SciDataCon 2025

Contribution ID: **269** Type: **Poster**

Leveraging Corpus Linguistics for Linguistic Research in Kazakh: A Data-Driven Approach

Monday 13 October 2025 19:10 (20 minutes)

The field of corpus linguistics has revolutionised linguistic research by providing data-driven insights into the structure, usage, and evolution of languages. By leveraging large-scale text corpora, researchers can uncover patterns in grammar, vocabulary, syntax, and language use that are not easily observable through traditional methods (Omarova et al., 2025). This data-driven approach offers a powerful tool for both theoretical and applied linguistic studies, particularly for languages such as Kazakh, which face challenges in terms of linguistic resources and computational tools. In our study, we explore the application of corpus linguistics to the Kazakh language, focusing on the creation, analysis, and implications of Kazakh language corpora for linguistic research. More specifically, our research lies in investigating the language contact called interference (Ormanova & Anafinova, 2022).

We present the methodologies employed in building Kazakh language corpora, including the selection of texts, the annotation process, and the use of computational tools for text analysis. We discuss the specific challenges encountered in working with Kazakh, including issues related to its agglutinative nature, complex morphology, and the lack of comprehensive, digitised language resources. A significant portion of the presentation focuses on how corpus linguistics has been utilized to investigate the borrowings from English and Russian due to the policy of trilingualism in Kazakhstan.

Furthermore, we will discuss the potential applications of Kazakh language corpora beyond academic research. These corpora hold significant promise for practical uses, such as in language education, machine translation, and speech recognition technologies. Our study outlines how corpus-based insights can be used to inform language teaching materials, contribute to the development of language resources for artificial intelligence, and support language preservation efforts for Kazakh, particularly in light of the ongoing sociolinguistic shifts within Kazakhstan.

Materials and methods. We generated a corpus of media texts in the Kazakh language (700 texts, 374087 words); we carried out a comparative analysis of statistical linguistic data (word occurrences) by using the computer program #LancsBox 6.0.

Results: A corpus analysis showed that borrowings from English and Russian are actively used in Kazakh (e.g., guide, speaker, PR, draft, team building, etc.). Along with widespread use, most borrowings are not included either in dictionaries of the Kazakh language or the official terminological base Termincom.kz.

There is a tendency in the use of borrowings when both borrowings and equivalent national variants are used in texts at the same time. The difference is observed in the number of occurrences. On the one hand, Kazakh words are dominant. For example, the Kazakh мәселе [masele] / problem has 696 occurrences in the corpus. However, along with this, there is a Russian translation equivalent проблема [problema] / problem with 109 occurrences in Kazakh texts. At the same time, there are cases where the dominance of a borrowed word is observed along with the already existing and approved Kazakh version. So, in the generated corpus, the borrowed word музей [muzei] / museum has 20 occurrences in Kazakh texts, while the Kazakh мұражай [murazhay] / museum has only 14 occurrences, which indicates the prevailing norms of a foreign language. Even though the difference is not significant, the fact of using the word indicates interference from the Russian language to the Kazakh vocabulary.

Thus, the implementation of the tools of corpus linguistics can enhance the research in linguistic data. A data-driven approach could highlight how the corpus data has been used to study the Kazakh language, providing insights into syntax, semantics, language variation, or language change. Through the creation and analysis of extensive corpora, linguists and researchers are better equipped to understand and document the linguistic intricacies, contributing to the broader field of corpus linguistics and the preservation of linguistic diversity in the digital age.

References:

Ormanova A. B. & Anafinova M.L. (2022) Linguistic Interference in Information Space Terms: A Corpus-Based Study in Kazakh. Theory and Practice in Language Studies, 12 (12), 2497-2507. DOI: https://doi.org/10.17507/tpls.1212.04 https://tpls.academypublication.com/index.php/tpls/article/view/5095

Omarova, S., Ospanova, D., Aitova, N., Tokenkyzy, G., Ormanova, A., & Alshynbekova, M. (2025) A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh. Forum for Linguistic Studies, 7(2), 869–881. https://doi.org/10.30564/fls.v7i2.8317

Primary author: ORMANOVA, assel (Astana IT University)

Presenter: ORMANOVA, assel (Astana IT University)

Session Classification: Poster Session

Track Classification: SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global