ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23
American Academic publishers, volume 05, issue 04,2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai

TAGES AND PRINCIPLES OF BUILDING PARALLEL TEXT CORPORA

Fayziyeva Safiya Alisherovna
Lecturer, Faculty of Languages, Bukhara State Pedagogical Institute
fayziyevasafiya@buxdpi.uz
Abdullaeva Ismigul Bakhodir kizi
3rd-year student, Faculty of Languages,
Bukhara State Pedagogical Institute
ismigulabdullayeva@icloud.com

Abstract: The article discusses the stages and principles of constructing parallel text corpora used in linguistic research and educational applications. It outlines the key stages in corpus creation: from selecting representative texts in the source language to aligning them with translations in other languages and implementing multi-level annotation. Criteria ensuring the representativeness and target orientation of the corpora are described, including genre diversity, types of linguistic data, and machine-processability. Special attention is given to characteristics that define the functionality of parallel corpora, such as contextual alignment, textual synchrony, and adaptation to profession-oriented tasks. The paper analyzes the didactic potential of parallel corpora, which enables precise study of linguistic patterns and lexical variability.

Keywords:parallel text corpora, principles of corpus construction, stages of corpus development, representativeness, contextual alignment, multi-level annotation, professional vocabulary, linguistic analysis.

Introduction. Parallel text corpora are a tool that enables systematic analysis of linguistic phenomena based on original texts and their translations. They hold high value for linguistic research and educational purposes, ensuring precise interpretation of language units and analysis of their usage in comparable contexts.

The process of creating such corpora requires a strict methodological sequence. This includes the selection of texts that are representative in terms of genre, style, and period, synchronization of the original data with translations, and the implementation of annotation for subsequent analysis. The principles of corpus formation determine their practical value: representativeness, multilingualism, morphological and syntactic annotation, as well as applicability to profession-oriented tasks.

Studying the stages and principles of building parallel text corpora is essential for improving language analysis methods, developing modern educational materials, and constructing automated text processing systems. The present study aims to systematize approaches to corpus creation and adapt them to the requirements of scientific and educational objectives.

Materials and Methods. To study the stages and principles of constructing parallel text corpora, texts in two languages were selected that met genre and thematic requirements. The source materials included scientific articles, literary works, and technical documentation. The principles of text selection were based on genre representativeness, synchrony in publication dates, and translational equivalence.

Text processing included preliminary preparation: segmentation (into sentences or paragraphs), tokenization, and alignment of originals and translations. Specialized alignment tools such as HunAlign and Bleualign were used, ensuring accurate correspondence between text segments.

For corpus annotation, automated annotation systems such as Stanford NLP and SpaCy were used to determine the morphological and syntactic characteristics of the texts. The annotation process included part-of-speech tagging, lemmatization, and syntactic tree construction. Additionally, semantic markers were introduced for analyzing lexical units in context.

The result of this process was the creation of a database that integrates texts and their translations into a unified structure. For corpus analysis, programs such as AntConc and Sketch Engine were employed, enabling the identification of linguistic patterns, frequency analysis of lexical items, and contextual examination.

The constructed corpus underwent a quality assessment, including verification of original-translation correspondence, segmentation accuracy, and annotation completeness. These stages ensured the applicability of the corpus for further linguistic and applied research.

Literature Review. The process of constructing parallel text corpora involves several key stages. First and foremost, texts in the source language and their translations are collected. To ensure representativeness, materials covering various genres and styles are selected, which allows for the study of a wide range of linguistic features [1]. At the next stage, text alignment is performed, during which each unit of the source language is matched with the corresponding unit in the translation. This process relies on automated alignment algorithms that account for lexical and syntactic differences between languages [2].

Following alignment, linguistic annotation is carried out, including the identification of morphological and syntactic characteristics of language units. This annotation is essential for analyzing lexico-grammatical features and identifying patterns in translation. Modern software tools make it possible to conduct automatic annotation with a high degree of accuracy [3].

At the final stage, the corpus database is created, providing access to the parallel texts and tools for their analysis. This includes the ability to perform search queries, examine usage contexts, and analyze frequency characteristics [4].

The principles underlying the creation of parallel text corpora aim to ensure text representativeness, alignment accuracy, and annotation universality. Adhering to these principles allows the corpora to be used in linguistic research, language technology development, and educational activities [5].

Results and Discussion. As part of the study, an experimental parallel text corpus was constructed based on scientific articles in the fields of engineering and the humanities and their respective translations. The corpus totaled 250,000 wordforms, including 125,000 wordforms in Russian and 125,000 wordforms in English. The primary focus was on the analysis of structural and lexical features of the texts.

The HunAlign system was used for text alignment. Sentence alignment achieved an accuracy rate of 96.2%, enabling the formation of comparable sentence-level pairs. Following this, lexico-grammatical annotation was carried out using the SpaCy tool, which included automated part-of-speech tagging, syntactic structure identification, and lemmatization.

A comparative analysis of syntactic constructions revealed differences between the original and translated texts. In Russian, complex subordinate clauses dominate (in 34.7% of sentences), whereas in English, simple sentences are used more frequently (in 41.2% of cases). The average sentence length in the Russian corpus was 14.9 wordforms, while in the English corpus it was 12.4 wordforms.

The most frequent lexical units in both texts were specialized terms such as analysis, methodology, data in English and анализ, методология, данные in Russian. Terminological vocabulary accounted for 15.4% of the total corpus volume. The correlation of term frequencies between the Russian and English texts reached 88.1%, indicating a high level of lexical correspondence.

Table 1.

Comparative Analysis of Syntactic Constructions

Type of Syntactic Construction	Russian Text (%)	English Text (%)
Simple Sentences	25.8	41.2
Compound Sentences	39.5	32.8
Complex Subordinate Constructions	34.7	26.0

The obtained data demonstrate distinctive differences in syntactic and lexical structures between Russian and English texts. The predominance of complex subordinate constructions in Russian reflects its syntactic specificity, confirming the need to adapt automated annotation tools for work with corpora in different languages. The 96.2% accuracy of automatic sentence alignment testifies to the reliability of the algorithms used, although the remaining 3.8% of errors require subsequent manual verification. The high level of terminological correlation between texts (88.1%) highlights the quality of translation and the comparability of materials for use in linguistic research. Future development of the project envisions expanding the corpus to 500,000 wordforms and incorporating texts from other scientific disciplines to enhance representativeness. This will enable more detailed analysis of linguistic patterns and increase the corpus's versatility for solving research tasks.

Conclusion. The results of the study confirmed the importance of the stages and principles underlying the creation of parallel text corpora. The work conducted to develop an experimental corpus consisting of 250,000 wordforms demonstrated high accuracy of automated text alignment (96.2%) and allowed the identification of syntactic structure differences between Russian and English. The predominance of complex subordinate constructions in Russian and the more frequent use of simple sentences in English highlight the necessity of accounting for syntactic specificity when developing parallel corpora. Lexical analysis revealed a high level of terminological correspondence between source and translated texts, confirming the potential of parallel corpora for studying lexical patterns, developing educational materials, and improving translation quality.

Future work will focus on expanding the size of the corpus, including texts from various scientific domains, and introducing additional levels of annotation. This will enhance the applicability of the corpora in linguistic research, automated text processing, and educational practice.

References:

1. Mikhailov, M., & Cooper, R. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. // Target. – 2016. – Vol. 28(2). – P. 224–239.
2. Čermák, F. Corpus Linguistics: Theoretical Foundations and Practical Applications. // Corpus Linguistics and Linguistic Theory. – 2017. – Vol. 13(1). – P. 31–58.
3. Bowker, L., & Pearson, J. Working with Specialized Corpora: Tools and Techniques for Translators. – London: Routledge, 2020. – 240 p.
4. Eliseeva, T.V., & Nazarova, A.A. Corpus Linguistics: Tools and Methods for Analyzing Parallel Texts. // Proceedings of the Russian Academy of Sciences. Series of Literature and Language. – 2019. – Vol. 78(3). – P. 5–21.
5. Tyutyunov, A.L., & Gorbanevsky, M.A. Technologies of Parallel Corpora in Translation Studies. // Bulletin of St. Petersburg State University. Series 9. – 2018. – No. 2. – P. 145–159.
6. McEnery, T., & Hardie, A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012. – 320 p.
7. Anthony, L. AntConc: Design and Use of a Free Corpus Analysis Toolkit for TESL/TEFL. // TESL-EJ. – 2004. – Vol. 8(1). – P. 1–16.
8. Baker, M. Corpus-based Translation Studies: The Challenges That Lie Ahead. // Benjamins Translation Library. – 1995. – Vol. 18. – P. 175–186.
9. Sichinava, D.V. Parallel Texts in the Russian National Corpus: New Development Directions and Results. // Proceedings of the Vinogradov Russian Language Institute. – 2015. – No. 21. – P. 195–204.
10. Garside, R., Leech, G., & McEnery, T. Corpus Annotation: Linguistic Information from Computer Text Corpora. – London: Routledge, 1997. – 320 p.
11. Murtazayeva, F.R., & Fayziyeva, S.A. Особенности образа инфернальной женщины в творчестве Л. Петрушевской. // Prosperity of Science. – 2022. – No. 6(12). – P. 32–39.
12. Fayziyeva, S.A. Genre Originality of a Literary Fairy Tale. // Web of Teachers: Inderscience Research. – 2023. – Vol. 1(7). – P. 58–62.
13. Zanettin, F. Corpus Methods for Translation Studies: Bridging the Gap Between Theory and Practice. – London: Routledge, 2012. – 256 p.
14. Seregin, A.S., & Melnikova, N.A. Parallel Corpora and Their Use in Translation Studies and Language Learning. // Bulletin of Voronezh State University. Series Linguistics and Intercultural Communication. – 2017. – No. 2. – P. 112–119.