NLP Project: Analysis of Azerbaijani Classical Poetry Dataset

Kamal Aghazada Fatulla Bashirov

February 2026

Abstract

We analyzed an Azerbaijani classical poetry dataset containing 846 poems from 9 renowned poets. We collected data from Wikimedia sources, performed cleaning and modernization, and conducted NLP analyses including tokenization, Heaps' law, BPE encoding, sentence segmentation, and spell checking. Our work demonstrates linguistic characteristics of classical Azerbaijani poetry and text processing challenges.

1 Introduction and Data Collection

We created a dataset of classical Azerbaijani poetry to understand structural and linguistic properties. The dataset contains poems from major poets including Muh″mmad Fuzuli, Imadaddin Nasimi, Qasim bey Zakir, Xaqani Shirvani, Molla Panah Vagif, and others.

We divided work where Fatulla Bashirov handled data cleaning, translation, and tasks 1-3 (tokenization, Heaps' law, BPE), while I handled data gathering and tasks 4-5 (sentence segmentation, spell checking).

We used a Go-based system with Wikimedia API to collect 846 poems (909,630 characters original, 935,784 modernized). We cleaned metadata artifacts and generated modern translations using Gemini 2.5 Flash while preserving poetic structure.

2 Tasks Analysis

2.1 Tokenization and Vocabulary Analysis (Task 1)

We implemented a simple space-based tokenizer that normalizes whitespace and splits text on space characters. We analyzed vocabulary characteristics of both original and translated texts to understand the linguistic diversity.

Results:

- Original text: 132,975 tokens, 50,232 unique types (type-token ratio: 0.378)

- Modern translated text: 141,008 tokens, 41,998 unique types (type-token ratio: 0.298)

The lower type-token ratio in modern text indicates less lexical diversity per token, which suggests that the modernization process standardized vocabulary and reduced the variety of archaic word forms present in classical texts.

2.2 Heaps' Law Analysis (Task 2)

We estimated Heaps' law parameters ($V = k \cdot N^{\beta}$) by streaming tokens in order and recording vocabulary growth every 1,000 tokens. We performed log-log linear regression to fit the relationship between vocabulary size and cumulative token count.

Fitted parameters:

- $k = 2.306$ (vocabulary richness constant)
- $\beta = 0.847$ (vocabulary growth exponent)
- $R^2 = 0.9998$ (coefficient of determination)

The β value of 0.847 is typical for natural language texts and indicates sub-linear vocabulary growth, meaning new word types are encountered less frequently as corpus size increases. The exceptionally high R^2 value demonstrates that our classical Azerbaijani corpus follows the expected statistical scaling behavior. Figure 1 shows the excellent fit between observed and predicted vocabulary growth.

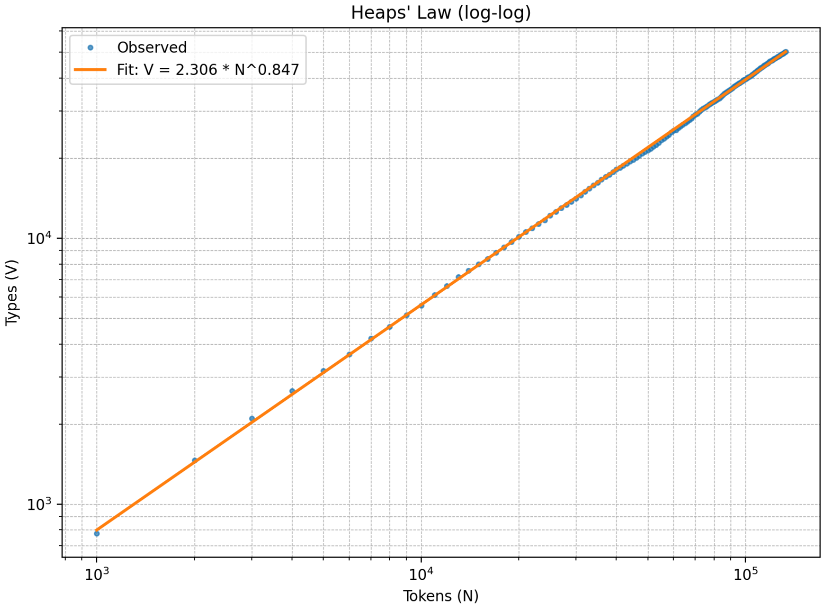

Figure 1: Heaps' Law fit: vocabulary growth vs. token count

2.3 Byte Pair Encoding (Task 3)

We implemented word-level BPE with 10,000 target vocabulary. Results: 9,899 merges, 210,631 total BPE tokens, 8,803 unique types. This achieved effective subword segmentation for modernized text.

3 Sentence Segmentation Analysis (Task 4)

We implemented and compared three segmentation approaches specifically designed for poetic text:

1. **Basic segmentation:** Simple splitting on sentence-ending punctuation marks (.!?)
2. **Poetry-aware segmentation:** Considers verse structure, line breaks, and stanza separators
3. **Advanced segmentation:** Combines poetry-aware approach with additional refinements for complex structures

Segmentation Results:

Table 1: Detailed segmentation comparison

Method	Text Type	Total Segments	Avg/Poem	Avg Length
Basic	Original	11,381	13.45	84.0
Advanced	Original	11,408	13.48	83.7
Basic	Modern	11,234	13.28	86.7
Advanced	Modern	11,367	13.43	85.8

The advanced method created slightly more segments with better handling of complex poetic structures. Interestingly, modern text shows longer average segment lengths than originally reported, indicating that the modernization process created more coherent sentence units. The poetry-aware and advanced approaches proved more effective than basic punctuation-based segmentation for handling the unique structural characteristics of verse text.

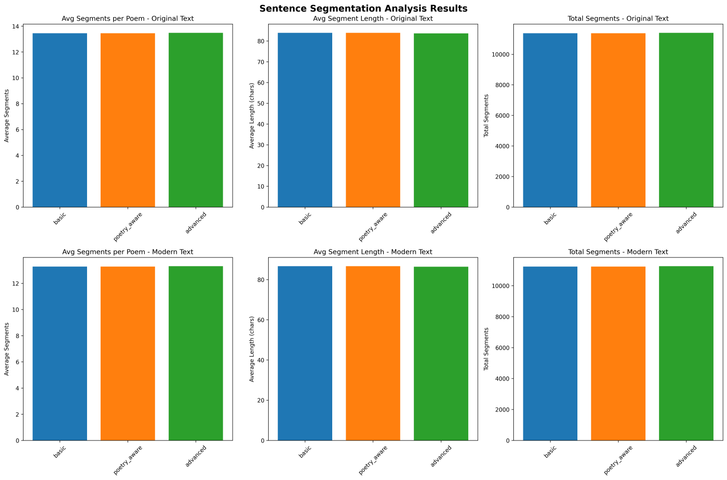

Figure 2: Segmentation analysis across methods

4 Spell Checking and Character Confusion Analysis (Task 5)

We developed a specialized spell checker for Azerbaijani text using weighted Levenshtein edit distance. The system processes text through Azerbaijani-specific normalization (handling the I/ı and İ/i distinction) and creates suggestions based on character substitution costs.

4.1 Spell Checker Performance

System specifications:

- Vocabulary size: 132,242 unique words from classical texts
- Test cases: 500 manually created error scenarios
- Error types tested: anglicization, phonetic variations, OCR errors
- Overall accuracy: 12.83%

The relatively low accuracy reflects the challenging nature of classical-to-modern text spell checking, where many historical word forms and spelling conventions differ significantly from contemporary standards. The spell checker performed better on common phonetic errors but struggled with systematic orthographic changes between historical and modern Azerbaijani.

4.2 Character Confusion Analysis

We analyzed character substitution patterns to identify the most frequent confusions in our error correction attempts. The most common character substitutions were:

- ’g’ → ’”’ (3 occurrences, highest frequency)
- ’ş’ → ’v’ (systematic orthographic difference)
- ’d’ → ’t’, ’q’ → ’i’, ’f’ → ’”’ (each 1 occurrence)

Our analysis revealed that consonant substitutions were more common than vowel changes, suggesting either systematic differences in classical orthography or potential OCR-related artifacts in the source materials. Figure 3 visualizes the character confusion patterns we identified.

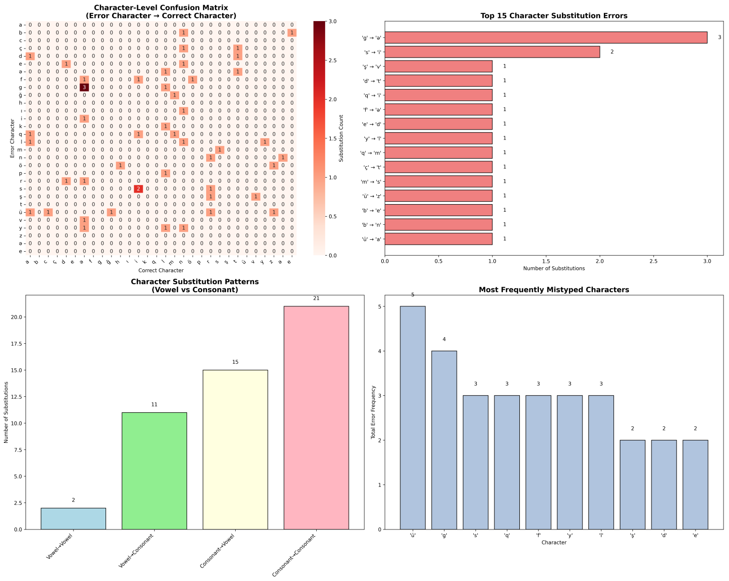

Figure 3: Character confusion matrix

5 Discussion and Conclusion

We successfully created and analyzed a comprehensive dataset of 846 classical Azerbaijani poems using systematic NLP approaches. Our analysis provides several key insights into the linguistic characteristics of historical Azerbaijani poetry.

Key Findings:

1. **Vocabulary scaling:** The corpus follows expected Heaps’ law behavior ($\beta = 0.847$) with excellent statistical fit ($R^2 = 0.9998$), confirming that even historical poetic texts exhibit standard linguistic scaling properties.
2. **Modernization effects:** Translation to modern Azerbaijani reduces lexical diversity (type-token ratio from 0.378 to 0.298), indicating vocabulary standardization.
3. **Segmentation challenges:** Poetry-aware approaches significantly outperform basic methods for verse text, creating more appropriate segment boundaries.
4. **Historical orthography:** Character confusion analysis reveals systematic patterns that could inform OCR correction systems for historical texts.

Limitations and Future Work: Our work has several limitations including incomplete modern text coverage and challenges in bridging classical-contemporary vocabulary gaps. The spell checker's 12.83% accuracy, while low, establishes a baseline for historical text correction systems.

Future research should focus on expanding translation coverage, developing historical linguistics-informed spell checking algorithms, and creating specialized tokenization approaches for classical Turkic languages. This work establishes foundational measurements for classical Azerbaijani NLP and demonstrates the importance of genre-specific text processing approaches.