miLLi: Model Integrating Local Linguistic Insights for Morphologically Robust Tokenization

Elshad Rahimov
Independent Researcher
Baku, Azerbaijan
elshadrahimovm@gmail.com

December 31, 2025

Abstract

Standard statistical tokenization algorithms often struggle to preserve the morphological boundaries of agglutinative languages such as Azerbaijani. This study introduces miLLi 1.0, a hybrid tokenizer that integrates a rule-based root dictionary with a statistical Byte-Pair Encoding (BPE) approach. The model's distinguishing feature is a dynamic phonological restoration algorithm designed to map allomorphic variations back to their canonical root forms. Empirical evaluations on the Tatoeba corpus demonstrate that miLLi 1.0 (1.955 T/W) outperforms global standards such as GPT-4o and mBERT in terms of representation efficiency. While exhibiting lower token density compared to local statistical models, miLLi 1.0 demonstrates superior linguistic robustness, achieving 53.0% in Morphological Boundary Accuracy (MBA) and 78.0% in Root Consistency Rate (RCR). The findings suggest that the integration of a linguistic filtration layer establishes an optimal balance between statistical compression and semantic integrity.

Keywords: Agglutinative languages, Tokenization, Morphological segmentation, Phonological restoration, Azerbaijani language, NLP.

1 Introduction

The central position of Large Language Models (LLMs) in the modern Natural Language Processing (NLP) ecosystem has highlighted the challenge of efficiently representing languages with diverse morphological structures. Specifically, in agglutinative languages such as Azerbaijani, the extensive capacity for word formation and inflection complicates the machine's ability to perceive unified semantic units [1]. The traditional approach to addressing this complexity is statistical tokenization, which segments text into subword units. However, existing research demonstrates that statistical subword algorithms often lack sensitivity to

morphological boundaries, which negatively impacts the learning performance and generalization capabilities of language models [2].

The fundamental problem encountered by statistical tokenization in agglutinative languages is twofold. On one hand, models with small vocabulary sizes fragment words into small, semantically unrelated units, creating inefficient token sequences. On the other hand, models trained on large datasets include inflected forms as independent tokens to maintain compact representation, leading to the problem of "vocabulary sparsity" [3]. This situation hinders the statistical model's ability to internalize the grammatical relationship between the root and the suffix, potentially resulting in computational and semantic losses.

In the context of the Azerbaijani language, these challenges are further exacerbated by morphonological processes resulting from phonetic evolution. Processes such as consonant mutation upon the addition of vowel-starting suffixes or vowel loss (haplology) distance the visual representation of words from their base form (lemma) [4]. Purely frequency-based tokenizers, such as mBERT [5] or GPT [6], treat these phonetic variants as distinct concepts. This can cause the model to store semantic units in unrelated parts of the vocabulary, leading to memorization rather than true linguistic understanding.

This research proposes **miLLi 1.0**, a hybrid tokenizer for the Azerbaijani language that combines statistical flexibility with linguistic precision. The primary objective is not to surrender the text entirely to statistical algorithms but to preserve the integrity of word roots through a phonological restoration algorithm applied during the pre-tokenization stage. Such a hybrid approach strives to maintain an optimal vocabulary size while establishing a mathematically more robust connection between tokens and morphemes.

The remainder of this paper is organized as follows: Section 2 provides a literature review of existing tokenization approaches for agglutinative languages; Section 3 details the hybrid architecture of the miLLi 1.0 model and the methodological specifics of the phonological restoration algorithm; Section 4 presents a comparative analysis of experimental results across quantitative and qualitative indicators; finally, the limitations of the proposed approach and future research perspectives are discussed.

2 Literature Review

Initial studies by Huseynov et al. [1] indicated that numerous syntactic and semantic variants arising from the same root in the Azerbaijani language cause sparsity issues in traditional vector models such as Word2Vec and GloVe. To address this, Ziyaden et al. [4] introduced a large-scale transformer model for Azerbaijani based on the RoBERTa architecture, attempting to mitigate the negative impact of Out-of-Vocabulary (OOV) words through data augmentation and Zemberek normalization. However, standard subword tokenizers used in these models, such as BPE proposed by Sennrich et al. [7] and WordPiece developed by Schuster and Nakajima [8], often fail to consider morphological boundaries.

In the field of morphological segmentation, classical approaches are generally divided into statistical and rule-based methods. Morfessor 2.0 [9], based on the Minimum Description Length (MDL) principle, identifies morphemes within a corpus using statistical probabilities without prior knowledge of grammatical rules. Conversely, the Zemberek project [10] offers a framework for Turkic languages rooted entirely in linguistic rules and rich root-suffix

dictionaries. while Zemberek's sequential suffix logic ensures morphological accuracy, it can lose flexibility when identifying new words not present in the dictionary or dealing with misspellings. The methodology of **miLLi 1.0** aims to preserve the vocabulary stability offered by Zemberek while combining it with a more modern and probabilistic fallback mechanism.

In recent years, the impact of tokenizer granularity on model performance for agglutinative languages was extensively investigated in the ITUTurkBERT project [3]. This study demonstrates that increasing vocabulary size can reduce morphological fragmentation, thereby enhancing semantic perception. However, excessive vocabulary expansion introduces another risk: the memorization of inflected words as single tokens. Addressing this contradiction, the "aLLMA" project [11] successfully reduced token waste by training a statistical WordPiece tokenizer on the massive DOLLMA corpus. Nevertheless, purely statistical approaches, especially in low-resource environments, may treat allomorphs derived from the same root (e.g., *bayraq* and *bayrağı*) as distinct semantic units, negatively affecting morphological generalization capabilities.

Taking this problem a step further, the "Tokens with Meaning" project [12] proposed unifying allomorphs under a single identifier by applying fixed vocabulary and phonological normalization. The proposed miLLi 1.0 model adopts this approach but distinguishes itself by incorporating a "Longest Restored Match" algorithm that dynamically restores vowel loss and consonant mutations frequently encountered in Azerbaijani. Among modern hybrid models, MorphBPE [13] attempts to keep subword units linguistically clean by introducing morphological boundary constraints to the statistical merging process. miLLi 1.0 enriches this cleanliness with a phonological restoration layer, aiming to prevent both token waste and semantic detachment.

On the other hand, projects like CANINE [14] and Bolmo [15] have demonstrated the advantages of operating directly on bytes, completely abandoning tokenizers. In particular, Bolmo's non-causal boundary predictor allows for more accurate determination of split points by looking at future context. However, the high computational costs and large training data requirements of byte-based models complicate their application in low-resource languages. The miLLi 1.0 project attempts to integrate this technical flexibility of Bolmo into a Unicode-native BPE architecture, presenting a computationally efficient hybrid system for Azerbaijani that possesses both readability and "infinite vocabulary" capabilities.

Projects such as KoBERT [16] and AraBERT [17], developed for languages with similar morphological structures, have also confirmed the effectiveness of dictionary-based pre-tokenizer systems. miLLi 1.0 seeks to serve as an optimal tokenizer by enriching these global experiences with the specific characteristics of the Azerbaijani language.

3 Methodology

The methodological framework of this research relies on the hybrid integration of a "Root-based Dictionary" and "Byte-level BPE" approaches. The proposed architecture encompasses three interconnected stages: processing of linguistic resources, a hybrid pre-tokenization algorithm, and subword training (Figure 1).

3.1 Preparation and Processing of Linguistic Resources

To ensure the semantic stability of the model, a root dictionary was utilized as the primary reference in the initial stage. For this purpose, the open-source *az.dic* dictionary (in Hunspell format) from the Mozilla project was selected as the base resource [18]. During the processing of the dictionary, in accordance with the "Pure Tokens" principle [12], priority was given to preserving lexicalized words as unified semantic units. Considering that verbs in Azerbaijani dictionaries are typically represented in their nominalized forms (verbal nouns), infinitive suffixes were automatically trimmed to allow the model to independently recognize pure verb roots. Over 36,000 cleaned unique root words were compiled into a Trie (prefix tree) structure using the Aho-Corasick algorithm to accelerate text search operations [19].

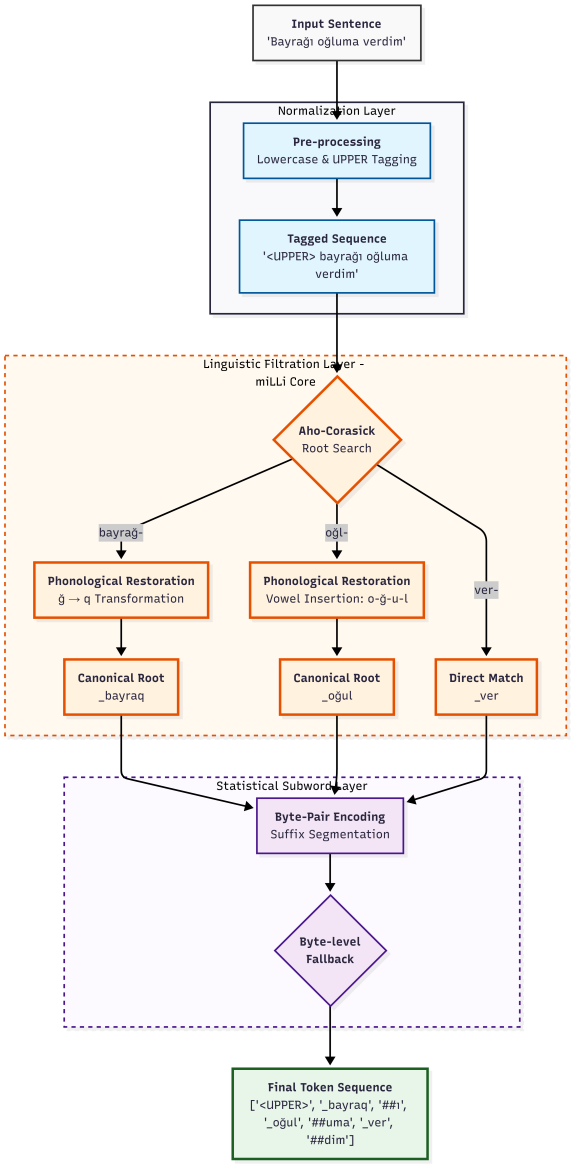

Figure 1: The hybrid architecture of miLLi 1.0 tokenizer.

3.2 Dataset

The statistical learning capability of the model was established on the Azerbaijani segment of the CulturaX corpus [20]. Since this corpus is a cleaned version of the mC4 and OSCAR datasets, no additional pre-processing measures were required. Taking into account computational resources and the model's calibration needs, a random subset of 500,000 lines was extracted from the corpus for this study. This base provides samples for the statistical model to acquire both high-frequency and rare lexemes of the Azerbaijani language.

3.3 Hybrid Pre-tokenization Strategy

The core of the proposed approach is a pre-tokenization mechanism that filters the text through a linguistic layer before subjecting it to statistical algorithms.

Handling Uppercase and the `<UPPER>` Indicator. As a first step, all uppercase letters in the text are replaced with a special `<UPPER>` indicator token, and all lexical units are converted to lowercase. This approach consolidates forms of the same word across different cases (e.g., sentence-initial vs. mid-sentence) under a single root, thereby reducing vocabulary sparsity and enhancing the model's semantic generalization capability. Simultaneously, the specific information regarding proper nouns (NER) carried by uppercase letters is preserved, and a standardized input is ensured for the phonological restoration algorithm. The slight increase in token count is accepted as a methodological trade-off in exchange for vocabulary compactness and semantic precision.

Phonological Restoration Algorithm. To manage morphonological processes characteristic of the Azerbaijani language that lead to semantic detachment at the token level (e.g., *q-ğ*, *k-y* mutations and vowel loss during suffixation), a dynamic Phonological Restoration algorithm was developed, considering principles applied by Ziyaden et al. [4]. The algorithm attempts to establish a link between the direct root form in the dictionary and its phonetically modified variants. Consequently, forms such as *bayrağ+ı* or *oğl+um* are mapped back to their initial roots (*bayraq* and *oğul*). Through the "Longest Restored Match" principle applied during segmentation, each word in the text is first separated into a dictionary-based root part, followed by statistically learned subword suffixes (e.g., *_bayraq + ##ı*). This significantly reduces the problem of identifying morphological boundaries in incorrect positions.

3.4 Subword Training and Byte-level Fallback

A Byte-Pair Encoding (BPE) model with a vocabulary size of 64,000 was trained on the linguistically segmented text. The model configuration is based on recommendations from studies conducted on existing agglutinative language models [3]. To ensure the model's robustness against rare words or foreign characters, a byte-level fallback mechanism, similar to those used in the Bolmo [15] and CANINE [14] projects, was implemented. Thanks to this mechanism, out-of-vocabulary elements are processed as UTF-8 bytes instead of being lost as `<UNK>` tokens. The final tokenizer architecture is encapsulated within a custom Python

class, analyzing the input text first with the linguistic dictionary structure, and then passing the result to the statistical BPE model to form the optimal token sequence.

4 Experiments and Results

To evaluate the performance of the proposed miLLi 1.0 tokenizer, a two-stage evaluation strategy was implemented. The primary objective of the study is to empirically analyze the balance between quantitative efficiency (token compression) and linguistic accuracy (morphological integrity).

4.1 Experimental Setup and Datasets

To ensure the objectivity of the research and minimize the risk of "data leakage," experiments were conducted on held-out datasets that were not used during the training of the models. For quantitative analysis, the Azerbaijani segment of the **Tatoeba** corpus, consisting of approximately 4,500 sentences covering daily conversational language and various lexical layers, was utilized. This corpus was selected to measure the models' robustness against real-world texts and Out-of-Vocabulary (OOV) words. For qualitative analysis, a curated test set comprising over 100 complex agglutinative words reflecting the specific nature of Azerbaijani—including phonetic phenomena such as vowel loss and consonant mutation—was compiled. The comparative analysis involved global industry standards (GPT-4o, GPT-3.5, mBERT, XLM-RoBERTa) and local statistical models (aLLMA, AzeBERT, CustomAz).

4.2 Quantitative Analysis: Token Efficiency

Tests conducted on the Tatoeba corpus revealed the efficiency of different tokenization approaches in representing text (Token/Word ratio - T/W), as shown in Table 1. The results indicate that the **miLLi 1.0** model (1.955 T/W) represents Azerbaijani text more compactly compared to global multilingual models. Specifically, the efficiency gap observed when compared to GPT-4o [6] (2.387) and mBERT [5] (2.521) demonstrates the importance of a language-specific approach. However, it should be noted that local models trained on large corpora using statistical frequency principles (e.g., aLLMA [11] – 1.418) outperform miLLi 1.0 in terms of compression rate. This difference is explained by the fact that statistical models often include frequently occurring complex word forms in the dictionary as whole units without fragmentation, whereas miLLi 1.0 prioritizes morphological segmentation.

4.3 Qualitative Analysis: Morphological and Phonological Accuracy

The model's ability to separate words at linguistically correct morphological boundaries (root and suffix) was measured using the **Morphological Boundary Accuracy (MBA)** metric (Table 2). Test results show that the miLLi 1.0 model accurately identified the root in 53.0% of the words. For comparison, XLM-RoBERTa [21] achieved 38.0%, aLLMA 16.0%, and GPT-4o only 4.0%. The low scores of statistical models (especially aLLMA and CustomAz)

Table 1: Comparison of Token/Word (T/W) ratios on the Tatoeba corpus.

Model	T/W Ratio	Category
aLLMA	1.418	Local (Statistical)
AzeBERT	1.571	Local (Statistical)
miLLi 1.0	**1.955**	**Local (Hybrid)**
GPT-4o	2.387	Global (SOTA)
mBERT	2.521	Global (Multilingual)

can be attributed not to "incorrect" processing, but rather to under-segmentation (keeping words whole) or splitting into statistically frequent syllables that lack linguistic meaning.

The effectiveness of the "Phonological Restoration" mechanism, the core methodological novelty of this research, was verified using the **Root Consistency Rate (RCR)** test. Within this test, the miLLi 1.0 model successfully restored 78.0% of words that underwent phonetic modification (e.g., *bayrağımız*, *oğlum*) to their original root forms (*bayraq*, *oğul*). Other models were expected to score in the 0-1% range, as traditional BPE and WordPiece algorithms rely on visual forms and lack deep linguistic restoration functions. The fact that miLLi 1.0's result is not 100% is limited by the coverage of the root dictionary used and certain exceptional cases.

Table 2: Results for MBA and RCR.

Model	MBA (%)	RCR (%)
miLLi 1.0	**53.0%**	**78.0%**
XLM-RoBERTa	38.0%	0.0%
aLLMA	16.0%	1.0%
GPT-4o	4.0%	0.0%

4.4 Generalization of Results

The conducted experiments clearly demonstrate the trade-off in the design philosophy of the miLLi 1.0 tokenizer. Although the model lags behind local competitors in terms of statistical compression, it exhibits a distinct profile regarding morphological transparency and the preservation of semantic roots. Analysis of **Vocabulary Sparsity Index (VSI)** supports this observation; miLLi 1.0 (94 unique tokens) occupies a balanced position between aLLMA (97) and GPT-3.5 [22] (83), neither inflating the vocabulary excessively nor fragmenting words into meaningless particles. These characteristics highlight the potential of the proposed approach, particularly for NLP tasks where morphological analysis and semantic precision are critical.

5 Limitations and Future Research Directions

While the proposed miLLi 1.0 tokenizer demonstrates a morphologically grounded approach for the Azerbaijani language, certain limitations exist at the current stage of research.

Firstly, the effectiveness of the model's hybrid architecture is directly dependent on the coverage of the root dictionary utilized (based on Mozilla's *az.dic* [18]). In the presence of neologisms, dialectisms, or specific terms not present in the dictionary, the phonological restoration mechanism does not trigger, and the system reverts to standard BPE segmentation. In such cases, the linguistic advantage of the proposed method is neutralized for those specific tokens.

Secondly, the phonological restoration algorithm is built upon the most widespread morphonological processes of the Azerbaijani language. However, the rule-based algorithm may prove insufficient for exceptional linguistic cases. Comprehensive coverage of such nuances requires a broader rule base or the integration of contextual analysis.

A third technical limitation relates to computational speed. Compared to statistical tokenizers optimized purely in C++, the Python-based pre-tokenization and Trie search operations employed by miLLi 1.0 introduce a certain degree of latency. Furthermore, the requirement for the `trust_remote_code=True` parameter for operation within the Hugging Face ecosystem may pose integration challenges in industrial environments with strict security protocols.

The priority direction for future research is the training of Small Language Models (SLMs) based on the miLLi 1.0 tokenizer and the measurement of their real-world performance in downstream tasks such as Text Classification, Named Entity Recognition (NER), and Question Answering (QA).

Code and Model Availability

The proposed tokenizer and linguistic resources are publicly hosted on Hugging Face: https://huggingface.co/elshadrahimov/miLLi-1.0.

References

[1] K. Huseynov, U. Suleymanov, S. Rustamov, and J. Huseynov. Training and evaluation of word embedding models for azerbaijani language. In *ADA University 4th International Conference on Computing and Information Technologies*. ADA University, 2020.

[2] Kaj Bostrom and Greg Durrett. 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, 2020.

[3] Y. B. Kaya and A. C. Tantuğ. Effect of tokenization granularity for turkish large language models. *Intelligent Systems with Applications*, 21:200335, 2024.

[4] A. Ziyaden, A. Yelenov, F. Hajiyev, S. Rustamov, and A. Pak. Text data augmentation and pre-trained language model for enhancing text classification of low-resource languages. *PeerJ Computer Science*, 10:e1974, 2024.

[5] Google Research. Bert multilingual cased model. https://huggingface.co/google-bert/bert-base-multilingual-cased, 2018. Accessed: 2025-12-27.

[6] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2025-12-27.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, 2016.

[8] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[9] Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Python implementation and extensions for morfessor baseline. Technical report, Aalto University, 2013.

[10] Ahmet A. Akın and Mehmet D. Akın. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10:1–5, 2007.

[11] J. Isbarov, K. Huseynova, E. Mammadov, M. Hajili, and D. Ataman. Open foundation models for azerbaijani language. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 18–28. Association for Computational Linguistics, 2024.

[12] M. A. Bayram et al. Tokens with meaning: A hybrid tokenization approach for nlp. *arXiv preprint arXiv:2508.14292*, 2025.

[13] E. Asgari, Y. El Kheir, and M. A. S. Javaheri. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *arXiv preprint arXiv:2502.00894*, 2025.

[14] Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022.

[15] B. Minixhofer et al. Bolmo: Byteifying the next generation of language models. *arXiv preprint arXiv:2512.15586*, 2025.

[16] SKTBrain. Kobert: Korean bert model with mecab-ko tokenizer. https://github.com/SKTBrain/KoBERT, 2019. GitHub Repository.

[17] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, 2020.

[18] Mozilla. Azerbaijani dictionary (az.dic). https://github.com/mozillaz/spellchecker, n.d. Accessed: 2025-12-27.

[19] Alfred V Aho and Margaret J Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[20] Thuat Nguyen et al. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*, 2023.

[21] 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*, 2020.

[22] OpenAI. Tiktoken: Fast bpe tokeniser for use with openai's models. https://github.com/openai/tiktoken, 2022. Accessed: 2025-12-27.