BG Pulse Journal USA
ISSN:1947-2536 http://www.bgpulseusa.com/index.php/bg

REVIEWER PROCESSES AND PARALLEL DICTIONARY CREATION IN ANNOTATING UZBEK DIALECTS FOR TRAINING ARTIFICIAL INTELLIGENCE MODELS ON THE EXAMPLE OF THE KHOREZM DIALECT

Shakhodat Qodirberganova
Department of Modern information technologies
Uzbekistan state world languages university
https://orcid.org/0009-0006-1247
https://doi.org/10.5281/zenodo.20520738

Abstract

This study presents the methodology for annotating the Khorezm dialect, a low-resource variety of the Uzbek language, and developing parallel dictionaries for training artificial intelligence (AI) models. The Khorezm dialect differs significantly from Standard Uzbek in its phonetic, lexical, and morphological features and, as a representative of the Oghuz dialect group, exhibits considerable linguistic affinity with the Turkish language. The article examines the role of reviewer processes in improving dataset quality and analyzes the impact of parallel dictionaries on the performance of Neural Machine Translation (NMT) systems. The findings highlight the importance of systematic review procedures and high-quality parallel lexical resources for enhancing the effectiveness of AI-based language technologies for dialectal Uzbek.

Keywords: Artificial Intelligence (AI), Uzbek language, dialect annotation, reviewer process, Khorezm dialect, Natural Language Processing (NLP).

Introduction

Artificial intelligence and natural language processing (NLP) systems need high-quality annotated data, but such databases are scarce for languages with limited resources, such as Uzbek. Among the regional dialects of the Uzbek language, the Khorezm dialect stands out with its unique identity. Since this dialect belongs to the Oghuz group of the Turkic language family, it has common phonetic and lexical features with Turkish and Azerbaijani languages, and is completely different from the literary Uzbek language (Qarluq group).

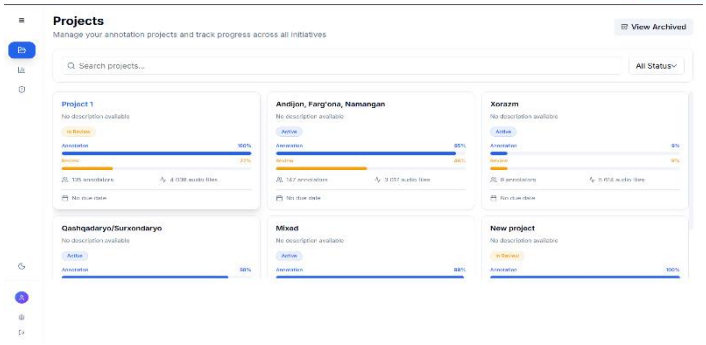
Currently, existing AI tools are mainly adapted to the literary language, which leads to errors in recognizing and translating dialect-related speech. The purpose of the research is to systematize reviewer processes in the annotation of the Khorezm dialect and increase the accuracy of AI models by creating parallel dictionaries "dialect → literary language".

Methods

The following methodological steps were used in the study to create a quality database:

- Data collection: Audio samples of everyday and formal speech were collected from different regions of the Khorezm oasis.
- Reviewer Process: A three-stage system of experts with excellent knowledge of the dialect was implemented to ensure the quality of the annotation:
    1. Initial annotation review;
    2. Peer review (agreement between several experts);
    3. Final approval.

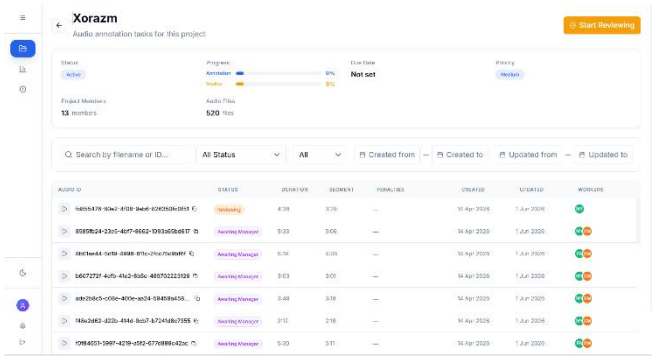
BG Pulse Journal USA
ISSN:1947-2536 http://www.bgpulseusa.com/index.php/bg
BG PULSE JOURNAL



- Technical tools: To manage processes and track error history mohirdev.labelcraft.org digital collaboration platform was used.
- Parallel dictionary formation: Khorezm dialect units were aligned with literary language synonyms and enriched with lexical-grammatical tags (noun, adjective, verb, etc.). In this, phonetic and morphological variants (for example, differences in verb tenses) indicating the dialect's proximity to the Turkish language were taken into account.

Results

The following indicators were achieved as a result of the study:



- Reviewer efficiency: Systematic control increased inter-annotator agreement and reduced subjective interpretations.
- AI model accuracy: Models enriched with parallel vocabulary (M1, M3 configurations) showed significantly higher results compared to the baseline models (M0). In particular, lexical accuracy increased and the TER (edit distance) coefficient decreased, which confirms the model's ability to correctly "normalize" dialect-specific words.

Phonetic compatibility: The changes made by the reviewers allowed AI systems to more accurately train the unique vowel sounds and pronunciation features of the Khorezm dialect.

Discussion

The complex morphological and phonetic structure of the Khorezm dialect, especially its features specific to the Oghuz dialect, poses challenges for traditional "rule-based" systems. The study shows that even in the era of neural translation, hand-crafted linguistic resources and expert review (reviewer) do not lose their importance.

The dialect's proximity to Turkish (e.g., similarities in the lexical layer) creates both a challenge and an opportunity in training AI models: on the one hand, it increases the difference from literary Uzbek, and on the other hand, it allows the use of general NLP tools created for Turkic languages (e.g., TurkicNLP). Open and consistent reviewer processes are key to building inclusive AI systems.

Conclusion

Reviewer processes are a key factor in ensuring the reliability of the database in the annotation of the Khorezm dialect. The created parallel dictionaries and systematic control mechanisms serve as a solid foundation for digitizing Uzbek dialects and developing culturally adaptive AI models.

References:

1. Ölmez, Mehmet. (2018). *Türk Dilleri ve Lehçeleri*. Ankara: Akçağ Yayınları.
2. Bender, Emily M.. (2019). The data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 7*, 587–604.
3. Zampieri, Marcos, et al. (2020). Language identification and dialect classification in low-resource settings. *Computer Speech & Language, 62*, 101088.
4. Yusupov, Erkin. (2019). Xorazm shevalarining fonetik va leksik xususiyatlari. *O‘zbek Tili va Adabiyoti*, 5, 45–53