CLIB 2026 · COMPUTATIONAL LINGUISTICS IN BULGARIA · SOFIA

POSTER

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

M. Ali Bayram[1] **Öner Aytaş**[2] **Tuğçe Şen**[2]

[1] Yıldız Technical University [2] Işık University malibayram20@gmail.com

+4.81
Pearson on **STSbTR** over the best BPE baseline (+4.31 Spearman)

39.57
best **TR-MTEB** average over 26 tasks (next best 35.92)

61.2
best **TurBLiMP** pair accuracy over 16 phenomena (next best 52.7)

2.91
tokens / word — the cost, vs 1.82 for the most compact BPE baseline

1 Tokenization is a variable, not a default

Turkish is agglutinative: productive suffixation, phonological alternation and long derived forms mean that purely frequency-driven subword merges can cut straight across morpheme boundaries.

Yet in sentence embedding pipelines the tokenizer is almost always frozen as a preprocessing detail — even though representation quality depends on how **consistently semantically related forms are segmented** before the encoder is ever trained.

> **If architecture, data and initialisation are held fixed, how much does the tokenizer *alone* change Turkish sentence embedding quality?**

Morpheme aligned	kitap	lar	ımız	dan
Frequency driven	kita	plar	ımızd	an

kitaplarımızdan ‘from our books’ = *kitap* (book) + *lar* (pl) + *ımız* (1pl.poss) + *dan* (abl)

Schematic contrast between morpheme-aligned and frequency-driven segmentation; the lower row illustrates the failure mode, it is not the output of a specific tokenizer.

2 Controlled setup: only the tokenizer moves

HELD FIXED

- EmbeddingGemma-300M encoder architecture
- ~300K-example Turkish subset of the Cosmos Turkish Corpus
- One seed-42 initialisation, sampled once and copied to every run
- Teacher embeddings precomputed before student training
- Shared 2048-token context limit

VARIED

- The tokenizer stream — and nothing else

`mft` `mursit` `cosmos` `tabi`

- `input_ids` prepared **offline** per tokenizer, so runtime tokenization differences cannot leak in
- Any example over 2048 tokens under **any** tokenizer is dropped from **all** streams

▼ four encoders, identical in every respect except segmentation ▼

STSbTR semantic similarity · TR-MTEB 26 tasks · TurBLiMP minimal-pair diagnostic

Training data is broad web / pretraining-style Turkish, not task-specific STS, retrieval or classification data — so all three benchmarks are out-of-domain with respect to the training source.

3 The tokenizer under test: MFT

MFT combines **root–suffix analysis**, **phonological normalisation**, and a **BPE fallback** for the cases where morphological decomposition is unreliable. Its aim is to reduce semantically opaque splits rather than to maximise compression.

MEASURE	MFT	BEST BPE
Turkish ratio TR%	**90.29**	**53.48**
Purity ratio Pure%	**85.80**	**31.45**
`decode(encode(w))`	99.48	—
Tokens / word	2.91	1.82

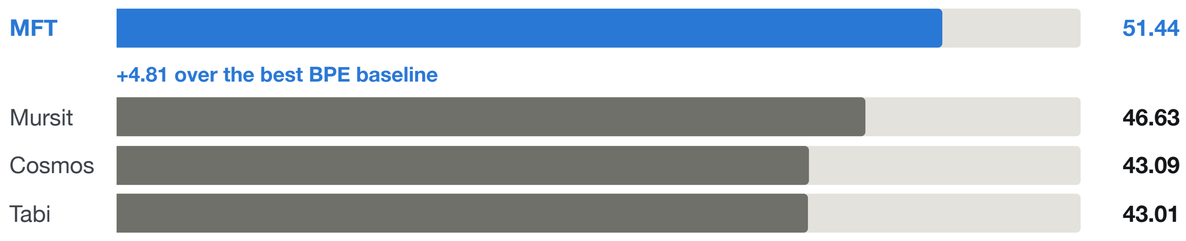
Tokenizer-side properties on TR-MMLU text with a 32,768-token vocabulary. The BPE column shows the strongest reported baseline value where relevant. TR% = share of language-appropriate tokens; Pure% = share that stay morphologically clean.

Lossless round-tripping holds for 99.48% of a 66,547-word sample. These are **internal** properties — the point of the controlled runs is to test whether they transfer downstream.

4 Results: MFT wins on all three levels

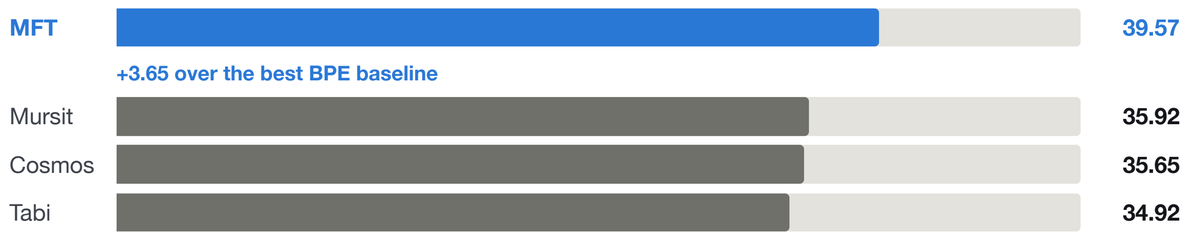
STSbTR — semantic textual similarity (test)

Pearson correlation ×100



TR-MTEB — 26-task Turkish embedding benchmark

Average score across all tasks



TurBLiMP — minimal-pair diagnostic

Mean pair accuracy over 16 phenomena, embedding centroid comparison

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

All four controlled runs. P and S are Pearson and Spearman on the STSbTR test split; the ranking is the same on the training split. Bars above are drawn from a zero baseline.

5 Reading the gains honestly

Where the advantage is largest. STS and retrieval-oriented settings drive most of the TR-MTEB gain.

Where it is not. Cosmos is slightly better on clustering, and the TurBLiMP phenomenon-level pattern is mixed — only the average favours MFT. Morphology-aware tokenization is a design choice with a clear overall benefit profile, **not** a universal win on every task.

> **The cost is real.** MFT spends 2.91 tokens per word against 1.82 for the most compact baseline — roughly **1.6× longer sequences** for the same text. The gains are not free; they are bought with sequence length.

6 Takeaways

1. For Turkish sentence embeddings, the tokenizer **still changes the learned space** when architecture, data and initialisation are all fixed.
2. Linguistically aligned segmentation should be treated as a **measurable design variable**, not a frozen preprocessing layer.
3. The contribution is methodological as much as empirical: tokenization can be **isolated, measured and compared** as a first-class factor in morphologically rich, lower-resource language technology.
4. **Limitation.** All controlled runs share a single seed; repeating the comparison across seeds is the immediate next step, alongside testing whether the pattern transfers to other morphologically rich and multilingual settings.

Code, models & contact

Contact
`malibayram20@gmail.com`

MFT tokenizer
`github.com/malibayram/turkish-tokenizer`

Models & evaluation artifacts
`huggingface.co/magibu`

tokenizer models

SELECTED REFERENCES

Bayram et al. (2025). Tokens with Meaning: A Hybrid Tokenization Approach for NLP. arXiv:2508.14292.

Toraman et al. (2023). Impact of Tokenization on Language Models: An Analysis for Turkish. TALLIP 22(4).

Schmidt et al. (2024). Tokenization Is More Than Compression. EMNLP 2024.

Beken Fikri et al. (2021). Semantic Similarity Based Evaluation for Abstractive News Summarization. GEM 2021.

Muennighoff et al. (2022). MTEB: Massive Text Embedding Benchmark. arXiv:2210.07316.

Baysan & Güngör (2025). TR-MTEB: A Comprehensive Benchmark and Embedding Model Suite for Turkish. Findings of EMNLP 2025.

Başar et al. (2025). TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs.

Schechter Vera et al. (2025). EmbeddingGemma: Powerful and Lightweight Text Representations. arXiv:2509.20354.