ACADEMIA ONE
PEER REVIEWED JOURNALS

Diversity Research: Journal of Analysis and Trends
Volume 4, Issue 04, April 2026
ISSN (E): 2810-6393
Website: https://academiaone.org/index.php/2

On The Principles Of Creating An Electronic Linguistic Database Of Translation Units In Literary Texts Within Corpus Linguistics In Turkology

A.R. Iskandarova
National University of Uzbekistan
Doctor of Philosophy (PhD) in Philology, Associate Professor
iskandarova@gmail.com

Annotation: This article examines the principles for constructing an electronic linguistic database of translation units derived from literary texts within the framework of corpus linguistics in Turkology. It elucidates the role of parallel corpora in the analysis of translation processes and in the identification of linguistic features across Turkic languages. The study delineates the principal stages of database development, including data collection, preprocessing, alignment, annotation, and system implementation. Particular emphasis is placed on the capabilities of Uzbekcorpus.uz as a significant resource for the study of literary texts and translation units. The findings indicate that such databases contribute to the improvement of translation quality, facilitate in-depth linguistic analysis, and promote the advancement of corpus-based research in Turkology.

Keywords: corpus linguistics, parallel corpora, literary texts, translation units, electronic linguistic database, Uzbekcorpus.uz

Introduction

In contemporary linguistics, corpus linguistics has evolved into an autonomous discipline, offering extensive opportunities for the empirical investigation of language units. Within the field of Turkology in particular, the significance of parallel corpora is steadily increasing in the study of interrelations among Turkic languages, translation processes, and their inherent linguistic features.

Literary texts constitute one of the most complex and linguistically rich strata of language, as they uniquely embody lexical, semantic, and stylistic dimensions. Consequently, the development of electronic linguistic databases for the systematic analysis of literary translations has emerged as a pertinent and essential scholarly task. The primary objective of the present study is to examine electronic linguistic databases of translation units based on literary texts within the framework of corpus linguistics in Turkology.

Methods

Within Turkology, substantial research has been conducted in the domain of corpus linguistics. In particular, the Turkish language corpus has been investigated by Deniz Aksan, Kemal Oflazer, and Umut Özge, while studies on the Uyghur language corpus have been carried out by Yusup Aibaidulla and Kim-Teng Lua. In addition, noteworthy research has been undertaken on the corpora of Bashkir, Crimean Tatar, and Tuvan languages.

At present, corpus linguistics in Turkology is undergoing steady development and has become a significant field for the construction and analysis of text corpora of Turkic languages. Scholars such as N. Yoqubova, M. Ayimbetov, S. Rizayev, and S. Muhamedov have made notable contributions in this area. Concurrently, considerable progress has been observed in computational linguistics. Research conducted by A. Pulatov, A. Rahimov, Z. Xolmanova, and N. Abdurahmonova is of substantial theoretical and practical significance, contributing to the study of Turkic languages through modern technologies.[1]

ACADEMIA ONE
PEER REVIEWED JOURNALS
Diversity Research: Journal of Analysis and Trends
Volume 4, Issue 04, April 2026
ISSN (E): 2810-6393
Website: https://academiaone.org/index.php/2

Professor G‘. Salomov defines translation as a creative process of reconstructing meaning through the means of another language, essentially regarding it as an art of words.[2] According to his perspective, knowledge of the dictionary meanings of words alone is insufficient for translation. A translator must apprehend the subtle contextual nuances of a word, including whether it is archaic or contemporary, figurative, honorific, euphemistic, ironic, sarcastic, emotive, or even offensive. Moreover, it is necessary to identify an equivalent expression in the target language that accurately conveys these semantic and stylistic characteristics. Consequently, translation requires not only linguistic competence but also an intuitive sensitivity to language.

National corpora developed within Turkology constitute essential resources for the scientific study of language. For example, the Turkish Language National Corpus (TUD) contains more than 50 million words and represents a comprehensive electronic database encompassing diverse genres and domains. It is based on both written and spoken samples of modern Turkish and provides users with advanced search and filtering functionalities.

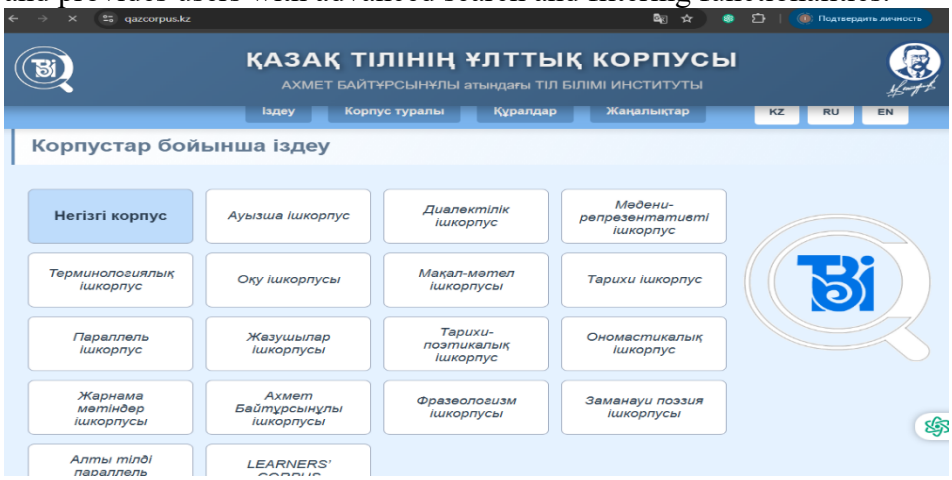

Figure 1. The Kazakh National Corpus

The Kazakh National Corpus, comprising 31 million words, encompasses the principal functional styles of the Kazakh language and enables users to analyze word usage across diverse contexts, as well as their grammatical characteristics.

The “Tügen Tel” [3] constitutes a large-scale corpus (180 million tokens) that covers texts from a wide range of genres and styles. It is distinguished by its detailed morphological annotation and supports searches based on both lexical and grammatical parameters. Furthermore, its socio-political subcorpus facilitates the analysis of language development and the formation of terminology.

These corpora provide a robust scientific and practical foundation for the investigation of authentic language use, translation processes, and linguistic research within Turkic languages.

ACADEMIA ONE
PEER REVIEWED JOURNALS
Diversity Research: Journal of Analysis and Trends
Volume 4, Issue 04, April 2026
ISSN (E): 2810-6393
Website: https://academiaone.org/index.php/2

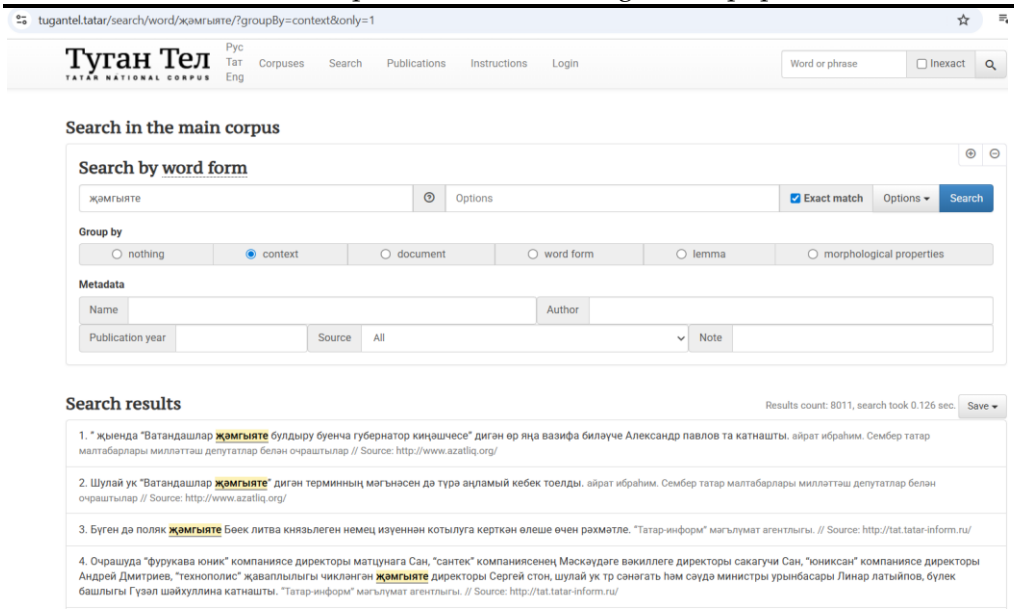

Figure 2. "Tugan Tel" Tatar National Corpus

Results

The creation of an electronic linguistic database of translation units in literary texts constitutes a process of automated analysis and systematic organization of translation units based on parallel corpora, while preserving their aesthetic, connotative, and cultural features.[4] This database enables the selection of phraseological, metaphorical, and stylistic units in accordance with contextual compatibility and facilitates the identification of translation equivalents between Uzbek and other languages (such as English).[5]

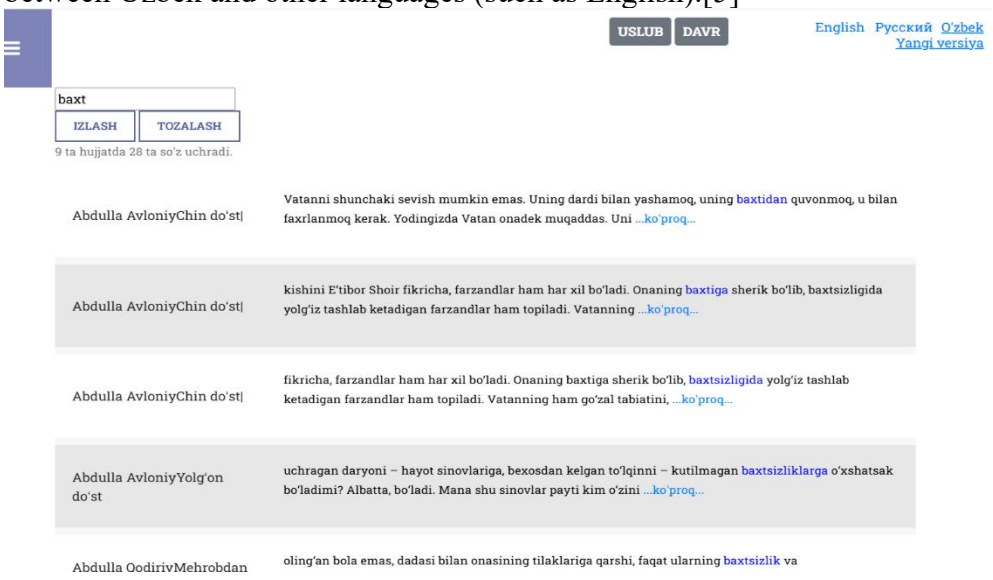

Figure 3. Uzbekcorpus.uz

Stages in the Creation of an Electronic Database of Literary Texts:

Data collection and digitization: The formation of electronic (textual) versions of both source texts and their translations.

Preprocessing: The cleaning, formatting, and standardization of texts to ensure consistency and usability.

Segmentation and alignment: The parallelization of texts at the level of sentences or translation units (such as phrases and word combinations) through the alignment of corresponding segments across two languages.

ACADEMIA ONE
PEER REVIEWED JOURNALS
Diversity Research: Journal of Analysis and Trends
Volume 4, Issue 04, April 2026
ISSN (E): 2810-6393
Website: https://academiaone.org/index.php/2

Annotation and encoding: The linguistic tagging and analytical processing of units, including phraseological expressions and metaphors.
Database construction: The integration of processed data into a system equipped with search and analytical functionalities.

Figure 4. Uzbekcorpus.uz

Uzbekcorpus.uz [7] constitutes the National Corpus of the Uzbek language, encompassing texts from a wide range of genres and styles, including a specialized database of literary texts. The literary texts database represents one of the most significant components of the corpus. It comprises prose and poetic works from both Uzbek and world literature, including novels, short stories, novellas, and dramatic texts.[6]

These materials are systematically organized according to criteria such as author, historical period, genre, and stylistic characteristics, thereby enabling structured and comprehensive analysis. The texts included in this database undergo linguistic processing and annotation. In particular, words are morphologically tagged, which allows users to examine their grammatical forms and observe their usage across diverse contexts. Furthermore, the search system provides users with the capability to retrieve authentic examples illustrating how specific words, phrases, or syntactic constructions are employed in literary texts.

Discussion

The obtained results indicate that an electronic linguistic database developed on the basis of literary texts provides extensive opportunities for the in-depth analysis of translation processes. In particular, this approach proves effective in identifying similarities and differences among Turkic languages.

The literary texts database plays a pivotal role in several respects, including the examination of the natural use of lexical and stylistic units, the analysis of phraseological and metaphorical expressions, the identification of translation equivalents, and the application of authentic materials in language teaching and linguistic research.[2] Overall, the literary texts database within Uzbekcorpus.uz constitutes a valuable electronic linguistic resource that reflects the semantic, stylistic, and aesthetic richness of the Uzbek language and is widely utilized in corpus linguistics, translation studies, and literary analysis.

Conclusion

The findings of the study have led to the formulation of both theoretical and practical principles for the creation of an electronic linguistic database of translation units based on literary texts within the framework of corpus linguistics in Turkology. This approach contributes to

ACADEMIA ONE
PEER REVIEWED JOURNALS
Diversity Research: Journal of Analysis and Trends
Volume 4, Issue 04, April 2026
ISSN (E): 2810-6393
Website: https://academiaone.org/index.php/2

improving translation quality, enhancing the depth of linguistic analysis, and strengthening the interconnections among Turkic languages. In perspective, the expansion and automation of this database will remain one of the key directions for further scientific research.

REFERENCES

1. Abdurakhmonova, N. Computer Models of the Electronic Corpus of the Uzbek Language (Monograph). – Tashkent, 2021. – 202 p.
2. Ergasheva, G., Khaitqulov, Z., Kuchimova, N. Automating Translation in a Parallel Corpus (on the Example of Collocations). // Foreign Languages in Uzbekistan, 2023, No. 5 (52), pp. 162–173. https://doi.org/10.36078/1697008599
3. Raimjonov, Oybek Khalimjon O‘G‘LI (2024). The Relevance of Creating Linguistic Support for Phraseological Units. Oriental Renaissance: Innovative, Educational, Natural and Social Sciences, 4(6), 29–37.
4. Salomov, G. Introduction to Translation Theory (Textbook). – Tashkent, 1978. – 219 p.
5. https://qazcorpus.kz/
6. https://tugantel.tatar/
7. Uzbekcorpus.uz