Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed


# ON THE PRINCIPLES OF CREATING AN ELECTRONIC LINGUISTIC DATABASE OF TRANSLATION UNITS OF LITERARY TEXTS IN GLOBAL CORPUS LINGUISTICS

**A.R. Iskandarova**
National University of Uzbekistan
Doctor of Philosophy (PhD) in Philology, Associate Professor
iskandarova@gmail.com
**Dildora Mirvaliyeva**
Teacher at the National University of Uzbekistan
E-mail: dildoramirvalieva07@gmail.com

**Annotation:** This article examines the principles underlying the creation of an electronic linguistic database of translation units in literary texts within the framework of corpus linguistics. The study employs a corpus-based approach to identify, classify, and systematically organize translation units across multiple linguistic levels. Particular emphasis is placed on segmentation, alignment models, and the extraction of translation equivalents through the use of parallel corpora. Furthermore, the research analyzes international corpus practices, with specific reference to the British National Corpus and the Russian National Corpus, highlighting their structural and methodological characteristics. The findings demonstrate that corpus-based methods provide a reliable empirical foundation for enhancing translation analysis and for the development of electronic linguistic resources.
**Keywords:** corpus linguistics; literary translation; translation units; parallel corpus; alignment; electronic linguistic database; annotation.

## Introduction

Corpus linguistics represents one of the most rapidly developing fields of modern linguistics, facilitating the investigation of linguistic phenomena on the basis of empirical data. In particular, parallel corpora play a pivotal role in the scientific analysis of translation processes in literary texts.

In global practice, the development of electronic linguistic databases incorporating translation units derived from literary texts has gained increasing significance at the intersection of translation studies, computational linguistics, and artificial intelligence. Such databases enable the identification of translation strategies, the comparison of alternative translation variants, and the enhancement of machine translation systems.

The primary objective of this article is to provide a systematic scientific analysis of the principles underlying the construction of electronic linguistic databases of literary translation units within the framework of global corpus linguistics

## Methods

In this study, the corpus analysis approach was adopted as the primary research method. This method is regarded as one of the most effective empirical approaches for the identification of translation units in literary texts and their systematic organization within an electronic linguistic database.

Corpus analysis is a scientific methodology that enables the investigation of linguistic phenomena based on authentic language data, thereby allowing both statistical and contextual

Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed


examination. In the present research, multiple levels of corpus analysis were employed specifically for the identification of translation units in literary texts.

First, a parallel corpus was selected as the primary research material. A parallel corpus is defined as a structured collection of texts in which the source text and its translation are systematically aligned. In this process, the experience of well-established national corpora in English, Russian, and Turkish was examined, and their structural characteristics and operational principles were analyzed. Particular attention was given to the mechanisms of segmentation and alignment in literary texts.

Second, the levels of segmentation within corpus analysis were defined. Literary texts were divided into the following units: [1]

paragraph level;

sentence level;

phrase and clause level;

lexical unit (word) level.

This multi-level segmentation enables the precise identification of translation units. In literary texts, sentence-level equivalence is often insufficient; therefore, analysis at the phrase level and at the level of micro-units is of particular significance.

Third, the alignment process was implemented. At this stage, segments of the source text and their corresponding translated counterparts were systematically linked. The following types of alignment were identified: [2]

one-to-one (1:1) alignment;

one-to-many (1:N) alignment;

many-to-one (N:1) alignment;

many-to-many (N:N) alignment.

This approach allows for the representation of the flexibility and structural transformations that are typically observed in literary translation.

Fourth, the extraction of translation equivalents was conducted within the framework of corpus analysis. At this stage, the following linguistic phenomena were identified:

direct equivalents;

contextual (conditional) equivalents;

transformed units (including modulation, transposition, equivalence, etc.).

As a result, a typology of translation units was developed.

Fifth, elements of statistical analysis were applied. The frequency of translation units within the corpus, their contextual usage, and the degree of variation were examined. This, in turn, facilitates the identification of priority units for inclusion in the electronic linguistic database.

Additionally, the use of concordance and collocation tools was integrated into the corpus analysis process. These tools make it possible to observe the authentic usage environments of translation units and to conduct a more in-depth analysis of their semantic properties.

In general, the corpus analysis method in this study served as the primary scientific instrument for: [2]

identifying translation units;

systematizing them;

developing a structural model for an electronic linguistic database.

Thus, it provided a comprehensive methodological foundation for achieving the objectives of the research.

**Results**

Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed


The results of the study demonstrate that the application of corpus analysis to literary texts enables the systematic identification and classification of translation units across multiple linguistic levels. The analysis of parallel corpora revealed that translation equivalence in literary texts is highly variable and cannot be restricted solely to sentence-level correspondence.


**Figure 1. British National Corpus.**

The findings indicate that phrase-level and micro-level units play a crucial role in achieving semantic and stylistic equivalence. In many cases, one-to-many (1:N) and many-to-one (N:1) alignments occur more frequently than strict one-to-one correspondences, reflecting the inherent flexibility of literary translation. [4]

The examination of international corpus practices demonstrates that well-established corpora such as the British National Corpus, the Russian National Corpus, [5] and the Turkish National Corpus provide advanced models for the structuring and processing of linguistic data. These corpora employ sophisticated segmentation, annotation, and alignment techniques, which can be effectively adapted for the development of electronic databases of literary translation units.

Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

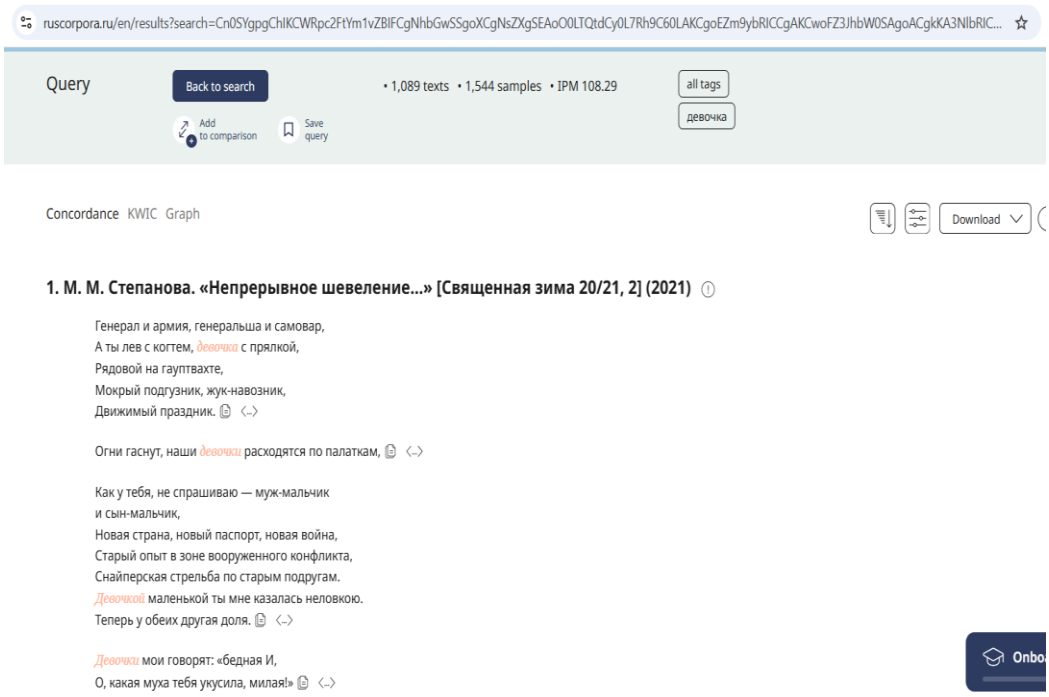




**Figure 2. Russian National Corpus.**

Furthermore, the extraction of translation equivalents revealed three dominant types: direct equivalents, contextual equivalents, and transformed units. Among these categories, contextual and transformed equivalents were particularly prominent in literary texts, underscoring the role of interpretation and creative transformation in the translation process. [6]

Statistical analysis confirmed that certain translation units demonstrate high frequency and relative stability across different texts, suggesting their potential inclusion as core components of an electronic linguistic database.

**Discussion**

The Russian National Corpus (RNC) is a large-scale, annotated linguistic database of the Russian language, comprising more than 2-billion-word forms. It was developed under the auspices of the Russian Academy of Sciences and has been publicly accessible since 2004. As a comprehensive and systematically organized resource, the RNC is widely recognized as one of the most authoritative corpora for the empirical study of the Russian language.

Structurally, the RNC is a multi-component system consisting of several interconnected subcorpora, including the main corpus, parallel corpora, a poetry corpus, a dialect corpus, and a multimedia corpus. Within this framework, literary texts occupy a significant and carefully balanced position. The main corpus incorporates a wide range of literary genres, including modern fiction, drama, memoirs, and biographical texts. Importantly, literary materials constitute no more than 40% of the total corpus volume, thereby ensuring the representativeness and balance of the dataset.

A particularly valuable component of the Russian National Corpus (RNC) is its specialized literary subcorpus known as “Russian Classics.” This collection comprises more than 35,000 texts, with an overall volume of approximately 26-million-word forms. It includes works by prominent authors such as Leo Tolstoy, Fyodor Dostoevsky, and Anton Chekhov, among

Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed


others. Chronologically, this subcorpus covers literary works from the 18th and 19th centuries and functions as a normative model of the Russian literary language. [6]

From a linguistic perspective, the RNC provides advanced analytical capabilities. The corpus is enriched with morphological and syntactic annotation, semantic tagging, and comprehensive metadata, including information on authorship, genre, and historical period. In addition, the availability of parallel corpora within the RNC significantly enhances its value for translation studies.

Overall, these features make the Russian National Corpus an essential resource for the study of literary texts and their translation, offering a robust empirical foundation for linguistic, stylistic, and comparative analysis.

When compared with the British National Corpus (BNC), several key differences become evident. In terms of size, the RNC significantly exceeds the BNC, containing over 2 billion tokens, whereas the BNC comprises approximately 100 million words. Regarding literary texts, the RNC includes both a dedicated literary subcorpus and literary materials integrated within its main corpus, while the BNC incorporates literary texts as part of a balanced overall corpus structure.

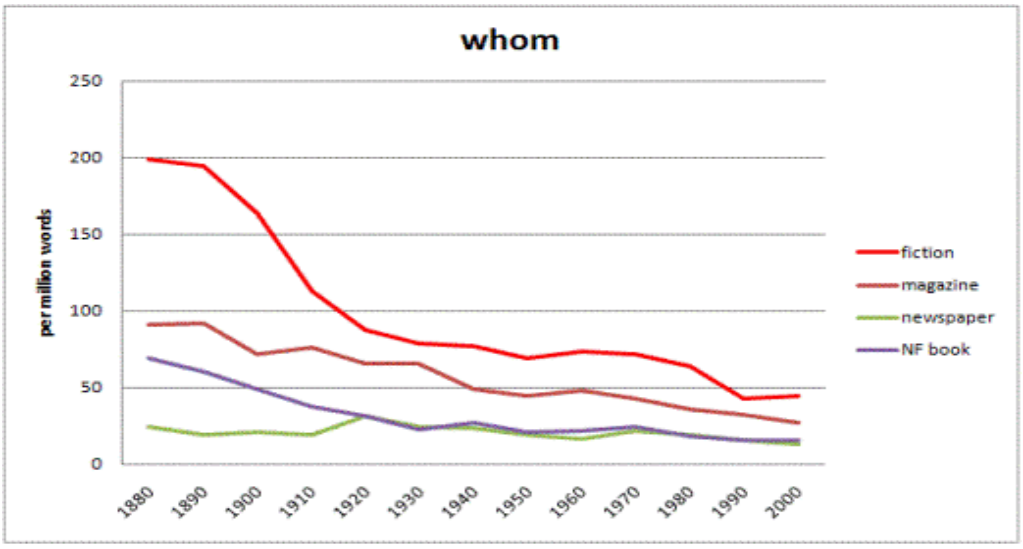
In terms of annotation, the RNC provides a more complex system, including morphological, syntactic, and semantic tagging, whereas the BNC primarily focuses on grammatical annotation. Another important distinction concerns the availability of parallel corpora: the RNC incorporates parallel data, whereas such resources are relatively limited in the BNC. [7] Finally, the historical scope also differs, as the RNC spans from the eighteenth century to the present, while the BNC predominantly represents late twentieth-century English usage.

In conclusion, the Russian National Corpus provides richly and deeply annotated linguistic data that is particularly valuable for translation studies and the stylistic analysis of literary texts. In contrast, the British National Corpus emphasizes representativeness and balance, offering a comprehensive model of general English usage. Both corpora are regarded as foundational resources in modern corpus linguistics and serve as essential tools for empirical linguistic research.

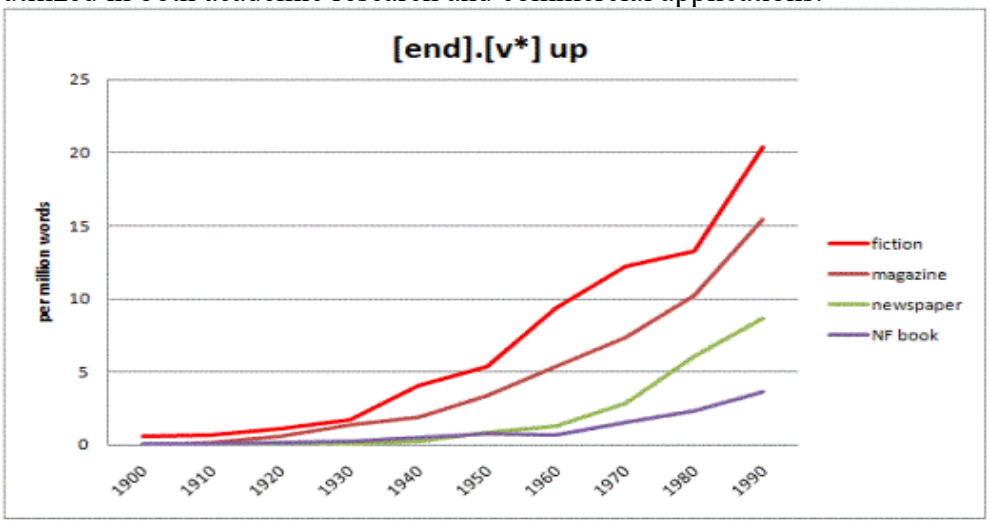
The British National Corpus (BNC) is a large, balanced corpus of the English language comprising approximately 100 million words. It includes both written and spoken language data, thereby constituting a representative resource for the study of authentic language use in real communicative contexts. The corpus was developed between 1991 and 1994 with the primary objective of enabling computational and empirical analysis of contemporary English. [3]

One of the defining features of the BNC is its balanced composition. It encompasses a wide range of genres, including fiction, journalism, and academic writing, as well as different modes of communication, such as spoken and written discourse. [3] Within this structure, literary texts constitute a significant component of the corpus. These include novels, short stories, and dramatic works, all of which contribute to the overall diversity and representativeness of the dataset.

Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed


From a scientific perspective, the BNC is characterized by several important features. First, it ensures a high degree of representativeness, reflecting the state of the English language in the late twentieth century. Second, the corpus is linguistically annotated, primarily through part-of-speech (grammatical) tagging and elements of syntactic analysis. Third, it maintains a balance between literary and non-literary texts, which enables more reliable comparative studies across genres. [1] Finally, the BNC is widely accessible and has been extensively utilized in both academic research and commercial applications.




**Conclusion**

In conclusion, this study confirms that corpus analysis constitutes a highly effective methodological framework for investigating translation units in literary texts and for constructing electronic linguistic databases. The integration of multi-level segmentation, flexible alignment models, and statistical analysis provides a comprehensive approach to capturing the complexity inherent in literary translation.

The analysis of international corpus practices demonstrates that existing large-scale corpora offer valuable methodological insights that can be adapted to the study of translation units. In particular, their approaches to annotation, segmentation, and data structuring serve as important reference models. The study also highlights that translation equivalence in literary texts is

Western European Journal of Linguistics and Education
Volume 4, Issue 4, April 2026
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed


predominantly context-dependent and frequently involves various transformation processes. Therefore, any electronic database of translation units must account for variability, multiplicity, and contextual factors. Overall, the proposed corpus-based approach contributes to the advancement of translation studies, computational linguistics, and artificial intelligence by providing a structured and empirically grounded model for the analysis and systematization of literary translation units.

## REFERENCES

1. Apresjan, Ju.; Boguslavsky, I.; Iomdin, B.; Iomdin, L.; Sannikov, A.; Sizov, V. (2006). A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. Proceedings of LREC. Genova, Italy. pp. 1378– 1381.
2. "Bilingual dictionaries to promote India's mother tongues". Times of Oman. 14 March 2012. Archived from the original on 2010-12-31. Retrieved 17 March 2012.
3. Burnard, Lou; Aston, Guy (1998). The BNC handbook: exploring the British National Corpus. Edinburgh: Edinburgh University Press. p. xiii. ISBN 0-7486-1055-3.
4. Hoffman, Sebastian; Lehmann, Hans Martin (2000). "Collocational Evidence from the British National Corpus". In Kirk, John M. (ed.). Corpora Galore: Analyses and Techniques in Describing English. Amsterdam: Rodopi. ISBN 9789042004191.
5. https://ruscorpora.ru/en
6. Korhonen, Anna (2002). "EVALUATION RESOURCES for English Subcategorization Acquisition Systems". Archived from the original on 2012-12-13. Retrieved 18 March 2012.
7. What is the BNC?. Retrieved 12 March 2012.