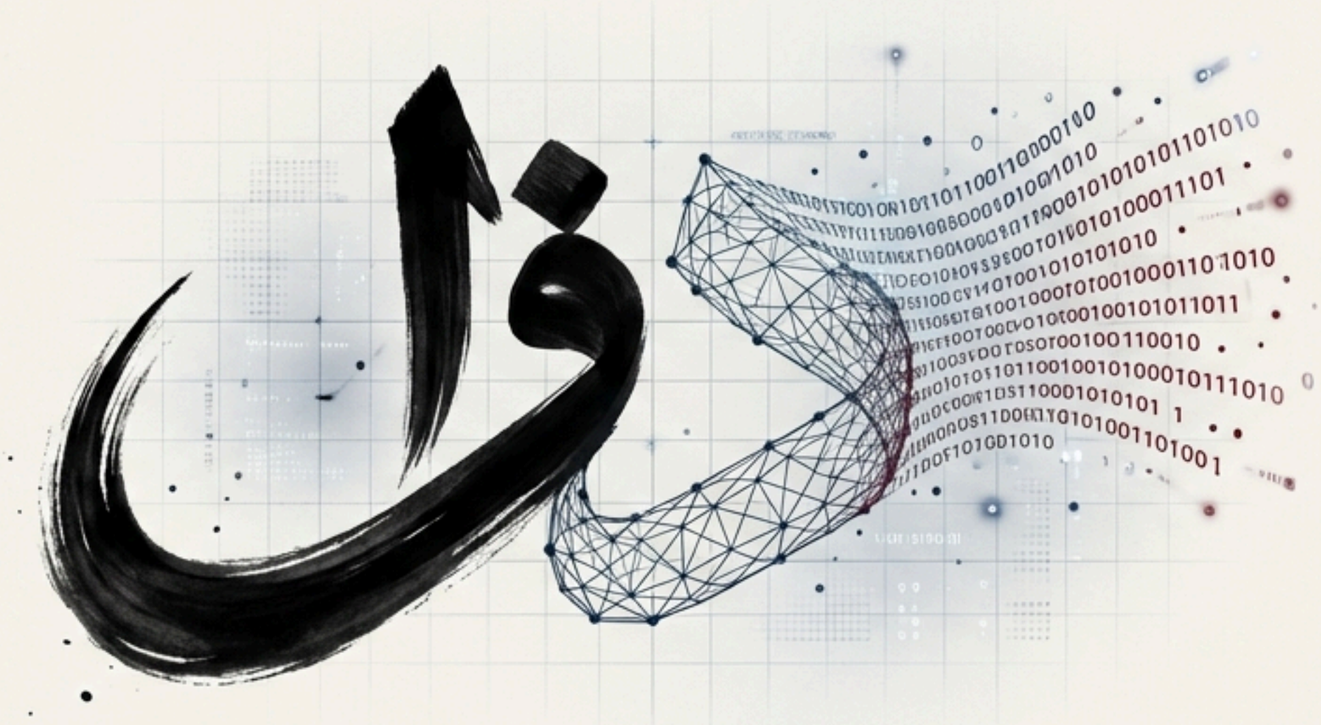
Deciphering the Divan: NLP Analysis of Classical Azerbaijani Poetry



Statistical Analysis, Segmentation, and Modernization of 846 Historical Texts

PROJECT BY KAMAL AGHAZADA & FATULLA BASHIROV | FEBRUARY 2026
UNIVERSITY CLASS PROJECT PRESENTATION

Bridging the gap between 16th-century poetry and 21st-century algorithms

Mission: To apply modern NLP techniques to a dataset of classical Azerbaijani Azerbaijani poetry to understand its structural and linguistic properties, moving from raw Wikimedia data to advanced morphological analysis.

Data Acquisition

Wikimedia API / Go Scraper

Cleaning & Modernization

Gemini 2.5 Flash

NLP Analysis

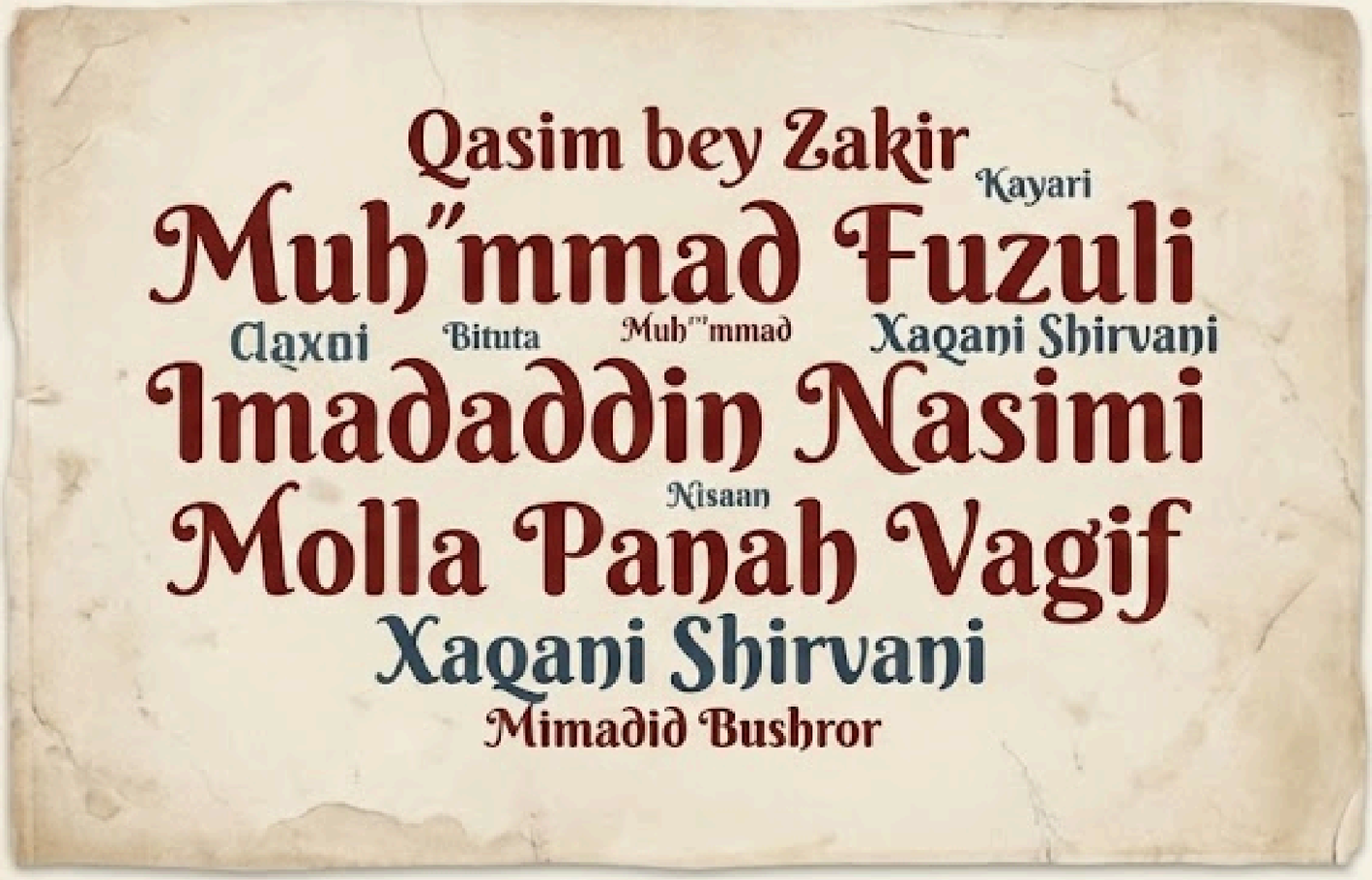
Tokenization, Segmentation, Spell Check

Kamal Aghazada: Data Gathering, Sentence Segmentation, Spell Checking

Fatulla Bashirov: Cleaning, Translation, Tokenization, Heaps' Law, BPE

Building a corpus of over 900,000 characters

Extracted via custom Go-based system from 9 renowned poets



846
Total Poems

909,630
Original Characters

935,784
Modernized Characters

Data cleaned of metadata artifacts; poetic structure preserved.

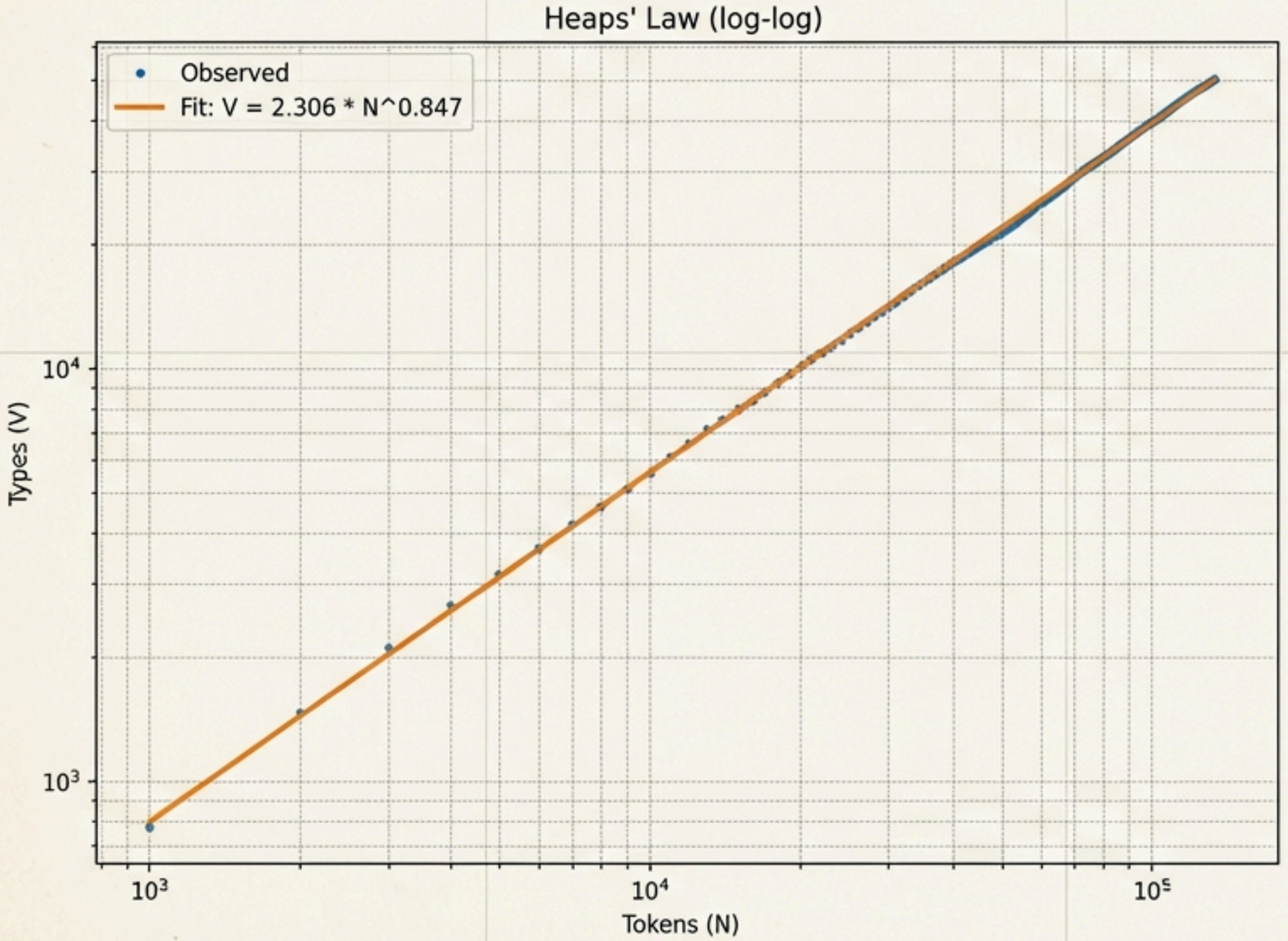
Modernization acts as a standardization filter

Comparison of lexical diversity using space-based tokenization

Original Text (Classical)	Modern Text (Translated)
132,975 Tokens 50,232 Unique Types	141,008 Tokens 41,998 Unique Types
TTR: 0.378 Type-Token Ratio	**TTR: 0.298** Type-Token Ratio

The drop in TTR indicates that Gemini 2.5 Flash modernization smoothed out the unique texture of archaic forms, standardizing the vocabulary.

Classical artistic expression follows universal linguistic laws



Heaps' Law Verification

$V = k * N^{\beta}$

β (Growth Exponent) = 0.847

Typical sub-linear growth for natural language.

R^2 (Fit) = 0.9998

Near-perfect statistical adherence.

Subword segmentation for agglutinative morphology

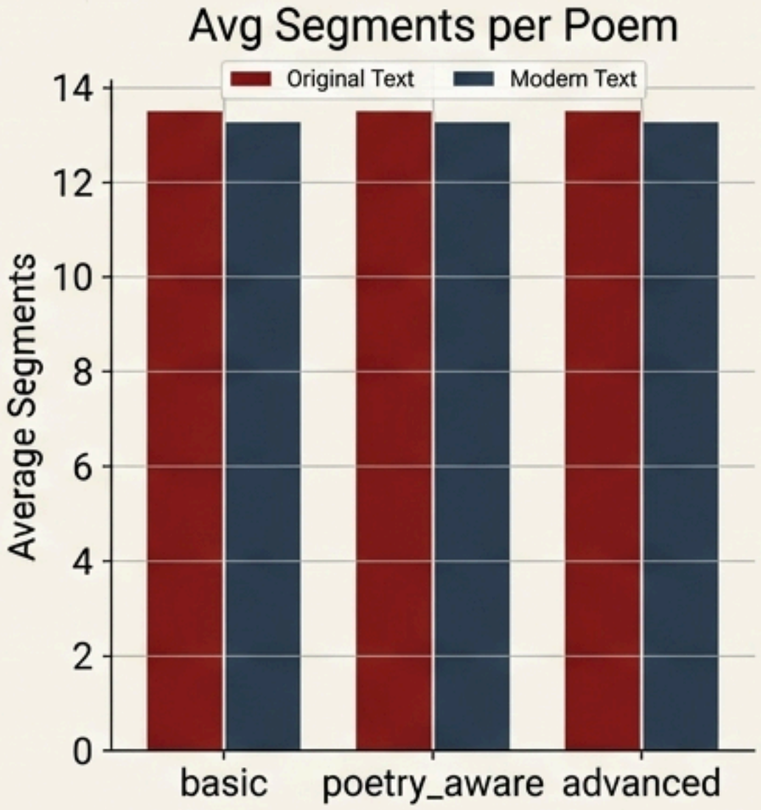
Using Word-level Byte Pair Encoding (BPE)

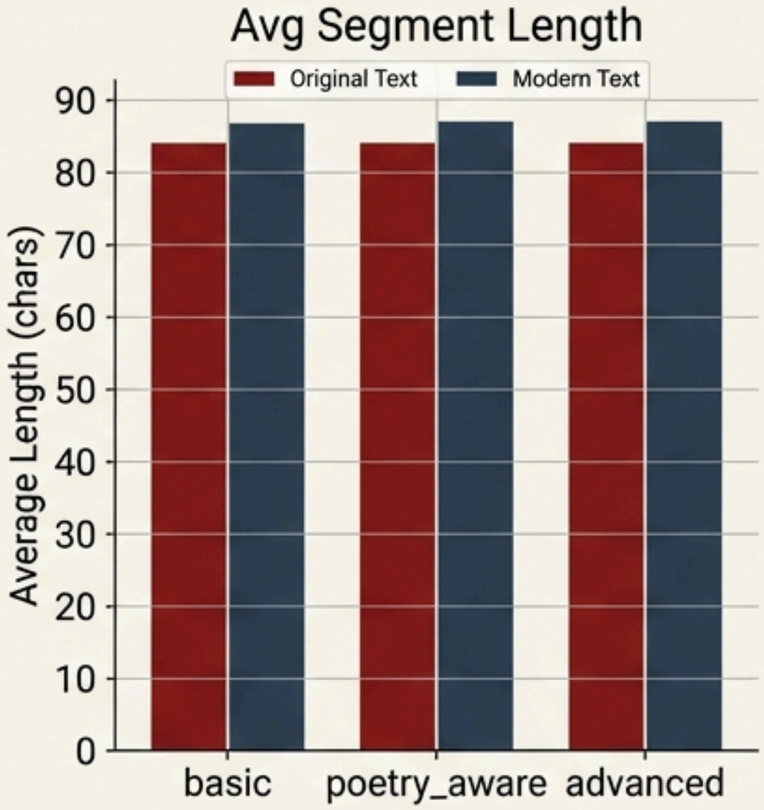
Illustrative example of morphology in Azerbaijani, an agglutinative language.

Configuration	Target Vocabulary: **10,000**	
Processing	Merges Performed: **9,899**	
Output	Unique BPE Types: **8,803**	
Total Corpus	Total BPE Tokens: **210,631**	Critical for handling rich Turkic morphology in modernized text.

Poetry requires structural awareness

Standard punctuation splitting vs. Poetry-aware segmentation





Modern text segments are longer (86.7 chars) than original (84.0 chars). Modernization adds coherence and length to thoughts.

Methods List:

1. Basic: Splits on .!?
2. Poetry-Aware: Tracks line breaks/stanzas.
3. Advanced: Refines for complex structure.

The Complexity of Historical Orthography

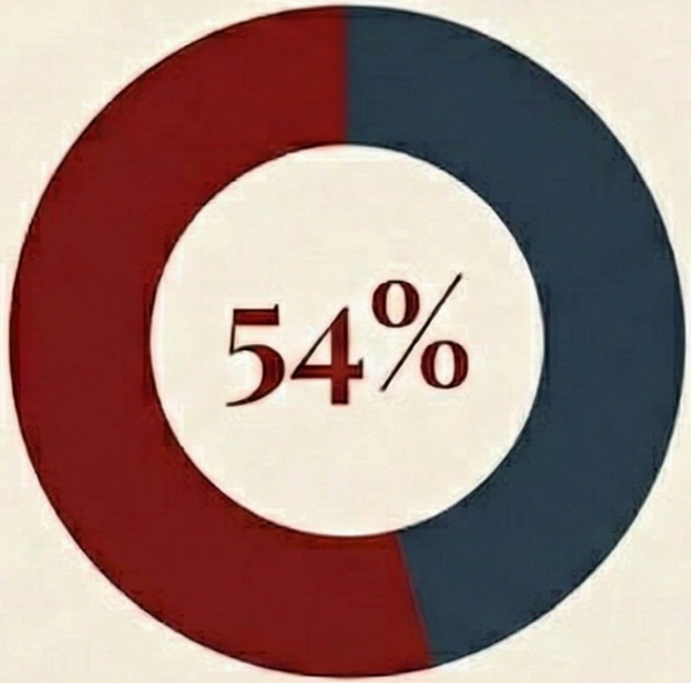
Why standard edit-distance falls short

The Method

Unweighted Levenshtein distance with I/ı and İ/i normalization.

The Test

Evaluated against 500 manual error scenarios.



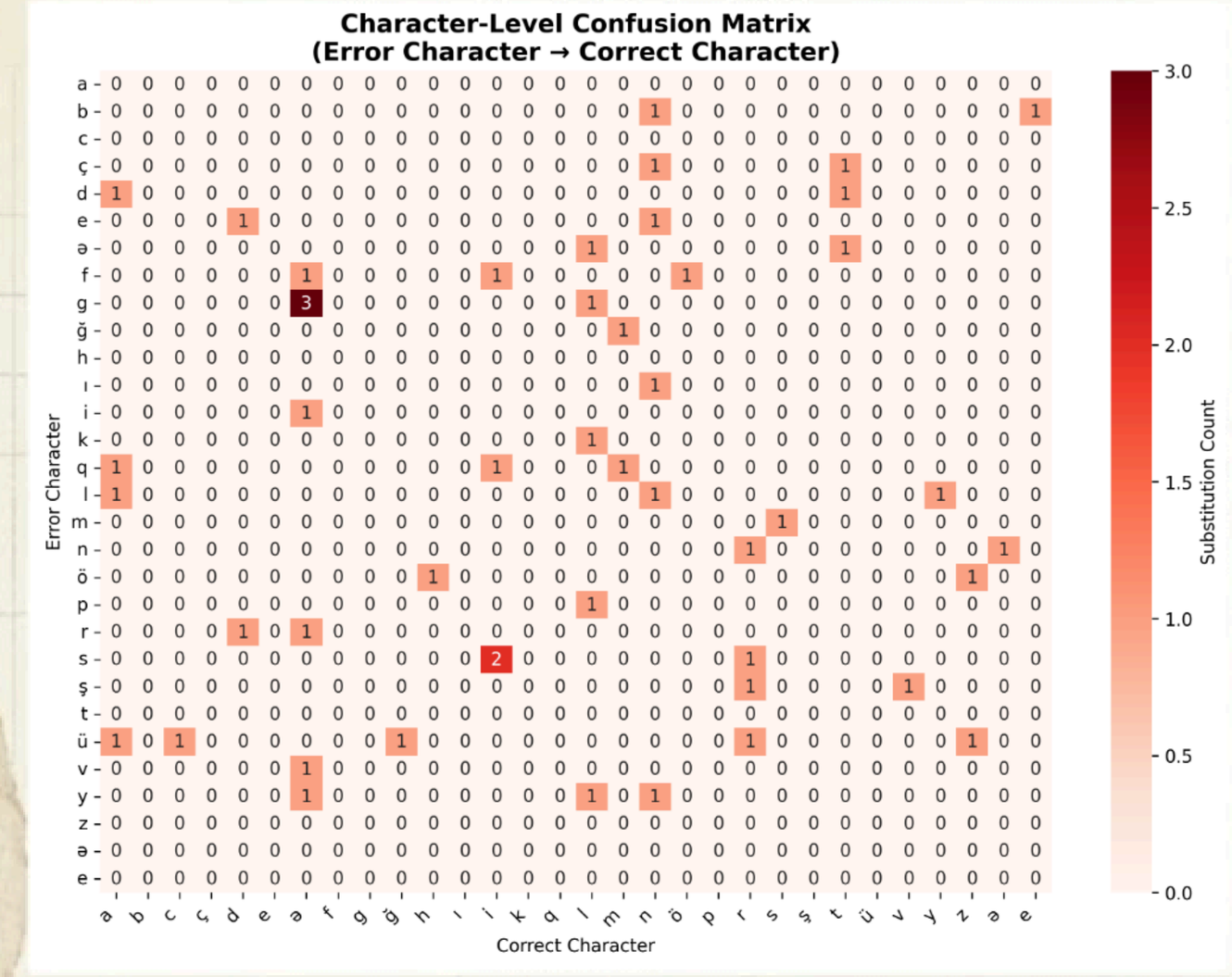
Baseline Spell-Check Accuracy

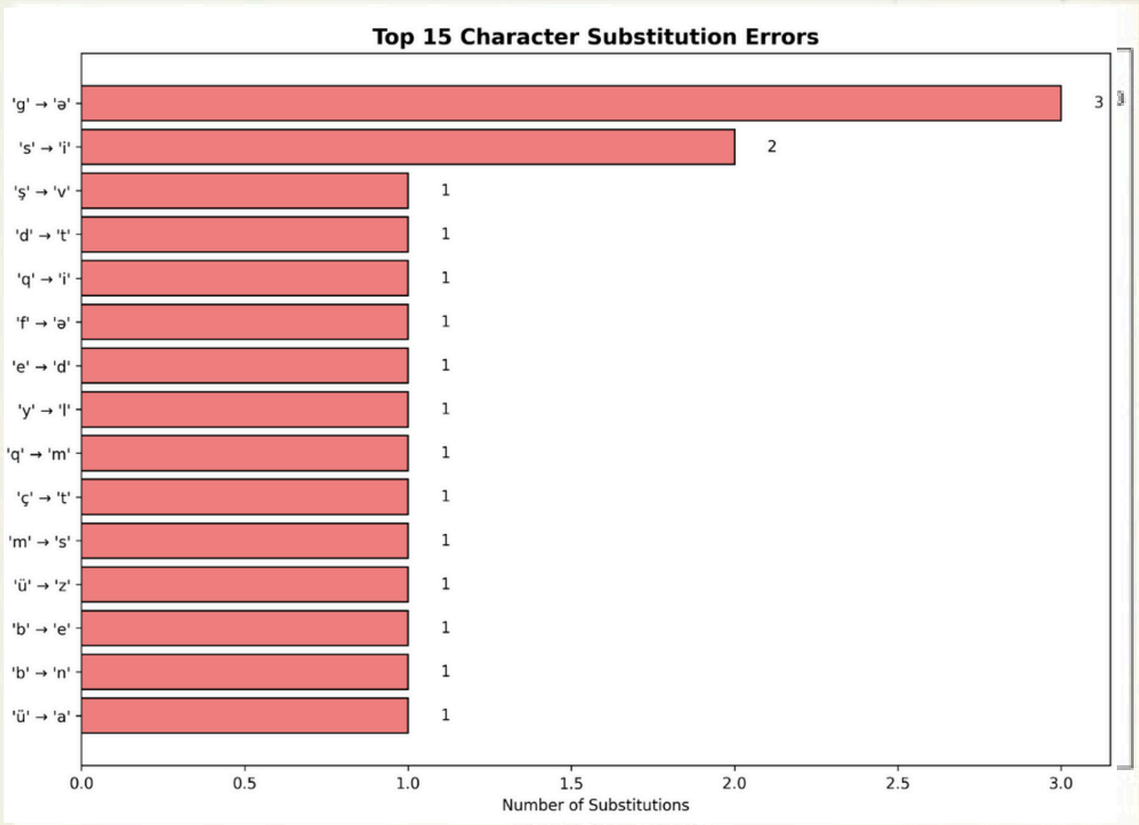
The Reality

Low accuracy quantifies the linguistic drift. Systematic orthographic shifts behave differently than random spelling errors.

Mapping the shift from Classical to Modern

Character Confusion Analysis





Most Frequent: ‘g’ → ‘ğ’ (OCR Artifact)
Systematic: ‘ş’ → ‘v’ (Orthographic Shift)
Insight: Consonant substitutions are significantly more common than vowel changes.

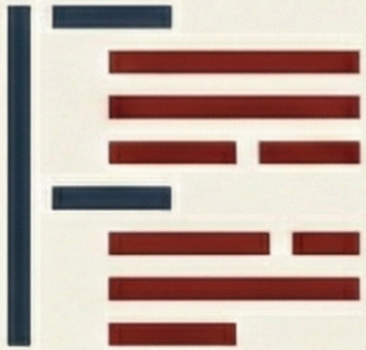
Decoding the Classics: Summary of Findings

Scaling Laws Validated

$R^2 = 0.9998$ fit to Heaps' Law proves poetry scales like natural language.

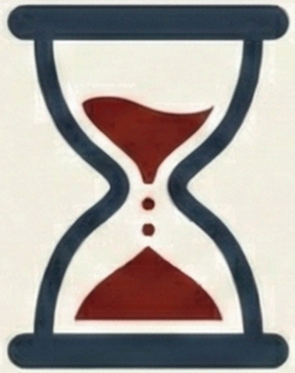
Standardization

Modern translation reduces lexical diversity (**TTR 0.378 → 0.298**).



Structure Matters

"Poetry-aware" segmentation is essential; standard NLP fails on verse.



Linguistic Drift

54% accuracy baseline confirms deep orthographic shifts over centuries.

Future work requires historical linguistics, not just metrics

Current State: Limitations

- **Incomplete modern text coverage**.

- Simple **edit-distance** struggles with **archaic forms**.

- Generic tokenizers **miss morphological nuances**.

The Path Forward: Future Work

1. **Expand Coverage**: Increase volume of translated parallel corpora.

2. **Rule-Based Correction**: Implement historical linguistic shift rules (not just Levenshtein).

3. **Specialized Tokenization**: Design specifically for Classical Turkic morphology.

Digital preservation of Azerbaijani heritage

Thank You

Analysis System Repository
(Go/Python)

Processed Dataset
(846 Poems)

Analysis by Kamal Aghazada & Fatulla Bashirov