4th Global Conference on Social Sciences
12-14 December 2025 I Dubai, United Arab Emirates

Error Taxonomy and Benchmark Framework for Kazakh Morphological Analyzers Annotation

Gauhar Munaitbas[1], Laura Baitenova[2], Gulnar Mukhamejanova[3]

[1]Turan University, Almaty, Kazakhstan
[1,2]Home Credit Bank JSC, Almaty, Kazakhstan
[3]Narxoz University, Almaty, Kazakhstan

ABSTRACT

Automating the morphological analysis of the Kazakh language faces constant difficulties due to its agglutinative structure, diverse affixation patterns, and high variability in surface forms. Despite the development of several linguistic tools and statistical resources, the evaluation of such systems remains fragmented and lacks methodological standardization. Existing publications demonstrate heterogeneous evaluation protocols, incompatible datasets, and local performance metrics, which hinder reproducibility and make cross-system comparison unreliable. This study addresses these limitations by proposing a unified error taxonomy and a benchmark system for evaluating Kazakh morphological analyzers. We present a four-level error classification – structural, syntactic, segmentation-based, and semantic – accompanied by a multi-stage testing protocol that evaluates surface-level recognition, lemmatization accuracy, and the reliability of morphological tags. The foundation was tested on a manually annotated corpus containing linguistically complex constructions marked by expert philologists. Three main paradigms were evaluated: rule-based systems, CRF-based models, and transformer architectures (mBERT, KazBERT, KazRoBERTa). For each model, we conduct a detailed evaluation based on errors and a statistical classification of failure cases. The results show that the proposed framework allows for an objective comparison of different models, identifies hidden algorithmic shortcomings, and supports targeted model improvement.

As far as we know, this is the first attempt to create a standardized error taxonomy and a reference protocol for the morphological analysis of the Kazakh language, which can also be applied to other Turkic languages. The proposed methodology has practical value for error-based learning, corpus development, and the evaluation of future NLP systems for low-resource agglutinative languages.

Keywords: Kazakh language, morphological analysis, error taxonomy, quality assessment, benchmark, NLP, agglutinative languages, transformational models