European International Journal of Multidisciplinary Research and Management Studies (ISSN: 2750-8587)

Received: 20 April 2026 | Accepted: 24 May 2026 | Published: 09 June 2026

DOI: 10.55640/eijmrms-06-06-07

RESEARCH ARTICLE

The Value of Text-Based Parallel Corpus as A Material for Comparative, Contrastive, And Translation Studies

Dilafruz Muhammadiyeva

Associate Professor at Toshkent State University of Uzbek Language and Literature, DSc, Uzbekistan

VOLUME: Vol.06 Issue06 2026

PAGE: 34-42

Abstract

A corpus is a collection of texts subjected to a search program for the purpose of identifying characteristics of language units, stored in electronic form in natural language, consisting of written or spoken materials placed on the basis of software with a computerized search system. Corpora are considered important sources for studying and analyzing the natural properties of language. They help to more deeply understand the vocabulary, grammar, phraseology, and semantics of a language. Corpora are crucial in linguistics, particularly in Natural Language Processing (NLP), lexicography, grammatical analysis, and studying the historical development of language.

KEYWORDS

Parallel corpus, corpus text linguistics, bitext, authorial corpora, translation quality, alignment, stylistic equivalence, adaptation.

INTRODUCTION

In world linguistics, targeted research in the field of corpora was initiated in the 1940s by Bloomfield, Fries, and Bondgers; N. Francis and G. Kučera first developed the principles of corpus construction. In Russian linguistics, V.P. Zakharov, A.B. Kutuzov, Y.V. Nedoshivina, V.V. Rikov, and V.A. Plungyan conducted research on corpora, their types, and the principles of corpus construction and tagging. When discussing the history of corpora, the Brown Corpus, created between 1961–1964, is mentioned first. This corpus was created at Brown University and contains 500 text fragments, each consisting of 2000 words [Sultanova, 2024, 44].

Corpus linguistics is a field concerning texts that contain large volumes of speech material, and it also studies issues of database formation, text processing, analysis and synthesis, razmetkalash (annotation) and tagging. Corpora enable the analysis of natural language phenomena and the observation of linguistic processes [Sultanova, 2024, 44].

Along with the concept of "corpus," the term "text corpus" is also used. A text corpus is a unified whole that can consist of phonemes, graphemes, morphemes, lexemes, sentences, and texts stored electronically. Corpora are essentially collections formed as databases to solve linguistics problems and serve as material for conducting research in various fields [Xolmanova, 2021, 55].

Corpus linguistics is quite developed worldwide, and today the following can be listed as actively functioning corpora: the Russian National Corpus, various authorial corpora, the English Corpus (Bank of English), the British National Corpus, the Hungarian National Corpus, the Spanish Corpus, the Latin Corpus, the CORIS/CODIS corpus of contemporary Italian, the modern Chinese Corpus, the Mannheim German Corpus (Institut für Deutsche Sprache, Mannheim, Germany), and corpora of Turkic, Tatar, Kazakh, Bashkir, Shor, and Uzbek languages.

From the perspective of expression language, corpora are divided into two groups: monolingual text corpora and parallel text corpora. All types of corpora belong to monolingual text corpora. Such corpora are subdivided into smaller groups: corpora comparing texts of various genres, corpora reflecting problems, titled corpora, research corpora, illustrative corpora, domain-specific corpora, authorial corpora, journalistic text corpora, and lingvodidactic (educational) corpora. Parallel text corpora can be bilingual (Uzbek-English, Russian-English) or multilingual (Russian-English-French) [Рахманова, 2021, 27].

Corpus linguistics is a modern direction oriented toward studying language based on real texts. This approach proposes studying language not through artificial examples but through real, contextual examples. A parallel corpus is a multilingual corpus in which the text of one language and its translation (or translations) are mutually aligned.

Parallel corpora are also crucial for linguistics and computer linguistics. Through them, it is possible to identify relationships between languages, study translation processes, develop Natural Language Processing (NLP) technologies, and conduct deep analyses in many other scientific fields. Parallel corpora are collections of texts containing identical or similar information in several languages. These corpora are primarily used in creating translation systems, language learning, conducting linguistic research, and automatic text analysis.

"Parallel corpus (English: parallel corpora) – the electronic analog of translation texts; wherein the original text and its one or several translations are systematically collected. Electronic texts in the corpus may be the original text or a part of it."

Another source provides the following definition of parallel corpus: "Parallel corpus – a corpus containing the original text and its translation."

D.O. Dobrovolskiy, Y. Tao, V. Zakharov, A.A. Kokoreva, and Y.P. Sosina conducted special research on the structure, composition, and capabilities of parallel corpora. D. Dobrovolskiy, a scholar who engaged in the theory and practice of parallel corpus construction, defines parallel corpus as follows:

"Parallel corpus – a corpus consisting of a collection of electronic texts in the original and translation. The original and translation texts are not merely placed side by side; rather, sentences (syntactic units) in both texts are divided into segments at the level of semantic correspondence, and these units stand side by side, referring and linking to each other. The fragment in the translation corresponding to the original text fragment is marked. Exactly this condition creates the possibility for performing various linguistic operations with such corpora" [Жўраева, 2022, 43].

The first parallel corpora worldwide emerged primarily for Natural Language Processing, machine translation, and linguistic research purposes. Parallel corpora are collections of mutually corresponding texts that help observe translation and language learning processes across different languages. These corpora serve not only to identify similarities and differences between languages but also to create machine translation systems and Natural Language Processing tools. The first parallel corpus emerged in the 1950–1960-years for developing machine translation systems. During that period, attempts began to automate translation from English to Russian or from Russian to English. For example, Canada's Texts and Translations (Canadian Hansard) parallel corpus was developed to test automatic translation systems by collecting Canadian parliamentary statements in English and French parallelly.

The first parallel corpora emerged for developing machine translation systems, linguistic research, comparative linguistics, and studying intercultural differences. These corpora were used not only by linguists but also served significant assistance to developers of automatic translation systems, multilingual technologies, and Natural Language Processing systems.

The corpus-based approach in linguistics is considered an integral part of modern linguistic analysis. However, this methodology is not absolutely new for the Uzbek language, literature, and culture. Historical sources indicate that samples possessing corpus characteristics had already emerged and were applied in practice during the formation and development of the Uzbek language. Although these corpora were not in electronic form as understood today, they approach modern language corpora in terms of content and purpose. Based on Z. Xolmanova's views [Xolmanova, 2021, 56], it can be stated that Mahmud Kashgari's work "Devonu lug'otit turk," created in the 11th century, is an important source of ancient Turkic language lexicology and textual studies, containing Turkic words along with their contextual

usage, text fragments, wise sayings, and samples of folk oral art. This work can be regarded as one of the initial examples of corpus linguistics in terms of content and functionality.

Looking at the history of literature, the "Khamsa" collections created by representatives of Persian-Tajik classical literature—Nizomi Ganjavi, Khusrav Dehlavi, and Abdurahmon Jomi—along with Alisher Navoiy's "Khamsa" works created in Turkic language constitute a vivid example of a parallel text corpus. Each dastan (epic) independently forms a text corpus possessing thematic and stylistic integrity. For instance, Alisher Navoiy's "Saddi Iskandariy" dastan consists of 7,315 couplets and can be studied as a large-scale text corpus. Furthermore, the phenomenon of intertextuality, namely, intertextual connectivity, is widely encountered in classical literary heritage. This condition was expressed in our ancient literary traditions through the "art of patterning." The inclusion of "story within story" or additional texts within the text that explain and expand the main reality is a manifestation of this technique. Today, this approach is explained in Western literary studies through the concepts of "intertextuality" and "precedent units" [Xolmanova, 2021, 57]. In fact, this phenomenon formed much earlier in Eastern literary heritage and was widely applied in practice.

The miniatures appended to Zahiriddin Muhammad Babur's "Baburnoma" work, as well as tables and drawings attached to other historical works, are examples of illustrative corpora providing visual interpretation to literary texts. Along with this, Alisher Navoiy's works "Majolis un-nafois" and "Nasoyim ul-muhabbat," forming an important part of his heritage, possess great scientific significance as an authorial corpus. Based on these historical and literary sources, it is possible to create a perfect classical text corpus by preparing electronic versions of the works, digitizing them, and forming a database. This will serve as a rich empirical basis not only for linguistics but also for literary studies, history, and cultural studies.

The first researches on parallel corpus construction in Uzbek linguistics are observed in the works of B. Mengliyev, R. Karimov, and Sh. Musulmonkulova.

Regardless of whatever linguistic or translation purpose parallel corpora pursue, alignment between the texts they contain is required. Such alignment is typically accomplished through special software technologies in bitext format. Bimatn (bitext) – texts containing the original language text and its translation in another language(s). Parallel corpora - partekst - bimatn - bitekst. The idea of creating bitext belongs to Brian Harris, who wrote research on the concept of bitext creation in 1988. Later, it was developed by a group of scholars and named RALI (Practical Research in Computational Linguistics). Bitexts are accomplished through special computer programs named "alignment tool" or "bitext tool." These programs align the original and translation text content in various syntactic units, primarily in the form of simple sentences. The collection of bitexts is named bitext database or bilingual corpus and serves as a database (справочник) enabling observation of various connections.

Aligning text with bitext involves establishing interconnection and ensuring compatibility of text units (at the level of sentences, paragraphs, or sections), thereby expanding the possibility of using the corpus for systematic analysis.

Sh. Xamroyeva, classifying linguistic corpora according to their structure and functional characteristics, divides them based on the degree of parallelicity into: monolingual, bilingual, and multilingual corpora [Хамроева, 2018, 29]. This classification serves the formation of a consistent approach in corpus linguistics. Specifically, according to the researcher's opinion, internal corpora are also formed within national corpora, and they are directed toward a specific research purpose. For example, through a corpus of literary works, characteristics of the literary language of a certain period or the linguistic style of a particular author can be analyzed. The parallel corpus, meanwhile, is a convenient tool for revealing intertextual correspondence based on translation, studying work texts from a comparative-linguistic perspective, and analyzing the semantic and stylistic features of translation. Therefore, parallel corpora are important not only for translation studies but also for stylistics, textual studies, historical grammar, and lexicography.

In contemporary linguistics, text-based parallel corpora create unparalleled scientific opportunities for translation studies, contrastive, and comparative linguistics. They serve to observe the usage of linguistic units in different languages in real context, identify how language structure and semantics change in translation, and systematically analyze how translation strategies are implemented.

In modern linguistic research, parallel corpora are widely used as an effective source for various practical purposes. In linguistics, parallel corpora enable comparison of the structure, function, and frequency of linguistic units in two (or

more) language systems. Through such corpora, scientifically grounded conclusions can be drawn regarding language structure, grammatical models, the scope of application of lexical units, and their contextual variants. Parallel corpora are particularly important in identifying formal and semantic similarities and differences between the source language (original text) and target language (translation text). This allows analysis to be conducted at various levels. These levels are typically classified as follows: Micro level (word, morpheme, word combinations), meso level or syntactic level (sentence and paragraph – this stage is also sometimes referred to as the meso level [Мельник, 2024]), and macro level (entire text, discourse, and genre). Based on the above arguments, it can be stated that parallel corpora serve systematic text-based analysis and cover linguistic phenomena at various levels.

In translation studies, using parallel corpora enables effective contrastive analysis of two or more languages. They assist translators in identifying translation equivalents between source and target languages, particularly in finding appropriate contextual solutions for units without direct counterparts. Furthermore, parallel corpora provide definitive data on the frequency of usage of words, combinations, phraseological units, metaphorical expressions, and stylistic transfers within both language systems. This empirical basis serves as a real tool for translators in developing context-appropriate translation strategies.

In recent years, the scope of using parallel corpora has further expanded, becoming one of the primary sources in developing automatic translation systems, training machine translation based on neural networks, and forming training bases for artificial intelligence applications. This is transforming parallel corpora into a resource of strategic significance not only for traditional linguistics and translation studies but also for modern directions such as computer linguistics, Natural Language Processing, and language technologies [Musulmankulova, 2024, 35].

Parallel corpora possess important scientific and practical significance in the fields of language learning, translation systems, and linguistic research. They enable studying multiple languages, analyzing translations, and examining various language structures. The scientific-theoretical foundations of parallel corpora, their analyzed components, and widespread application are necessary for achieving significant accomplishments in linguistics. Furthermore, the role of parallel corpora in enhancing the effectiveness of computer linguistics and automatic translation systems is crucial.

Parallel corpora encompass several scientific-theoretical foundations in linguistics and linguistic research. Among them are language learning, translation research, syntax and semantics, morphology, and other fields. Parallel corpus, in turn, possesses the following advantages:

1. Creates numerous conveniences in language learning and translation. Parallel corpora, especially, play an important role in creating automatic translation systems and language learning programs. They are used in creating translation systems and enhancing their effectiveness. By analyzing similarities in source and translation texts within the corpus, automatic translation systems are improved. Today, many popular translation systems (Google Translate, DeepL) operate based on parallel corpora. Parallel corpora also assist in language learning processes, particularly in studying language rules and vocabularies through mutual translation. This is especially an effective method for those learning a new language. Through the corpus, peculiarities of grammar and syntax are demonstrated.

2. Parallel corpora are an important source for linguistic research. They enable analyzing grammatical differences between languages. For example, through a parallel corpus, morphological, syntactic, and semantic differences in languages can be studied. Along with this, parallel texts assist in understanding linguistic units and their changes in translation. Through the corpus, similarities and differences between translations can be identified, and semantic correspondence can be verified. Through parallel corpora, syntactic and semantic relationships between languages are studied. Using corpora, words and phrases in languages can be analyzed in context.

3. Semantic analysis is conducted through parallel corpus. When semantic analysis is performed through a parallel corpus, meaning relationships between texts and translation accuracy are examined. Semantic analysis creates the possibility of studying meaning changes and word nuances in language. This process is particularly necessary for improving the quality of translation systems. For example, by translating semantic peculiarities in existing texts into other languages, the database can be expanded.

4. Parallel corpora are also the primary tool in corpus linguistics. Corpus linguistics is a field directed toward using large volumes of texts to identify and analyze various characteristics of language, and parallel corpora enable performing this process across multiple languages.

In today's conditions of globalization, along with multilingualism, intercultural communication, and the expansion of translation activities, the role of empirical bases in scientific research is increasing. Particularly, corpus linguistics, computer linguistics, and artificial intelligence fields are creating great opportunities in this regard. Parallel corpora are an integral part of this process, bringing translation studies to a new stage. Through them:

- translation laws are revealed;
- new methods are developed in comparative linguistics;
- precise, fact-based conclusions are obtained in linguistic research.

Parallel corpora reveal the characteristics of mutual correspondence or difference of grammatical and syntactic structures for comparative linguistics. Through this, the possibility arises to identify phenomena such as grammatical parallelicity between two related languages, structural changes, lexical variants, stylistic changes, and functional equivalence. These differences identified through corpora allow deriving generalizations based on clear, empirical evidence regarding theoretical language models and translation types.

In contrastive linguistics, parallel corpora enable contextual identification of semantic, syntactic, and stylistic differences emerging between two or more languages. Analyzing how linguistic units are used in real texts, how they undergo transformation in translation, and the reasons for these changes enriches the main approaches of contrastive methodology. Particularly, for research conducted from the perspective of cultural differences, pragmatic markers, and stylistic adaptations, parallel corpora create an important empirical field.

In translation studies, parallel corpora serve as a reliable foundation for filling translation theory with practical examples, empirically confirming categories of translation strategies, and investigating translation quality, alignment, stylistic equivalence, and compensation phenomena. Particularly, in analyzing stylistic and emotional differences of translation, real context is important, and in this process, the contribution of parallel corpora is especially noteworthy.

Furthermore, along with the development of automatic translation systems based on artificial intelligence, parallel corpora are also used as a primary linguistic resource in training, testing, and analyzing machine translation models. This enables not only linking corpus linguistics with traditional scientific research but also integrating it with technological innovations, Natural Language Processing, and artificial intelligence fields.

True parallel corpora (translation corpora) – corpora containing numerous original texts written in a certain language and their translations into one or several other languages.

In the preparation of true parallel corpora and the development of programs for their processing, the problem of alignment arises – establishing correspondence between fragments of the original text and translation texts. To solve this problem, various methods of automatic text correction are applied: by sentences, by clauses (grammatical structures), by word combinations, and by words.

Comparable corpora (comparable corpora) – corpora whose texts are united based on certain criteria. In such corpora, for example, texts within a certain subject matter in Uzbek and English languages are reflected (the same thematic field, dialects of one language, regional variants, for example, Uzbek as a native language and Uzbek as a foreign language, and others).

Parallel corpora differ from comparative (comparative) corpora in several aspects. Parallel corpora encompass translations of one text in various languages and enable studying translation correspondence and analyzing semantic equivalence. In comparable corpora, however, texts that differ in language and style but are created within the same subject matter are collected. That is, in this type of corpus, how texts in various languages related to a certain subject are reflected in a certain discursive context is analyzed. Comparable corpora are primarily created for performing pragmatic and cognitive analysis and identifying how certain social-cultural reality is expressed through various language tools. From this perspective, such corpora form an important research base for

the contrastive and typological directions of linguistics. Russian researcher V.I. Melnik interprets parallel corpus as a collection of corresponding texts based on translation, while explaining comparable corpus as a tool for analyzing texts created in various languages but on the same topic [Каримов, 2022, 12]. Therefore, in scientific research methodology, the necessity arises to clearly define the functional boundaries of these two types of corpora.

Z. Xolmanova states that parallel text corpus can be created in two directions based on the characteristics of the original text and target language [Xolmanova, 2021, 57]:

1. Parallel text corpus of related languages.

2. Parallel text corpus of languages belonging to different families.

Parallel text corpus of related languages – crucial in analyzing general Turkic development stages and identifying subsequent development characteristics of Turkic languages. In matters of the degree of preservation of synharmonism phenomenon characteristic of Turkic languages, the usage status of oldest phonetic phenomena, lexical and morphological elements in later periods, corpora also serve as a practical source. Corpora created by domains provide necessary material for linguists. Primarily, they create the possibility of investigating domain terminology. In identifying terms and their sources of formation, domain-specific corpora provide practical assistance to researchers [Xolmanova, 2021, 57].

N. Jo'rayeva divides parallel corpora into two types according to their function [Жўраева, 2022, 45]:

1. Text corpus where one is the translation of the other;

2. Text corpus in two languages on one topic.

The first type of corpus is named "parallel corpus" (parallel corpora) and is used for studying various aspects of a certain translation. A parallel corpus (parallel corpora) is a linguistic resource that plays an important role in deeply studying various aspects of the translation process. This type of corpus typically contains texts in one language and their translations into another language. Parallel corpora are divided into two main types: aligned and not aligned corpora.

In aligned parallel corpora, translation units (sentences, phrases, or expressions) are connected in a form that mutually corresponds precisely. In this case, the exact correspondence of each translation unit in the original text is indicated. This characteristic makes this type of corpus very useful in translation research, especially for translators. The reason is that such corpora contain a unique linguistic resource called "translation memory." Through this memory, translators gain the opportunity to translate new texts based on previously used translation variants. As an example of multilingual, aligned parallel corpora, the European Union's Acquis Communautaire database can be cited. This corpus is considered one of the largest parallel corpora in the world, distinguished by its two main advantages—openness to users and free provision.

Unaligned parallel corpora, in some literature, are also called comparable corpora. In this type of corpus, there is correspondence based on general similarity in content, though not precisely defined between original text and translation. The alignment process can be accomplished through automatic tools or manually by humans. Although the automated alignment method is fast and convenient, it sometimes makes errors. For example, a simple sentence in the original text may be expressed as a complex sentence in the translation.

The scientific and practical value of parallel corpora is primarily determined by their volume and the number of languages covered. Therefore, such corpora possess great significance for translation technologies, machine translation systems, and linguistic research. Specialists define the importance of these corpora as follows: 1) revealing typical translation methods and transformations; 2) studying statistics of automatic translation systems; 3) creating single and multilingual dictionaries; 4) studying and evaluating data storage and transmission programs; 5) automatic checking of translation accuracy; 6) facilitating translator's work through wide possibility of equivalent selection. The second type of corpus is called "translation corpora" and is important for studying the expression of the same idea in various languages [Жўраева, 2022, 45].

"In contemporary corpus linguistics, parallel corpus has forms such as Multilingual Corpora and Translation Corpora. The structural composition of such a corpus, depending on its purpose, can be: 1) in ordinary text format referring to translation; 2) in the form of 'mirror texts' convenient for comparison; 3) in database format" [Каримов, 2022, 13].

Today, one of the popular parallel corpora is the Europarl

corpus—consisting of texts from official discussions of the European Parliament, containing multiple languages. This corpus was collected from parliamentary speeches between 1996 and 2007. Europarl is primarily widely used in processing multilingual texts. The Europarl corpus contains 21 European languages, including English, French, German, Spanish, Italian, Russian, and others. The corpus volume increases annually. In large versions, there are texts consisting of hundreds of millions of words. The corpus uses precise and formal forms of language used in the European Parliament. Europarl is used in creating multilingual translation models, developing systems like Google Translate, and automatic text translation. It is also useful for language teaching, syntactic and semantic research. Due to the official and political characteristics of texts, the opportunity to study everyday language is limited.

Let us analyze BTEC, one of the parallel corpora in Turkic languages. BTEC is a parallel corpus created by Bilkent University in Turkey, primarily containing translations in business, technical, scientific, and general subjects. The corpus primarily uses formal and academic language. The BTEC corpus covers large volumes of texts, containing over 1 million words of immutable, parallel texts. The BTEC corpus is useful in developing automatic translation systems between Turkic and English languages, including creating Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) methods. This corpus is used in identifying similarities between languages and evaluating translation quality. Due to the presence of material related to formal, scientific, and technical language, this corpus can also be used for Natural Language Processing and syntactic, semantic research. The corpus has a structure convenient for testing machine translation and Natural Language Processing systems, but material related to everyday language is limited. Its disadvantages include limited opportunities for studying informal language or daily conversations.

Tilde-Turkish Parallel Corpus is a parallel corpus created for translation and Natural Language Processing among Turkic languages. The corpus was created for translation and text analysis between Turkish and other Turkic languages (such as Kyrgyz, Uzbek, Turkmen, Tatar, and others). This corpus was developed by the Tilde organization and contains parallel texts available in many Turkic languages. The corpus contains texts on various topics, including: news and articles, materials related to politics and economics, technical and scientific texts, daily conversations and informal expressions. These texts are often obtained from various resources, most of which belong to formal and academic fields, though some informal texts are also available. The Tilde-Turkish Parallel Corpus is useful in translation analysis and Natural Language Processing among Turkic languages. Through this corpus, it is possible to develop machine translation systems and identify characteristics necessary for creating automatic translation in other Turkic languages. The corpus can be used in developing translation systems among Turkic languages, for example, Statistical Machine Translation (SMT) or Neural Machine Translation (NMT) methods. The corpus is available in various languages, creating opportunities for studying translation problems between different languages, as well as for linguistic research and multilingual algorithm studies. The Tilde-Turkish Parallel Corpus is an important tool applied in translation and Natural Language Processing among Turkic languages. Through this corpus, it is possible to create and study translation systems between various languages. This is particularly useful in developing multilingual machine translation systems, linguistic research, and language learning. Despite its disadvantages, its coverage of various languages and topics increases its scientific value in the field.

In parallel corpora of works, existing translations are typically used, but machine translation also plays an important role in working with parallel corpora. Translation of works and parallel corpora are interconnected and complement each other. Parallel corpora can contain both ordinary translations (performed by humans) and automatic translations. Several programs exist for analyzing works through existing translations in parallel corpora.

Most translations of works are performed by specific individuals—translators. For example, translations of famous literary works like "Don Quixote" or "The Little Prince" were created by translators. We have information about the existence of parallel corpora for these works, but since internet addresses were not accessible, they were not included in the work.

Some variants of corpora were created by machine translation systems. Machine translation, as automatic translation of initial works, is used in parallel corpora. Machine learning algorithms and neural networks are used to accomplish this process. Although these translations may not be accurate and

reliable, some parallel corpora provide machine translation in several languages.

Machine translation systems often rely on parallel corpora because parallel corpora are useful for ensuring high quality of the translation process. Machine translation systems (such as Google Translate or DeepL) are trained based on parallel corpora, meaning they analyze parallel texts (human translation or machine translation) for translating from one language to another. Machine translation systems use the data in parallel corpora to translate from one language to another. These systems apply neural network-based methods to help improve the translation process.

Parallel corpora are particularly important for evaluating and analyzing translation quality. Parallel corpora created for works, for example, together with translations of classical literary works, are also used to examine the performance of machine translation systems. For example: If a work's translation exists in many languages (such as " Слово о полку Игореве "), its translation and translation quality can be compared using machine translation systems. Parallel corpora provide these comparison possibilities because they may contain both human-created translations and automatic (machine) translations in their composition. Some parallel corpora are created solely by machine translation systems. Such corpora are primarily used for Natural Language Processing, language learning, and training translation systems. These corpora are intended more for creating and regulating automatic translation systems.

In conclusion, it can be stated that in parallel corpora, existing translations typically use translations performed by human translators, but machine translation also finds its place in parallel corpora because it is used for creating automatic translations and evaluating translation quality. Machine translation systems learn from parallel corpora and are used to improve the translation process, but for high-quality translations, works created by translators play a very important role in parallel corpora.

Text-based parallel corpora serve as the primary source at multi-layered (grammatical, semantic, stylistic, discursive) analysis levels of linguistic research. Through them, not only the usage characteristics of linguistic units in the translation process are understood, but also the cultural and communicative context behind these units is deeply comprehended. This serves as one of the main factors in bringing modern translation studies and comparative linguistics to a new methodological stage.

For aligning translation to the original text, the following methods are primarily applied in parallel corpus: align (alignment), POS tagging, syntactic tagging, semantic tagging, and annotating ambiguities. These methods assist in better understanding relationships between translation and original text, evaluating translation quality, and conducting linguistic analysis. The tagging process is particularly necessary for improving machine translation systems and analyzing the translation quality of works.

REFERENCES

1. Dilafruz, M. (2023). "Creating an Electronic Platform for 'Baburnoma' – A Demand of the Era." Роль наследия Захириддина Мухаммада Бабура в развитии восточной государственности и культуры, 1(1).
2. Jo'raeva, N. Sh. (2022). About Multilingual Parallel Corpora. Education and Innovative Research, 3, 43–47.
3. Karimov, R. (2022). Linguistic and Software Issues of Creating Uzbek-English Parallel Corpus (PhD Abstract). Bukhara.
4. Khamroeva, Sh. (2018). Linguistic Foundations of Creating Uzbek Language Authorial Corpus (PhD Dissertation). Bukhara.
5. Kholmanova, Z. (2021). Theoretical and Practical Issues of Creating Uzbek National and Educational Corpora. In Proceedings of International Scientific-Practical Conference (pp. 55–59). Tashkent.
6. Melnik, V. I. (2006). Text and Its Structure. Moscow.
7. Mukhammadiyeva, D. (2021). Reflection of Paremiologies in Turkic Language Translations of "Baburnoma". Uzbekistan: Language and Culture, 4(4).
8. Mukhammadiyeva, D. (2022). Principles of Translating Proverbs and Adages in "Baburname". Turkology, 1.
9. Mukhammadiyeva, D. (2022). Problems Related to Translation of "Baburnoma" into Turkic Languages. Turkologiya, 121, 65–84.
10. Mukhammadiyeva, D. (2023). Translation Issues in Creating Uzbek-Turkish-Kazakh Parallel Corpus of "Baburnoma". In The Role of Zahiriddin Muhammad

Babur's Heritage in the Development of Eastern Statehood and Culture (p. 190).

11. Mukhammadiyeva, D. (2025, November). Translation Principles in Creating "Baburnoma" Parallel Corpus. In Conferences, 1(01).
12. Musulmankulova, Sh. (2024). Linguistic Provision of Uzbek and Turkic Frazeme Corpus (PhD Dissertation, p. 35). Tashkent.
13. Rakhmanova, A. (2021). Computer Methods in Creating Uzbek National Corpus (PhD Dissertation, p. 63). Tashkent.
14. Sultanova, A. (2024). Theoretical Foundations of Creating Karakalpak Educational Corpus. In Modern Technologies of Computer Linguistics (CTCL) (p. 44). Tashkent.
15. Tareva, E. G. (2024). Dialog Cultures in the Context of Globalization. Moscow.
16. Zakharov, V.P., Kutuzov, A.B., Nedoshivina, Y.V., Rikov, V.V., Plungyan, V.A. (Various years). Research on Corpora in Russian Linguistics.
17. Francis, N., Kučera, G. (1961-1964). Principles of Corpus Construction. Brown University.
18. Bloomfield, Fries, Bondgers (1940s). Initiation of Targeted Research in Corpus Field.
19. http://ruscorpora.ru
20. http://www.inforeg.ru/electron/concord/concord.htm
21. https://tuebingen.de/b1/en/korpora.html
22. http://www.collins.co.uk/Corpus/CorpusSearch.aspx
23. http://sara.natcorp.ox.ac.uk/ http://corpus.nytud.hu/mnsz/
24. http://www.corpusdelespanol.org/
25. http://www.perseus.tufts.edu
26. http://www.cilta.unibo.it/ricerca.html
27. http://www.rcl.cityu.edu.hk/livac/
28. http://corpora.ids-mannheim.de/~cosmas/http://korpus.juls.savba.sk
29. http://ucnk.ff.cuni.cz
30. https://www.tnc.org.tr
31. http://web-corpora.net/TatarCorpus/search/index.php?interface_language=ru
32. http://web-corpora.net/KazakhCorpus/search/?interface_language=ru
33. http://web-corpora.net/bashcorpus/search/index.php?interface_language=ru http://uzschoolcorpara.uz