings where language-specific alignment must coexist with shared vocabularies.

References

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, Savaş Yıldırım, and Demircan Çelik. 2025. Tokens with Meaning: A Hybrid Tokenization Approach for NLP. ArXiv: 2508.14292.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16495–16510, Suzhou, China. Association for Computational Linguistics.

Figen Beken Fikri, Kemal Oflazer, and Berrin Yanikoglu. 2021. Semantic Similarity Based Evaluation for Abstractive News Summarization. In *Proceedings of the First Workshop on Natural Language*

Generation, Evaluation, and Metrics (GEM), pages 24–33, Online. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. ArXiv: 2210.07316.

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.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of Tokenization on Language Models: An Analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and lightweight text representations. ArXiv: 2509.20354.