Computational Documentation of Laki: A Novel Framework for an Unwritten Endangered Language

Mahboobeh Fathollahimiragha

Abstract

Laki, an endangered Northwestern Iranian language, faces critical preservation challenges due to absent standardized writing system and limited computational resources. This paper presents first comprehensive computational framework for Laki documentation, introducing Laki-Poet dataset of 140 poems (1,400 lines) from Delfan region with phoneme-based writing system. Methodology includes custom tokenization, morphological analysis, metaphor identification achieving 85% precision. Work establishes foundational Laki computational linguistics resources with transferable methods for unwritten languages.

Keywords: Laki language, endangered languages, computational linguistics, low-resource NLP

1. Introduction

Linguistic diversity erosion represents pressing challenge in cultural heritage and computational linguistics. Among 7,000+ global languages, nearly half endangered, many lacking writing systems. Laki, Northwestern Iranian language, exemplifies crisis-rich poetic traditions facing loss risk from computational absence. Language presents unique computational challenges unaddressed in NLP literature. Rich morphology, metaphorical poetry, writing system absence create preservation obstacles. Existing low-resource approaches assume orthographic consistency, failing oral languages. We propose novel framework combining linguistic expertise with computational innovation. Contributions: phoneme-based writing system, Laki-Net digital corpus, baseline NLP tools for Laki structure.

2. Related Work

Low-resource language research advanced significantly, gaps remain unwritten languages. Previous work: General low-resource NLP-transfer learning, multilingual BERT, data augmentation assume basic orthography. Iranian language processing-Persian NLP matured, Kurdish dialects work exists, Laki computational approaches nonexistent. Endangered language documentation-computational documentary linguistics developed preservation tools, focus audio recording not full NLP poetic analysis. Our work bridges domains addressing computational methods development for language lacking digital resources and standardized orthography.

3. Methodology

3.1 Data Collection and Annotation

Laki-Poet corpus represents decade fieldwork, data collection initiated 2012, digital annotation through 2024. Corpus: 140 poems (1,400 lines) family manuscripts Delfan region, Central Laki dialects. Longitudinal approach combines traditional documentation modern methods. Case Study: Frog riddle poem analysis-auditory personification, behavioral metaphor, visual simile, ecological contextualization. Annotation: 2012-2015 initial transcription, 2020-2024 digital processing, phonetic transcription IPA-based, metaphor labeling, structural analysis, cultural context.

3.2 Orthographic System Development

Novel writing system addresses: vowel length distinctions, consonantal inventory (pharyngeal /ħ/, uvular /q/), suprasegmentals poetic recitation, native speaker validation

3.3 Computational Pipeline

Processing framework: tokenizer agglutinative morphology support, morphological analyzer suffix decomposition, metaphor detector cultural context rules, evaluation gold standard.

4. Experiments and Results

4.1 Metaphor Identification

System identified four metaphorical patterns 85% precision: personification natural phenomena 100%, visual simile 80%, ecological metaphor 75%, behavioral metaphors 85%.

4.2 Tokenization and Morphological Analysis

Tokenization accuracy 92%, morphological constructions 88%, agglutinative suffixes 78%.

4.3 Performance Metrics

TABLE 1: COMPUTATIONAL PERFORMANCE

Task	Precision	Recall	F1-Score	Baseline
Tokenization	0.92 ± 0.03	0.89 ± 0.04	0.90 ± 0.02	0.76
Metaphor Detect	0.85 ± 0.05	0.82 ± 0.06	0.83 ± 0.04	0.62
Morph Analysis	0.78 ± 0.04	0.75 ± 0.05	0.76 ± 0.03	0.58

4.4 Writing System Evaluation

Native speaker acceptance 92%, inter-annotator agreement κ=0.87, cross-dialect consistency 85%.

5. Comparative Analysis

TABLE 2: INTERNATIONAL PROJECT COMPARISON

Metric	Our Work	Navajo NLP	Inuktitut	NLLB
Data Volume	140 poems	50 texts	70 stories	1M+ sentences
Token Accuracy	92%	85%	88%	95%
Writing System	Novel	Existing	Existing	Existing
Metaphor Proc	Yes	No	No	No
Languages	1	1	1	200+
Community	15 spk	8 spk	12 spk	Limited

5.1 Key Innovations

First framework completely unwritten languages-solves fundamental challenge absent writing systems, provides solution 3,000+ similar languages. Unique combination linguistics AI-deep expertise advanced methods, preserves cultural nuances automated processing. Community-centered approach-15 native speakers design evaluation, addresses real linguistic community needs.

5.2 Limitations and Future Work

Scalability larger corpora, dialectal variations handling, neural approaches limited data, audio processing oral traditions integration.

6. Broader Impact and Conclusion

6.1 Broader Impact

Immediate Impact: Academic-first Laki computational resources, Cultural-digital preservation endangered traditions, Community-tools language revitalization, Methodological-blueprint unwritten languages. Laki Community Impact: Empowerment-digital tools local community, Cultural Valorization-pride revival documentation, Education-digital materials foundation, Continuity-intergenerational

knowledge transfer. Global Significance-3,000+ endangered languages lack resources, framework offers scalable documentation, community-academic collaboration, technical-cultural balance.

6.2 Future Work Timeline

2025: Expand 500+ poems audio recordings

2026: Develop Laki speech recognition

2027: Create educational tools teaching

2028: Extend framework Iranian languages

6.3 Conclusion

Research establishes first computational Laki framework, new paradigm preserving unwritten endangered languages. Immediate consequences: Academic Impact-founding unwritten language processing field, Social Impact-template 3,000+ endangered languages, Methodological Impact-innovative oral tradition digital technology combination. Future vision: framework standard endangered language projects, lasting contribution global linguistic diversity maintenance.

References

[1] Bird, S. (2020). Decolonising computational linguistics for endangered languages. Language Documentation & Conservation.

[2] Joshi, P., et al. (2020). The State and Fate of Linguistic Diversity in NLP. ACL 2020.

[3] McCurdy, K., & Risam, R. (2021). Digital Humanities and Endangered Language Preservation. DH Quarterly.

[4] Sproat, R. (2021). Computational Methods for Unwritten Languages. Computational Linguistics.

[5] Zeman, D., et al. (2023). Universal Dependencies for Low-Resource Languages. LREC 2023.

[6] Anastasopoulos, A. (2022). Computational Language Documentation. ACM Transactions.

[7] Fathollahimiragha, M. (2025). Laki-Poet Dataset. Personal Collection.