ACADEMIA ONE
PEER REVIEWED JOURNALS

Progress Annals: Journal of Progressive Research
Volume 4, Issue 4, April 2026
ISSN (E): 2810-6466
Website: https://academiaone.org/index.php/8

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

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

Annotation: This article examines the principles of creating an electronic linguistic database of translation units in literary texts within the framework of Uzbek corpus linguistics. The study focuses on the methodological foundations of corpus creation, including segmentation, annotation, and alignment processes. Special attention is given to the development of Uzbek language corpora and their role in translation studies and computational linguistics. The research demonstrates that corpus-based approaches provide effective tools for identifying and systematizing translation units. The results highlight the importance of empirical data in improving translation quality and developing linguistic resources.

Keywords: Uzbek corpus linguistics; literary texts; translation units; parallel corpus; annotation; electronic linguistic database; alignment

Introduction

Corpus linguistics has become one of the key directions in modern Uzbek linguistics, especially since the 2010s, when theoretical research began to evolve into practical corpus creation projects. The development of the Uzbek National Corpus and educational corpora has significantly contributed to the digitalization of linguistic resources and the advancement of applied linguistics.

In Uzbek corpus linguistics, the creation of structured and annotated text databases is considered essential for linguistic analysis, translation studies, and natural language processing. Particular importance is attached to literary texts, as they reflect the richness, stylistic diversity, and cultural depth of the language[1].

The need to develop an electronic linguistic database of translation units in literary texts arises from the increasing demand for empirical, data-driven approaches in translation studies. Such databases enable systematic analysis of translation equivalence, variability, and transformation processes. The aim of this study is to analyze the principles of creating an electronic linguistic database of translation units in literary texts within the context of Uzbek corpus linguistics.

Methods

The present study is grounded in the corpus analysis method, which has become one of the leading empirical approaches in modern Uzbek computer linguistics. As noted by N.Abdurakhmonova, corpus linguistics enables the systematic investigation of language phenomena on the basis of large-scale, structured textual data, ensuring both objectivity and reproducibility in linguistic research. In this regard, the corpus-based approach provides a reliable methodological foundation for identifying and analyzing translation units in literary texts.

At the initial stage, particular attention is devoted to corpus design principles, including data selection, representativeness, and balance. According to N.Abdurakhmonova, the effectiveness of any linguistic corpus depends on the degree to which it reflects real language use across different genres and functional styles. In Uzbek corpus studies, the processes of text standardization and linguistic annotation are considered essential, as they ensure the scientific validity and internal consistency of the corpus[2]. Subsequently, the study applies a multi-level segmentation model. The corpus material is divided into paragraph, sentence, phrase, and

ACADEMIA ONE
PEER REVIEWED JOURNALS
Progress Annals: Journal of Progressive Research
Volume 4, Issue 4, April 2026
ISSN (E): 2810-6466
Website: https://academiaone.org/index.php/8

lexical units. This approach corresponds to the views of Shavkat Rahmatullayev, who emphasizes that meaning in language is realized not only at the sentence level but also through smaller structural units. Such hierarchical segmentation is particularly important in the analysis of literary texts, where semantic and stylistic nuances are often distributed across multiple levels of linguistic structure.

The next stage involves the implementation of annotation procedures. Uzbek linguistic corpora typically include morphological tagging, syntactic annotation, and metadata (such as author, genre, and historical period). Annotation allows researchers to transform raw textual data into a structured and searchable resource.

Furthermore, alignment techniques are employed to establish correspondences between source texts and their translations[3]. The study utilizes different types of alignment, including one-to-one (1:1), one-to-many (1: N), many-to-one (N:1), and many-to-many (N: N). In Uzbek translation studies, as discussed by B.Abduvohidov[4], such flexible alignment models are necessary to capture the variability and transformation processes inherent in literary translation.

Finally, the research incorporates statistical and concordance-based methods for analyzing translation units, their frequency, and contextual usage. Concordance analysis enables the observation of lexical units in real usage environments, which is crucial for identifying semantic relationships and translation equivalence. At the same time, statistical methods help to determine the frequency and stability of translation units within the corpus. In general, the integration of these methodological procedures—corpus design, multi-level segmentation, annotation, alignment, and statistical analysis—reflects the current trends in Uzbek computer linguistics and provides a scientifically grounded framework for the study and systematization of translation units in literary texts.

Results

The results of the study show that Uzbek corpus linguistics has made significant progress in recent years, particularly in the development of educational and national corpora. These corpora provide a foundation for analyzing literary texts and their translations. The application of corpus analysis enabled the identification of translation units at different linguistic levels. It was found that phrase-level and contextual units are more relevant for literary translation than sentence-level equivalents[5].

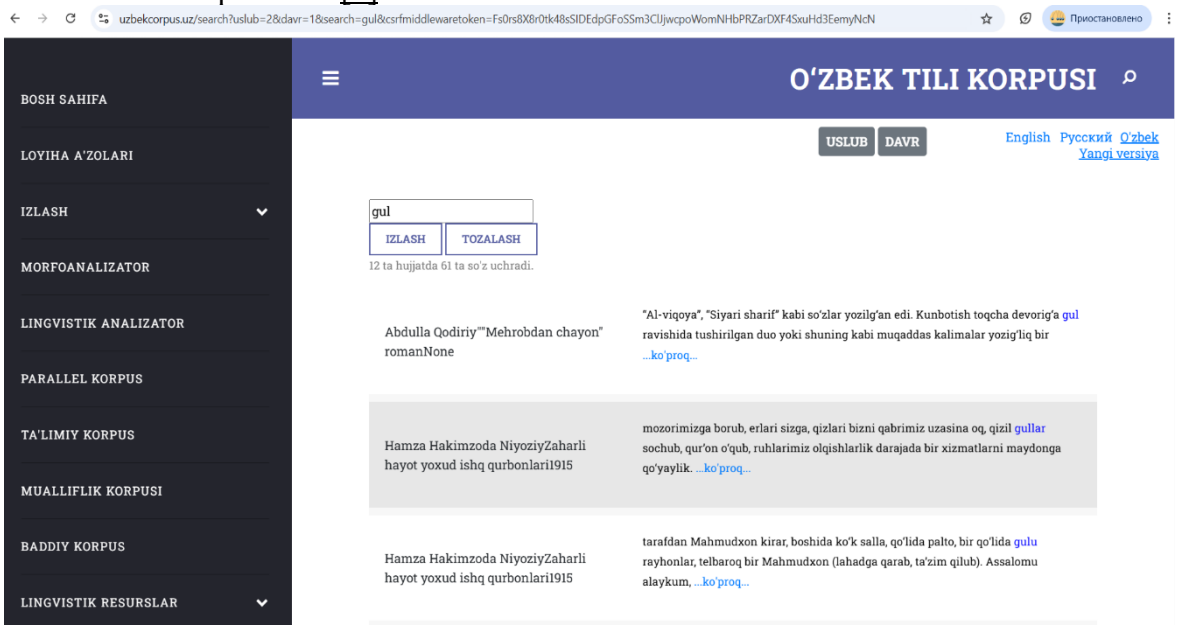

Figure 1. Uzbek corpus

The study also revealed that annotated corpora significantly improve the accuracy of linguistic analysis. The inclusion of morphological and syntactic tagging allows researchers to better

ACADEMIA ONE
PEER REVIEWED JOURNALS
Progress Annals: Journal of Progressive Research
Volume 4, Issue 4, April 2026
ISSN (E): 2810-6466
Website: https://academiaone.org/index.php/8

understand the structure and meaning of translation units. Furthermore, the use of concordance tools made it possible to analyze real usage patterns of words and phrases, which is essential for identifying translation equivalents.

The findings confirm that the creation of an electronic linguistic database facilitates:

- systematization of translation units;
- identification of translation strategies;
- improvement of machine translation systems.

Discussion

The development of Uzbek corpus linguistics is closely linked with global trends in computational linguistics and artificial intelligence. Although Uzbek is considered a low-resource language, recent efforts in corpus creation have significantly improved its digital representation. Research shows that Uzbek corpora are still in the process of expansion and standardization. Compared to large international corpora, Uzbek corpora face challenges such as limited data volume, insufficient annotation, and lack of parallel corpora[6]. The analysis of the Uzbek National Corpus demonstrates that Uzbek corpus linguistics is actively evolving in line with global trends, particularly in the creation of electronic linguistic resources for literary texts and translation studies[7]. The platform represents a modern, dynamically developing corpus system that integrates key principles of corpus design, annotation, and search functionality.

First, the Uzbek National Corpus reflects the principle of representativeness, as it incorporates texts from various functional styles, including literary works, journalistic materials, and scientific texts. Within this structure, literary texts occupy a significant place, providing valuable material for analyzing stylistic features and translation units[8]. However, compared to large-scale corpora such as the Russian National Corpus and the British National Corpus, the volume of Uzbek corpus data remains relatively limited, which affects the statistical robustness of certain analyses.

Second, the corpus demonstrates ongoing efforts in linguistic annotation, particularly morphological tagging. This aligns with the general тенденции in Uzbek computer linguistics, where the development of annotated corpora is considered a priority. At the same time, deeper levels of annotation—such as syntactic and semantic tagging—are still under development. This limitation somewhat constrains the detailed analysis of complex translation phenomena in literary texts.

Third, from the perspective of translation studies, the Uzbek National Corpus currently shows a need for more developed parallel and aligned subcorpora[9]. While some bilingual resources exist, a systematically aligned literary corpus (with sentence- and phrase-level alignment) is not yet fully realized. This contrasts with international practices, where parallel corpora play a central role in identifying translation equivalence and strategies.

Moreover, the corpus interface provides search and concordance tools, which enable users to analyze lexical units in context. These tools are particularly useful for examining translation units in literary texts, as they allow researchers to observe usage patterns and stylistic variation. Nevertheless, the analytical tools could be further enhanced by integrating advanced statistical modules and visualization features.

The Uzbek National Corpus illustrates both the achievements and challenges of Uzbek corpus linguistics. On the one hand, it establishes a solid foundation for the creation of electronic linguistic databases of literary translation units. On the other hand, its future development requires expanding corpus volume, improving multi-level annotation, and creating fully aligned parallel corpora[10]. Addressing these issues will significantly increase the corpus's potential for international-level research in translation studies and computational linguistics.

Progress Annals: Journal of Progressive Research
Volume 4, Issue 4, April 2026
ISSN (E): 2810-6466
Website: https://academiaone.org/index.php/8

However, studies indicate that the creation of educational and national corpora has laid the foundation for further development.

The inclusion of literary texts in Uzbek corpora is particularly important, as they provide rich material for studying translation processes, stylistic features, and semantic transformations[11]. Moreover, the integration of corpus-based methods into translation studies allows for a more objective and data-driven analysis, reducing reliance on subjective interpretation. Thus, the creation of an electronic linguistic database of translation units represents an important step toward the development of Uzbek computational linguistics and translation technologies.

Conclusion

In conclusion, this study demonstrates that corpus linguistics provides an effective methodological framework for analyzing translation units in literary texts within Uzbek linguistics. The research confirms that:

- multi-level segmentation improves the accuracy of translation unit identification
    - annotation enhances linguistic analysis
    - alignment models reflect the complexity of literary translation
    - The creation of an electronic linguistic database is essential for:
    - systematizing translation units
    - supporting translation studies
- developing artificial intelligence applications

Overall, the study contributes to the advancement of Uzbek corpus linguistics and highlights the importance of integrating modern technologies into linguistic research.

References:

1. Usmonova, M. “UZBEK CORPUS LINGUISTICS: AN ANALYSIS OF PRACTICAL RESULTS AND SCIENTIFIC ACHIEVEMENTS”. 2025. International Journal of Artificial Intelligence 5 (11): 2326-28. https://www.academicpublishers.org/journals/index.php/ijai/article/view/8266.
2. Abdurahmonova, N. (2023). Corpus Linguistics. GlobeEdit.
3. Sharipov, M. et al. (2022). Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language.
4. B.Abduvohidov. ON CORPUS LINGUISTICS AND THE NATIONAL CORPUS. (2023). Modern Science and Research, 2(12), 883-885. https://inlibrary.uz/index.php/science-research/article/view/27264
5. Suleymanov D., Gatiatullin A., Prokopyev N., Abdurakhmonova N.Z. Turkic Morpheme Web Portal as a Platform for Turkology Research / International Conference on Information Science and Communications Technologies ICISCT 2020 (Indexing Scopus) Applications, Trends and Opportunities 4th, 5th and 6th of November 2020, Tashkent Uzbekistan https://ieeexplore.ieee.org/document/9351500
6. Kadirova Z. O‘ZBEK TILI KORPUSINI YARATISHNING NAZARIY ASOSLARI. (2025). Multidisciplinary Journal of Science and Technology, 5(12), 1571-1573. https://www.mjstjournal.com/index.php/mjst/article/view/6404
7. https://uzbekcorpus.uz/
8. Abdurahmonova N. A two-level morphological analysis of the Uzbek corpus /МАТЕРИАЛЫ IV международного научного конгресса иностранная филология. социальная и национальная вариативность языка и литературы / 2019/4/1 Симферополь, -P. 425-431
9. Abdurakhmonova N. Modeling word combinations in terms of parts of speech in the process English-Uzbek machine translation / 5th international conference on Computer processing of Turkic languages “TurkLang. 2017” (18-21 октябрь)

ACADEMIA ONE
PEER REVIEWED JOURNALS
Progress Annals: Journal of Progressive Research
Volume 4, Issue 4, April 2026
ISSN (E): 2810-6466
Website: https://academiaone.org/index.php/8

10. Alessandro A, Usmanov T., Khamdamov U., Abdurakhmonova N., Mamasaidov M. UZWORDNET: A Lexical-Semantic Database for the Uzbek Language / 11th International Global Wordnet Conference (GWC2021) will be held from 18 to 21 January 2021 (Indexing Scopus)) https://www.aclweb.org/anthology/2021.gwc-1.2.pdf
11. Kadirova Z. O‘ZBEK TILI KORPUSINI YARATISHNING NAZARIY ASOSLARI. (2025). Multidisciplinary Journal of Science and Technology, 5(12), 1571-1573. https://www.mjstjournal.com/index.php/mjst/article/view/6404