PLTK: Morphology-Aware BPE Tokenization for Polish Language Models

Rafał Adamczyk and Tianxiang Lu and Maja Popovic
IU International University of Applied Sciences
Erfurt, Germany
rafal.adamczyk@iu-study.org
{tianxiang.lu,maja.popovic}@iu.org

Abstract

Subword tokenization is a critical but often overlooked component in the development of language models for morphologically rich languages. Standard byte-pair encoding (BPE) tokenizers are frequency-driven and do not account for the inflectional structure of languages such as Polish. We introduce PLTK (PoLish morphological ToKenizer), a morphology-aware BPE tokenizer for Polish that enforces morpheme boundary constraints during the merge procedure using part-of-speech-specific trie structures. We train two GPT-2 decoder-only language models from scratch on identical Polish corpora: one tokenized with PLTK, one with a standard SentencePiece BPE baseline, and evaluate both on the KLEJ benchmark. The PLTK model achieves a higher average KLEJ score (0.823 vs. 0.813) and a lower final pre-training loss, suggesting that morphology-aware tokenization provide measurable benefits for decoder-only language models in the Polish setting.

Keywords: Polish NLP, BPE tokenization, morphology-aware tokenization, decoder-only language models, morphologically rich languages

1 Introduction

Subword tokenization is a foundational component of modern language models. Methods such as byte-pair encoding (BPE), often applied via implementations like SentencePiece, are widely used because they provide a simple and efficient compromise between character- and word-level modeling. However, these approaches are primarily frequency-driven and largely agnostic to linguistic structure. For morphologically rich languages, this can lead to segmentations that split stems inconsistently or merge inflectional affixes across morpheme boundaries. For example, the Polish word *napisałem* ("I wrote") could be split by a standard BPE tokenizer as *nap* | *isał* | *em*, fusing part of the prefix *na-* into the following token, whereas a morphology-aware split such as *na* | *pisał* | *em* preserves the prefix boundary.

This limitation is particularly relevant for Polish, a highly inflected Slavic language with rich nominal and verbal morphology. A single lemma may appear in many surface forms depending on case, number, gender, aspect, or person. Standard BPE training may therefore fragment morphologically related forms in ways that reduce token sharing and lexical consistency.

Recent work has explored morphology-aware tokenization for morphologically rich languages, including Slovak and multilingual settings (Držík and Forgáč, 2024; Asgari et al., 2025). However, Polish has received comparatively little attention in this context, particularly for decoder-only language models. Existing Polish decoder-only models typically inherit tokenizers designed for English-centric training pipelines.

In this work, we aim to answer the following question: *"Does blocking BPE merges across morpheme boundaries during tokenizer training improve model performance for Polish?"*

We introduce PLTK (PoLish morphological ToKenizer), a morphology-aware BPE tokenizer for Polish that constrains merge operations across morpheme boundaries using part-of-speech (POS)-aware affix tries. Unlike approaches that require full morphological analyzers during tokenization, PLTK modifies the BPE training procedure itself by preventing merges that cross selected protected affix boundaries.

To evaluate the impact of morphology-aware tokenization, we train two GPT-2-style decoder-only language models from scratch on the same Polish corpus and with the same architecture and optimization settings. The only experimental difference is the tokenizer: PLTK versus a standard SentencePiece BPE (SPM) baseline. We evaluate both models on the KLEJ benchmark as well as on intrinsic

tokenization metrics.

Our experiments show that PLTK produces segmentations with substantially higher morphological consistency and achieves lower pre-training loss together with modest but consistent downstream improvements on the evaluated KLEJ tasks. These findings suggest that preserving morpheme boundaries during subword vocabulary construction can benefit decoder-only language models for morphologically rich languages such as Polish.

Our contributions are:

- We introduce PLTK, a morphology-aware BPE tokenizer for Polish based on POS-aware morpheme boundary constraints.
- We provide a controlled tokenizer ablation study for Polish decoder-only language models trained from scratch.
- We show that morphology-aware tokenization improves intrinsic morphology metrics, is associated with lower pre-training loss, and yields consistent directional gains on downstream Polish NLP tasks.

2 Related Work

Subword tokenization. Byte-pair encoding (BPE), introduced as a data compression algorithm by Gage (1994) and adapted to open-vocabulary neural machine translation by Sennrich et al. (2016), iteratively merges the most frequent character pair in a corpus until a target vocabulary size is reached. Being purely frequency-driven, it does not encode linguistic structure explicitly. Frequent merges often align with morpheme boundaries, but this is a side effect of the data rather than a constraint of the algorithm: boundaries are crossed whenever the merge condition is met, producing segmentations that may split stems or mix affixes with unrelated tokens.

While morphology-aware tokenization can improve linguistic consistency, purely statistical tokenization methods optimize important practical objectives including compression efficiency, shorter token sequences, vocabulary agnosticism, and simple deployment. These advantages are particularly important for large-scale multilingual systems where language-specific resources are unavailable. PLTK therefore does not aim to replace statistical tokenization as a general-purpose solution, but investigates whether additional linguistic structure provides measurable benefits for a morphologically rich language such as Polish.

Several works have attempted to address that. Vilar and Federico (2021) propose a statistically motivated variant that defines a probability distribution over subword units and derives a data-driven stopping criterion, producing more morphologically coherent segmentations on rich languages without explicit linguistic resources.

Polish NLP. Polish has benefited from a growing ecosystem of language resources and benchmarks. Rybak et al. (2020) introduced the KLEJ benchmark for Polish language understanding. Mroczkowski et al. (2021) released HerBERT, a BERT-based encoder model (0.903 avg KLEJ). Dadas et al. (2020) released Polish RoBERTa (0.895 avg). However, these widely used Polish pretrained models are encoder-only and rely on standard subword tokenization.

Existing Polish decoder-only models, including Bielik (Ociepa et al., 2025c,a)and PLLuM (Kocoń et al., 2025), typically rely on general-purpose subword tokenization strategies rather than explicit morphology-aware vocabulary construction. Our work examines whether incorporating morphological constraints during vocabulary construction can benefit decoder-only language models for Polish.

Beyond pretrained models, Polish NLP provides mature linguistic resources. Formal grammar resources developed within the ParGram project (Butt et al., 2002), including the Polish POLFIE grammar (Patejuk and Przepiórkowski, 2017), provide symbolic descriptions of Polish syntax and grammatical structure. Morphological analyzers such as Morfeusz 2 provide detailed inflectional analyses of Polish word forms (Kieraś and Woliński, 2017). PLTK instead adopts lightweight POS-aware affix constraints to incorporate selected morphological information during BPE training.

Morphology-aware tokenization. Držík and Forgáč (2024) introduced SKMT, a morphology-aware BPE tokenizer for Slovak, achieving a 3.5% F1 improvement over standard BPE on downstream classification. Asgari et al. (2025) proposed MorphBPE, which integrates morphological dictionaries into BPE training to preserve morpheme boundaries. Our work applies morphological constraints for Polish, yielding a 1.2% relative improvement in average KLEJ score as well as lower training and validation loss values in pre-training phases.

Multilingual tokenization. Recent large-scale language models predominantly rely on multilingual pretraining and shared subword vocabularies, including models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021). These approaches optimize cross-lingual transfer by sharing vocabulary capacity across many languages, often using data-driven tokenization strategies such as SentencePiece. However, shared multilingual vocabularies may allocate fewer tokenization resources to individual languages, particularly morphologically rich low-resource languages. PLTK focuses instead on the monolingual setting, investigating whether explicitly preserving morphological structure during vocabulary construction benefits Polish decoder-only language models. Extending morphology-aware tokenization to multilingual models remains an open research direction.

3 Methodology

3.1 Architecture and Workflow

The experiment is structured in two parts as explained in (Figure 1). The upper branch of the workflow evaluates tokenization quality in isolation, while the lower branch examines whether the morphological focus of the proposed tokenizer translates into measurable gains during pre-training and fine-tuning. In both branches, PLTK and SPM start from the same preprocessed corpus; the only difference between the two experimental conditions is the tokenizer used for vocabulary construction and corpus encoding.

During the second phase of the experiment, two language models are trained on datasets encoded with different tokenizers and later fine-tuned on downstream tasks. The baseline tokenizer is standard SentencePiece BPE (Kudo and Richardson, 2018), referred to in this research as SPM. Fine-tuning is performed with LoRA (Hu et al., 2022), a parameter-efficient adaptation method that freezes the pre-trained weights and trains low-rank update matrices.

The central hypothesis of this study is not that PLTK improves tokenization quality alone, but that improved morphological consistency in the vocabulary construction stage can influence language model learning. Therefore, the evaluation follows a controlled tokenizer ablation: all model components, training data, and optimization parameters are fixed, while only the tokenizer differs.

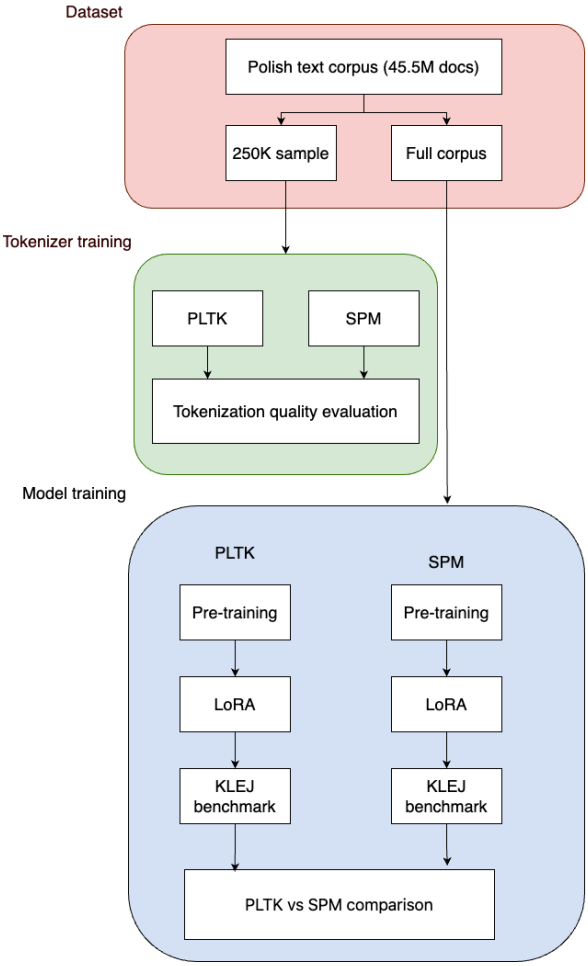

Figure 1: High-level overview of the experiment setup. A 250k-document sample of the Polish corpus is used to train both tokenizers (PLTK and the SPM baseline), which are first compared in isolation on intrinsic tokenization metrics (green block). The full 45.5M-document corpus is then encoded separately with each tokenizer, and two GPT-2-style decoder-only models with identical architecture and hyper-parameters are pre-trained from scratch, one per tokenizer (blue block). Both models are subsequently fine-tuned with LoRA on each KLEJ task and evaluated. The tokenizer is the only variable that differs between the two branches.

3.2 Corpus and Vocabulary

The training corpus combines three Polish text sources: Polish Wikipedia (`chrisociepa/wikipedia-pl-20230401`), OPUS-OpenSubtitles v2024, and OPUS-ParaCrawl v5. The preprocessing steps were selected to reduce noise specific to the source domains while preserving linguistic variation. Subtitle formatting markers such as leading dashes were removed because they represent dialogue structure rather than linguistic content. Wikipedia references were removed because they introduce

non-natural token sequences. No stemming, lemmatization, or morphological normalization was applied, since the objective is to evaluate whether the tokenizer itself can exploit naturally occurring morphology. Documents shorter than 50 characters are discarded. The same preprocessed corpus serves both purposes in this study: a sample of it trains the tokenizers, and the full set pre-trains the language models. We therefore filter once, keeping only documents long enough to be useful for both. The 50-character threshold was chosen heuristically and not tuned. The final corpus contains 45.5M documents.

Tokenizer training uses a 250k document sample (60% OpenSubtitles, 25% ParaCrawl, 15% Wikipedia) to cover conversational, web, and encyclopedic text. LM pre-training uses the full corpus.

Both tokenizers encode the same corpus. PLTK produces 1.55B tokens, SPM produces 1.39B tokens. The ratio (1.12) matches their fertility ratio (1.608 / 1.432), confirming the difference comes from segmentation, not corpus size. PLTK produces more tokens per word because protected morpheme boundaries prevent certain merges that would otherwise reduce token sequence length. This increased fertility is an expected consequence of the design rather than an optimization target.

3.3 PLTK Tokenizer

PLTK is a morphology-aware BPE tokenizer trained on the Polish corpus. Its core modification to standard BPE is the introduction of morpheme boundary constraints: before training, each word is analyzed with spaCy's `pl_core_news_sm` model to determine its part of speech (POS). Although full morphological analyzers such as Morfeusz 2 provide richer inflectional analyses, PLTK intentionally uses lightweight POS-aware boundary constraints instead of complete morphological parsing. POS-specific trie structures, populated with manually curated Polish inflectional affixes, detect prefix and suffix boundaries for verbs, nouns, adjectives, and adverbs. Each trie stores prefix patterns via forward traversal and suffix patterns via reversed traversal, returning the longest match. A boundary is marked as protected if (1) the detected affix belongs to the POS-appropriate set, (2) the remaining stem is at least two characters, and (3) the stem contains a Polish vowel; protected boundaries are excluded from BPE merge candidates. Word-initial space is represented with the SentencePiece (Kudo and Richardson, 2018) marker. A protected stop-word list is excluded from boundary analysis. The SPM baseline is trained on the same corpus using standard SentencePiece BPE with no morphological constraints. The detailed procedure is specified in Algorithm 1.

Algorithm 1 PLTK Training

Require: Corpus $\mathcal{C}$, target vocabulary size V
Ensure: Vocabulary $\mathcal{V}$, merge list $\mathcal{M}$

```
Count word frequencies in C
POS-tag all unique words with spaCy
Initialize V with special tokens and characters
for each word w do
    Detect morpheme boundaries using POS-specific tries
end for
while |V| < V do
    Count adjacent pairs, skip pairs crossing boundaries
    Select most frequent pair (t1, t2)
    Merge (t1, t2) into t1t2 in all words
    Add t1t2 to V
    Add (t1, t2) to M
end while
return V, M
```

Existing Polish morphological resources, such as Morfeusz 2, provide comprehensive analyses of Polish inflectional forms (Kieraś and Woliński, 2017). However, PLTK does not aim to perform full morphological analysis or replace established linguistic resources. Instead, it requires only boundary information relevant for BPE merge decisions. Therefore, PLTK adopts a lightweight POS-aware affix representation rather than integrating a complete morphological analyzer, which keeps the tokenizer training procedure focused on vocabulary construction while still incorporating selected morphological constraints.

3.4 Model Hyperparameters

Both models share an identical decoder-only transformer architecture (Vaswani et al., 2017): 12 layers, 12 attention heads, hidden dimension 768, feed-forward dimension 3072, dropout 0.1, and context length 512. Vocabulary size is 30,005 tokens for both. Pre-training uses the AdamW optimizer with learning rate 10^{-4}, batch size 64, for 4 epochs. Fine-tuning on KLEJ is performed using LoRA (Hu et al., 2022) with rank $r = 8$, $\alpha = 16$, learning rate 10^{-4}, batch size 16, for 5 epochs.

3.5 Evaluation Metrics

We evaluate tokenization quality using four intrinsic metrics. Fertility rate (FR) measures the average number of tokens produced per word and reflects segmentation granularity (Rust et al., 2021). STRR (Subword Tokenization Regularity Ratio) measures consistency of token segmentation across morphological variants (Nayeem et al., 2025). MC (Morphological Consistency) quantifies the extent to which morphologically related word forms share subword structure (Arnett and Bergen, 2025). MC-F1 is the harmonic mean-based variant of MC used in recent morphology-aware tokenization studies (Asgari et al., 2025).

We evaluate downstream performance using the KLEJ benchmark (Rybak et al., 2020): named entity recognition (NKJP-NER; F1), semantic textual similarity (CDSC-R; Spearman ρ), textual entailment (CDSC-E; accuracy), sentiment analysis in-domain (PolEmo-in; accuracy), sentiment analysis out-of-domain (PolEmo-out; accuracy), and review rating prediction (AR; $1-\text{wMAE}$). The overall KLEJ score is computed as the macro-average across all six tasks. We use the following abbreviations in Table 2 for compact presentation: CDSC-E (C-E), CDSC-R (C-R), PolEmo-in (IN), and PolEmo-out (OUT).

4 Results and discussion

4.1 Tokenization Quality

We compare PLTK against three reference tokenizers: the SPM baseline and two externally trained tokenizers, Bielik (APT4)(Ociepa et al., 2025b) and HerBERT. The comparison is intended to characterize existing Polish tokenization behavior rather than provide a controlled vocabulary-size comparison. We evaluate all four on fertility rate (Rust et al., 2021), STRR (Nayeem et al., 2025), morphological consistency (Arnett and Bergen, 2025), and MC-F1 (Asgari et al., 2025) as presented in Table 1. PLTK exhibits a higher fertility rate than the baselines as a consequence of morpheme boundary preservation, which produces 12% more tokens per word. This increase reflects a deliberate trade-off: PLTK sacrifices compression efficiency to preserve morphological structure.

Peak GPU memory allocation remained similar under the fixed training configuration used in our experiments. PLTK scores higher on MC and MC-F1, indicating that the tokenizer preserves stem-sharing structure across inflected forms.

Tokenizer	FR	STRR	MC	MC-F1
PLTK (ours)	1.608	54.2	**54.9**	**50.7**
SPM Baseline	1.432	69.6	33.6	16.1
Bielik (APT4)	1.438	69.4	29.9	13.3
HerBERT	**1.387**	**72.8**	23.3	7.7

Table 1: Tokenization quality metrics. PLTK and SPM use approximately 30k vocabulary size; Bielik and HerBERT use their original vocabulary sizes (32k and 50k, respectively). STRR, MC, and MC-F1 in percent.

4.2 Pre-training results

Training loss is tracked across all steps for both models across three independent runs. All three PLTK runs converge to a lower final training loss (approximately 3.1) than all three SPM runs (approximately 3.5 - 3.6), within the same number of optimization steps (Figure 2). However, PLTK processes more tokens due to its higher fertility. The SPM runs follow different initializations across seeds but converge to a consistent range, above the PLTK runs throughout training. The gap between the two groups is stable from early training onward and is not an artefact of the different starting values, which reflect the larger token count per sequence under PLTK's higher fertility rate. Peak GPU memory allocation on a single A100 (80 GB) was approximately 48 GB for both models, indicating that the choice of tokenizer has no significant impact on memory footprint under identical training configuration.

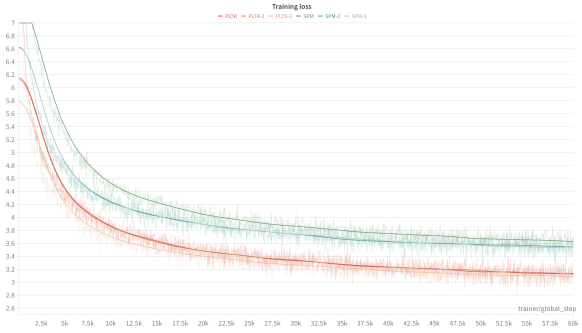

Figure 2: Training loss for PLTK (red) and SPM (green) models over 60k steps across three independent runs each. Solid lines are smoothed; shaded areas show per-step variance. PLTK converges consistently lower (≈3.1) than SPM (≈3.5–3.6) across all runs.

4.3 Fine-tuning and Downstream Evaluation

Both models are fine-tuned on KLEJ (Rybak et al., 2020). We evaluate on six tasks: named entity recognition (NKJP-NER; F1), semantic textual sim-

ilarity (CDSC-R; Spearman ρ), textual entailment (CDSC-E; accuracy), sentiment analysis in-domain (PolEmo-in; accuracy), sentiment analysis out-of-domain (PolEmo-out; accuracy), and review rating prediction (AR; $1-$wMAE). The overall KLEJ score is the average across all six task scores. The PLTK model scores higher than the SPM baseline on all evaluated tasks (Table 2), though the per-task margins are small and, on NER and AR in particular, fall within the range that seed variation could plausibly account for; we report only a single fine-tuning run per task and do not establish statistical significance, as discussed in the Limitations section.

For reference, base-scale encoder models on the full KLEJ leaderboard achieve higher scores (HerBERT: 0.903, Polish RoBERTa: 0.895 (Mroczkowski et al., 2021; Dadas et al., 2020)), but differ in architecture (encoder-only, MLM objective) and are not directly comparable.

Model	Semantic			Sentiment		AR	Avg
	NER	C-E	C-R	IN	OUT		
PLTK	0.842	0.922	0.917	0.833	0.605	0.820	**0.823**
SPM	0.839	0.906	0.902	0.817	0.595	0.817	0.813

Table 2: KLEJ results for decoder-only GPT-2 models (tokenizer ablation).

The largest gains are observed on CDSC-E, CDSC-R, and PolEmo-in, while smaller improvements appear on PolEmo-out, AR, and NER. Although the precise source of these gains requires further investigation, the consistent direction of improvement suggests that morphology-aware segmentation may improve lexical generalization for Polish decoder-only models.

These findings are consistent with results reported for Slovak, where SKMT yielded a 3.5% F1 improvement over standard BPE (Držík and Forgáč, 2024). To our knowledge, this is the first systematic evaluation of morphology-aware BPE tokenization to Polish in a decoder-only setting.

4.4 Generation Quality

As a qualitative analysis, we additionally inspect completions on prompts ending mid-stem or requiring agreement-sensitive continuation (Table 3).

These examples were selected by the authors to illustrate characteristic differences and are intended as a qualitative illustration. In these manually selected examples, PLTK produces continuations that remain morphologically consistent with the prompt context, including inflected noun and adjective forms as well as subject–verb agreement. In contrast, the SPM baseline produces in several examples less morphologically consistent continuations, agreement mismatches, or repetitive generations. In the mid-stem prompts, PLTK produces morphologically well-formed continuations, including a genitive-feminine adjective (*odległej*), an accusative noun (*książkę*), and a locative plural (*wzgórzach*). In contrast, the SPM baseline produces continuations with incorrect or inconsistent inflectional forms in each case.

On the agreement-sensitive prompt, PLTK generates the correct third-person plural form (*czytają*), while SPM produces a singular form (*czyta*). On the prefixed-verb prompt, PLTK produces a diverse continuation (*cokolwiek pomoże*), whereas SPM collapses to repeated generation of the same token (*narobić*).

4.5 Efficiency and Generality Cost

Morpheme boundary preservation is not free. Because PLTK blocks selected merges across protected boundaries, it produces more, shorter tokens than a standard BPE tokenizer: fertility rises from 1.432 to 1.608, a 12% increase in tokens per word (fertility ratio 1.12). Encoding the full corpus generates 1.55B tokens with PLTK against 1.39B with SPM. Compression is the primary objective of standard BPE, and PLTK deliberately relaxes it.

This has a direct downstream cost. Under a fixed context length and batch size, a 12% increase in tokens per word implies a proportional increase in average sequence length and a corresponding reduction in training and inference throughput. We did not measure throughput or sequence length directly, so we use the token-count ratio as a proxy for this slowdown. Memory footprint, by contrast, was approximately unchanged under our experimental configuration: peak GPU allocation on a single A100 (80 GB) was approximately 48 GB for both models, since the vocabulary size and model architecture are identical.

The boundary detection step adds a one-time preprocessing cost: each unique word is POS-tagged and passed through the affix tries before tokenizer training begins. This cost is incurred once, at training time, and does not affect inference latency once the merge list is fixed.

Finally, PLTK trades BPE's language agnosticism for language-specific structure. Standard BPE requires only raw text, whereas PLTK depends on

a Polish POS tagger, manually curated inflectional affix lists, and a stopword list. The method is therefore Polish-specific by construction and does not transfer to a new language without equivalent resources. Whether this reduced generality is worth it depends on the gains it buys. For Polish, the improvements in morphological metrics, pre-training loss, and KLEJ scores suggest the trade is worthwhile.

Prompt	**PLTK**	**SPM**
Jestem przybyszem z odległ... (‘I am a visitor from a distant...’)	*odległej galaktyki* (‘distant galaxy’, genitive feminine)	*odległości* (‘distance’)
Napisałem książk... (‘I wrote a book...’)	*książkę o tym jak* (‘book, accusative, about how’)	*książk, żeby* (malformed stem, ‘so that’)
Dom stoi na wzgórz... (‘The house stands on a hill...’)	*wzgórzach* (‘hills’, locative plural)	*wzgórze* (‘hill’, nominative singular)
Ona czyta książkę. On czyta książkę. My czytamy książkę. Oni (‘She reads a book. He reads a book. We read a book. They’)	*czytają książkę* (‘read a book’, third person plural)	*czyta książkę* (‘reads a book’, third person singular)
robić, zrobić, przerobić, narobić, (‘to do, to do (perfective), to redo, to make a lot of,’)	*cokolwiek pomoże* (‘whatever helps’)	*narobić, narobić, narobić, narobić* (repetition)

Table 3: Representative generation examples comparing PLTK and SPM outputs, with English glosses. Examples were selected manually to illustrate characteristic differences and are not a random sample; they are intended as qualitative illustration only.

5 Limitations

PLTK combines multiple components (POS-aware boundary detection, trie-based affix matching, stopword exclusions); the present work does not ablate their individual contributions.

We do not report variance across fine-tuning seeds: each KLEJ result comes from a single fine-tuning run, so the observed score differences, several of which are small, may not be statistically significant. Establishing significance would require repeating each fine-tuning run across multiple seeds and reporting the resulting variance.

Both models operate on the GPT-2 (Radford et al., 2019) base scale and are not directly comparable to Polish production LLMs trained on significantly larger corpora.

Although multilingual pretrained models such as XLM-R and mT5 demonstrate the effectiveness of shared multilingual vocabularies, this study focuses on isolating the effect of morphology-aware tokenization in a controlled monolingual Polish decoder-only setting. A direct comparison would require controlling for differences in architecture, training objectives, and multilingual data composition. Investigating morphology-aware extensions for multilingual tokenizers is an important direction for future work.

Furthermore, the higher fertility rate of PLTK changes the token distribution and sequence length characteristics of the training data, which may also contribute to the observed differences independently of morphology preservation itself.

6 Conclusion

We presented PLTK, a morphology-aware BPE tokenizer for Polish that constrains merge operations across morpheme boundaries using POS-specific affix tries. In a controlled comparison against a standard SentencePiece BPE tokenizer, PLTK produced more morphologically consistent segmentations, achieved lower pre-training loss, and yielded higher scores across the evaluated KLEJ tasks for GPT-2-style decoder-only language models.

Although the observed gains are modest, the results suggest that incorporating morphological structure into subword vocabulary construction can benefit decoder-only language models for morphologically rich languages such as Polish. PLTK remains compatible with standard transformer training pipelines and does not require architectural modifications.

Future work should examine larger-scale training regimes, multilingual settings, and controlled ablation studies of individual tokenizer components.

Artifact Availability

The implementation, training scripts, and evaluation resources are publicly available online. [1]

[1] https://github.com/Prof-it/tokenizer_pol

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *COLING-02: Grammar Engineering and Evaluation*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Evaluation of sentence representations in Polish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1674–1680, Marseille, France. European Language Resources Association.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Dávid Držík and František Forgáč. 2024. Slovak morphological tokenizer using the byte-pair encoding algorithm. *PeerJ Computer Science*, 10:e2465.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Witold Kieraś and Marcin Woliński. 2017. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, 97(1):75–83.

Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, Julia Moska, Tomasz Naskręt, Bartosz Walkowiak, Mateusz Gniewkowski, Kamil Szyc, Dawid Motyka, Dawid Banach, Jonatan Dalasiński, Ewa Rudnicka, and 80 others. 2025. Pllum: A family of polish large language models. *Preprint*, arXiv:2511.03823.

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, Brussels, Belgium. Association for Computational Linguistics.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Mir Tafseer Nayeem, Sawsan Alqahtani, Md Tahmid Rahman Laskar, Tasnim Mohiuddin, and M Saiful Bari. 2025. Beyond fertility: Analyzing strr as a metric for multilingual tokenization evaluation. *Preprint*, arXiv:2510.09947.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025a. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025b. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2025c. Bielik 11b v2 technical report. *Preprint*, arXiv:2505.02410.

Agnieszka Patejuk and Adam Przepiórkowski. 2017. Polfie: wspólczesna gramatyka formalna języka polskiego. *Język Polski*, 97(1):47–63.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive

benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Vilar and Marcello Federico. 2021. A statistical extension of byte-pair encoding. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 263–275, Bangkok, Thailand (online). Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.