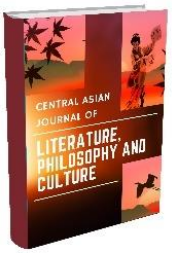

CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY AND CULTURE

eISSN: 2660-6828 | Volume: 04 Issue: 12 Dec 2023
https://cajlpc.centralasianstudies.org

Theoretical Basis of the Formation of Uzbek-Turkish Parallel Corpus

Shaxribon Musurmankulova
PhD student at Gulistan state university

Received 4th Oct 2023, Accepted 5th Nov 2023, Online 22th Dec 2023

ANNOTATION

The article provides information on what a corpus is, the formation of linguistic corpora in world linguistics, machine translation and the work carried out in this regard, parallel corpora and types of corpora. Bilingual and multilingual texts, the formation of common corpora between Turkic languages, and the creation of parallel corpora ensure the development of languages and their improvement as a technical language. The purpose of the idiom parallel corpus, forms of creation are explained. The stages of forming the database of the parallel corpus of phrases in the Uzbek and Turkish languages are described. Processes performed in steps are explained. The stages of text processing, scanning, editing, and formatting are analyzed. In the process of automatic translation, the problems related to the translation of stable compounds and idioms are highlighted. The requirements for the creation of the linguistic support corpus of the Uzbek-Turkish parallel corpus are indicated. The role of the Uzbek-Turkish parallel corpus in the implementation of scientific research, automatic translation, the formation of national corpus and educational corpus is shown. The practical importance of the Uzbek-Turkish parallel corpus in teaching the Uzbek language as a foreign language, learning the Turkish language, and translating artistic sources in the Uzbek and Turkish languages has been shown. It is stated that the created corpus performs the function of material and linguistic support in the formation of parallel corpora, educational corpora, and national corpora.

It is highlighted that the Uzbek-Turkish parallel corpus is a very necessary resource for overcoming difficulties in translation processes, parallel corpuses occupy a central place in translation studies and comparative linguistics.

KEYWORDS: phrase, Uzbek-Turkish parallel corpus, corpus, parallel corpus, parallel corpus of Uzbek-Turkish phrases, stages of corpus formation, text processing, editing, analysis, graphematic analysis, sorting, tagging, search engine , concordance.

Introduction: Computational linguistics has been formed as a science since the 50s of the 20th century and is a field that has been developing in world linguistics until now. Computational linguistics is developing as a modern field in world linguistics. It is considered a new direction in Uzbek linguistics and deals with the issues of solving language problems within the limits of computer capabilities. The following can be highlighted about the features of computer linguistics, its role in social spheres, and its importance in linguistic research:

1	Computer linguistics makes it possible to compile and systematize the vocabulary of the language
2	Computational linguistics serves to analyze linguistic issues by means of a computer, to illuminate the possibilities of language, to reflect it using the computer-informational method.
3	Computer linguistics lays the foundation for the formation of new areas such as corpus linguistics and computer lexicography
4	Computational linguistics provides linguistic research with material that is diverse in terms of content, allows to collect, systematize, compare, and contrast data.
5	Solving the issues of linguoculturalology, linguostatistics, comparative linguistics, cross-linguistics using a computer will have practical value.

Corpus linguistics was formed as a field of computational linguistics and is currently developing as an independent field in world science. Corpus linguistics deals with the technology, methods and database formation of corpora. Corpora are databases that illuminate the possibilities of a language.

Corpus (corpus) is a Latin word that means "body". "Corpus is a collection of texts in electronic form that means finding words, word meanings, word combinations, grammatical forms through a specific search engine" [http://rusorpora.ru].

Corpora are also valuable as a resource for improving linguistic research. The role of corpora in improving research is reflected in the following:

1. A large amount of linguistic information is collected in corpora.	2. All words of the traditional layer within one language - old layer: archaisms, historicisms; modern layer; new layer words - neologisms are included.	3. As a result of subjecting the corpora to the search system, it is possible to find information quickly and eas	4. Corpora perform statistical analysis of language units and linguistic processes with the help of concorders.	5. The corpus of texts performs the function of providing factual material and sources.

The formation of corpus linguistics on a global scale dates back to the 60s of the 20th century. Formation of corpus linguistics is related to machine translation. A.N. Baranov [Baranov, 2001], V.P. Zakharov [Zakharov, 2005], Y.V. Nedoshivina [Nedoshivina, 2006], K. Boyarskiy [Boyarskiy, 2013], N. Kozlova [Kozlova, 2013] corpus, its types, itself carried out research on the principles of corpus composition.

World-famous Russian, British, Slovenian, Hungarian, Spanish, Italian, Chinese, and German language corpora are widespread.

Corpus types are described in the studies of A. Baranov, V. Zakharov. Corpus according to the form of storage (audio, written, mixed), according to the language of the text (monolingual and multilingual); according to genre affiliation (literary, dialectal, oral, journalistic, mixed), according to access to the corpus (free, commercial corpora, closed); according to the purpose (research, illustrative); according to its dynamics

(dynamic (monitor), steady); divided into groups such as annotated (tagged) and untagged according to the presence of additional information [Zakarov, 2005:48].

Corpus are of two types according to the language of expression of the text: corpus of texts in one language; corpus of parallel texts.

The corpus of texts in one language is divided into small groups: corpus of titles, corpus of research, illustrative corpus, corpus related to the field, corpus of authorship, corpus of research, corpus of journalistic texts, linguodidactic (educational) corpus.

Research object and used methods

Uzbek and Turkish idioms were taken as the object of research. Uzbek phraseology was based on the reworked and expanded edition of "Dictionary of Uzbek language phraseology" published by Professor Shavkat Rahmatullayev in 1992 [11]. Idioms in the Turkish language Ömer Asim Aksoy's "Proverbs and Sayings Sözlüğü II". Retrieved from "Dictionary of Idioms" [1]. Comparative-historical, cross-sectional, and statistical analysis methods were used to illuminate the research topic.

The obtained results and their analysis

Formation of a database for the corpus of parallel mats. A corpus of parallel texts is an electronic representation of works of art, manuals, mass media, various documents in two or more languages. Currently, parallel corpora have been created that reflect the text features of English, German, Japanese, Finnish, and Slovak languages with Russian [Rakhmonova, 2020:30].

A parallel corpus is a corpus containing sets of original texts in L1 and their translated sets of texts in languages L2 ... Ln. In most cases, a parallel corpus contains information in only two languages.

Closely related to parallel corpora are "comparative corpora (comparable corpora)", which are two or more works that are similar in genre, subject, but not identical in content and form. consists of texts in more than one language. The corpus of parallel texts can be bilingual (Uzbek-English, Russian-English) and multilingual (Russian-English-French). In other words, they consist of texts in two or more languages. They can be one-way, two-way or multi-way:

Send feedback

Side panels

1. Unidirectional - corpora aimed at translation into one language only; for example, an English text translated into German;	2. Bilateral - corpora adapted to mutual translation between two languages; for example, texts translated from German to English and vice versa from English to German;	3. Multidirectional - multilingual, corpora adapted for translation into different languages; for example, English texts translated into German, Spanish, French and other languages.

In the world experience, there is no corpus of parallel texts related to Turkic languages. In recent years, the progress made in Kazakh and Turkish computer linguistics laid the foundation for the creation of all-Turkish corpora. The corpus of parallel texts based on Turkic languages makes it possible to spread the content of artistic texts specific to the period of all-Turkic development, to master written sources related to the next

stages of language development. Common corpora of Turkic languages are formed on the basis of parallel corpora of developed languages. All-Turkish corpora are important as a source of enrichment of studies in the fields of comparative linguistics, etymology, epistemology, geneology, dialectology, folklore studies, textual studies, translation theory, literary studies. The creation of the common corpus of Turkic languages will help to analyze the characteristics of the Turkic languages during the development of the Turkic language, to clarify the mutual genetic and synchronic relations of the Turkic languages, to solve the issues related to the phonetic-structural, grammatical, semantic changes in the vocabulary of the Turkic languages, to the further development of the Turkic languages. serves. The common corpus of Turkic languages, parallel corpora serve as the basis for improving a small number of Turkic languages such as Kumyk, Gagauz, Balkar, Altai (Uyrot), Khakas, Khalach, Karaim [Erdoğan, 2011:74]. Common monuments of Turkic languages play an important role in studying and promoting the common cultural heritage [Rakhmonova, 2020:31].

The corpus of Uzbek-Turkish parallel texts contains a large amount of information. It is difficult to form the database of this corpus directly and in its entirety. It is desirable to form a database for corpora in terms of seasonal-semantic groups, genres or styles. Therefore, we limited the dictionary of phrases in Uzbek and Turkish languages for the database of Uzbek-Turkish parallel corpora.

The corpus of Uzbek-Turkish phrases is one of the first examples of a parallel corpus between Turkic languages. We tentatively named this corpus Uzturkfraz parallel corpus. Like other corpora, the texts of the sentence corpus are selected according to specific criteria that depend on the purpose of its creation. In particular, the compilers decide whether to include a static set of Uzbek and Turkish phrases and full-length texts with their participation. Phrases in the Uzbek language were selected based on the "Dictionary of Uzbek Language Phraseologisms" published by Professor Shavkat Rahmatullayev. The dictionary was also published in 1978, and the corpus of parallel texts was based on the reworked, supplemented edition of 1992 [Rahmatullayev, 1992]. Idioms in the Turkish language Ömer Asim Aksoy's "Proverbs and Sayings Sözlüğü II". Taken from "Dictionary of Sayings" [Aksoy, 1984].

Uzturkfraz parallel corpus is based on the text of the explanatory dictionary of phrases in the Uzbek and Turkish languages. Among the requirements for creating and tagging a parallel corpus is the selection of appropriate texts, paying attention to the issues of author, volume, topic, genre, and style. In this respect, Uzturkfraz meets the requirements. Dictionary texts of both languages on the same topic, i.e. idioms, were selected. Linguistic support of the Uzturkfraz corpus was formed based on the requirements for creating a corpus:

1. Corpus should be based on natural language data. Uzturkfraz is based on the text of Uzbek and Turkish idiom dictionaries, these languages are natural languages with a history of many thousands of years of ancient development and their own stages of development. As a natural language, these languages are performing a communicative, emotional-expressive, accumulative function. The number of Turkish-speaking peoples in the world is increasing. In particular, learning and using the Turkish language in the process of communication is spreading widely in Uzbekistan. The level of knowledge acquired in the Turkish language is formalized with an international certificate, which provides certain opportunities to the owner of the language. The issue of developing the Uzbek language at this level and raising its status as a state language has become one of the most urgent topics.

2. The corpus should be representative, that is, it should contain elements of various speech styles. In this respect, the parallel corpus of Uzturkfraz meets the requirements and covers texts in colloquial and artistic style. Partially, official and scientific style texts are also reflected.

The technology of creating a corpus is covered in scientific studies. A. Baranov provided information on the basic concepts of corpus linguistics, text corpus, problem area, database, corpus data storage units, research corpus, illustrative corpora, methods of displaying and storing dynamic and static text corpus [Baranov, 200:8].

Stages of formation of a parallel corpus of phrases of the Uzbek and Turkish languages. V. Zakharov noted 9 stages as a technological process of forming corpuses [Zakharov, 2011:36]. The formation of the Uzturkfraz parallel corpus database was carried out on the basis of these stages.

1. Providing text input based on the mentioned sources. The text of the explanatory dictionary of phrases in the Uzbek and Turkish languages was used to form the linguistic support of the Uzturkfraz parallel corpus. "Dictionary of phraseology of the Uzbek language" authored by Professor Shavkat Rahmatullayev and "Atasözleri ve Deyimler Sözlüğü II" by Ömer Asim Aksoy. Phrases taken from the sources of "Deyimler Sözlüğü" and their explanation, compatibility of form and content (phrases that are compatible in terms of form and content), partial compatibility of form and complete content (phrases that differ phonetically, structurally and grammatically, but fully compatible in content), were placed on the basis of the principles of content compatibility (phrases inconsistent in form, but compatible in content). Phrases that are not given examples in dictionaries have been enriched with examples taken from artistic sources. In addition, 378 phrases not recorded in dictionaries were included.

2. Reformatting in machine-readable form. "Dictionary of phraseology of the Uzbek language" authored by Sh. Rakhmatullayev was published in the Cyrillic alphabet.

The selected texts were reworked in forming a parallel corpus. Automatic methods are used in the formation of corpus texts. The most active automatic method of text formation is scanning. The phrase was processed using parallel corpus text scanning. Scans were made using special technical means. The following defects on the scanned text have been eliminated:

- letters that have changed to another view have been fixed;
- extra letters appearing in words were removed;
- "-" hyphen and "-" hyphen symbols were distinguished.

After these processes, the text was converted to the Latin alphabet using the TransEdit.exe program. Ömer Asim Aksoy's "Proverbs and Sayings Dictionary II. The text of Deyimler Sözlüğü" was given in the same form.

3. Analysis and preliminary processing of texts. The text has been edited for spelling and style.

4. Conversion and graphematic analysis. Graphematic analysis is a program for analysis based on the relationship in natural text. The text of the dictionaries intended for the linguistic support of the corpus was analyzed as follows:

- idioms and their comments, examples are separated in the form of a table;
- non-standard (non-lexical) elements were separated;
- the paragraph, the main units of the dictionary article, the title, comments are separated.

When forming a database for a parallel corpus, tokenization, lemmatization, stemming, parsing are distinguished as the main procedures of natural language processing.

5. Text formatting. Labels are of practical importance in highlighting the features of the text and in quickly finding the necessary information. Parallel corpus is modeled in two ways according to the theory of text parsing:

1) extralinguistic information was provided. B. Bazarova cited extralinguistic information as additional information and noted text name, author's name, author's gender, year of birth, year of creation as the main parameters of the text [Bazarova, 2016:24]. We used parameters in a creative way.

Send feedback

Side panels

In the extralinguistic information part of the parallel corpus - information about the structure of the text, related to the entire text, a meta-description of the corpus was given:

Corpus name: Uzturkfraz

Sources for the corpus text:

Dictionary of Uzbek phraseology/ Proverbs and Sayings II. Dictionary of Idioms.

Year of publication: 1992/1984.

Authors: Shavkat Rahmatullayev/Omer Aksoy

The language of the work: Uzbek/Turkish

Text genre: dictionary.

Text size: 694/347

2) linguistic information. Linguistic additional information is information that describes text elements. In the parallel corpus, phrasemes were classified structurally and morphologically. Such a classification serves as factual material for studies focused on the process of formation of phrases, structural-grammatical structure, component analysis. In order to show the contextual features of phrasemes, a tagging system was created for the selected sources.

6. Corrected post-sorting errors and eliminated inconsistencies during the editing phase of automatic sorting results (manual and semi-automatic).

7. A database was formed and referred to the programmer to implement the process of converting classified texts into a special linguistic-informational-search system (corpus manager) that provides rapid multi-faceted search and statistical processing.

8. The stage of providing access to the case is carried out after the parallel case is ready. After the parallel corpus is formed with the help of a computer programmer, it is first distributed on a CD. After the examination, it is planned to be placed in the global network. Provision of access to the parallel corpus will be reviewed taking into account the age, training and capabilities of the users.

9. A document supply is created describing the aspects of creating and using the parallel corpus of idioms, a table that allows searching by metadata, and information about the query language of the corpus-manager is provided [Zaharov, 2011:36].

Summary. In the process of parallel corpus text processing, alternative variants of Uzbek phrases in Turkish were identified. Alignment of parallel texts is important because during the translation process the text may be manipulated by the translator, certain parts may be combined, deleted, added or rearranged to create a "natural" translation. To compare the original text and its translation, it is necessary to establish an alternative between the texts. Proportions of proper nouns, numbers, and other aspects are often used as points of reference during matching. The degree of concordance between parallel corpus texts varies depending on the type of text. For example, a fiction text may give the translator more freedom than a formal text.

The idiom parallel corpus can be used for various practical purposes. This corpus also serves to compare the structural-semantic features of phrases and their frequency of use in two languages. The Uzturkfraz parallel corpus is also used to study similarities and differences between the source Uzbek language and the target language Turkish. This allows for systematic, text-based contrastive and comparative studies at different levels of analysis. In this way, the Uzturkfraz parallel corpus is based on the typological and cultural differences and similarities between the Uzbek and Turkish languages and provides detailed information about these languages. Parallel corpora are considered an important stage in the development of comparative studies, and a number of issues of comparative linguistics are solved using parallel corpora. The Uzturkfraz parallel corpus provides an opportunity to compare the Uzbek and Turkish languages in translation studies. Parallel corpus also helps translators to find translation equivalents between Uzbek and Turkish. Uzturkfraz parallel corpus provides information about the frequency of idioms and compounds that are homonymous with idioms. This corpus also provides practical assistance to translators in developing systematic translation strategies for words or phrases that do not have a direct equivalent in Turkish. Recently, parallel corpora have been increasingly used in the development of bases for automatic translation systems.

Parallel corpora are widely used in teachers' activities. With the help of the parallel corpus, they can identify frequent linguistic phenomena in the language, enrich their knowledge of the language, design educational materials, and perform analysis on the original source during the teaching process. Parallel corpora are actively used, especially in the process of language learning.

The Uzturkfraz parallel corpus is of practical importance in teaching Uzbek as a foreign language, learning Turkish, and translating artistic sources in Uzbek and Turkish languages. This corpus serves as material and linguistic support in the formation of parallel corpora, educational corpora, and national corpora.

Uzturkfraz parallel corpus also eliminates difficulties in translation processes. The most difficult situation in the process of translation is to find the exact alternative of phrases and fixed combinations in the language being translated. Complete translation equivalence of stable combinations between source and target language texts is rare. Uzbek and Turkish idioms, although they have similar aspects, may belong to different semantic circles or contexts. The difficulty of translating idioms can be explained as follows:

1. Phrases have a complex structural-grammatical character, that is, they consist of two or more components. This is more complicated for machine translation than single word translation.

2. The phraseological component is related in terms of form and content, in most cases it is connected through certain grammatical forms.

3. From the meaning of the components of idioms, a coherent new lexical meaning emerges. Understanding this meaning is difficult for the automatic translation process.

4. Phrases express cultural signs, views, and values. This aspect also affects the meaning.

Parallel corpora are central to translation studies and comparative linguistics. As with most parallel corpora, the phrase parallel corpus is accessible through easy-to-use concordances. The search in the concordance system facilitates the analysis of events in the Uzbek and Turkish languages. This capability of the corpus ensures that it serves as a rich source of material for language teaching.

Uzturkfraz parallel corpus serves as training data for statistical machine translation systems.

It is important for linguists to use parallel corpora to be aware of subtle differences in meaning. Parallel corpora are increasingly used to design corpus-based (bilingual) dictionaries.

References

1. Baranov A.N. Introduction to applied linguistics. - M.: Editorial URSS, 2001.
2. Boyarsky K. Introduction to computer linguistics. - St. Petersburg, 2013.
3. Bazarova B.B. Introduction to corpus linguistics. - Ulan-Ude, 2016.
4. Erdoğan Boz, Yemen Ertuğrul. Turkish language for universities (written and spoken). - Ankara, Savaş Yaynevi, 2011.
5. Fisun Özgenç idioms in Turkish. Aquarium Publishing House, 2005
6. Zakharov V.P. Corpus Linguistics. Uchebno-methodicheskoe posobie. - St. Petersburg, 2005.
7. Zakharov V.P., Bogdanova S.Yu. Corpus Linguistics. -Irkutsk: IGLU, 2011.
8. Kozlova N.V. Lingvisticheskie corpus: definition of basic concepts and typology. Vestnik NGU. Series: Lingvistichekaya i mejkulturnaya communication. 2013. T. 11. Vypusk 1.
9. Nedoshivina E.V. Programs for working with text corpus: overview of basic corpus management. Uchebno-methodicheskoe posobie. - St. Petersburg. - 2006.
10. Rakhmonova A. Computer methods for creating the national corpus of the Uzbek language. Philol. science. fals.doc.diss... – Tashkent, 2020.
11. Shavkat Rahmatullayev, Nizomiddin Mahmudov, Zulkhumor Kholmanova, Iqbal Orazova, Kamola Rikhsiyeva. Explanatory dictionary of Uzbek phraseology. Tashkent, 2023.
12. http://rusorpora.ru.

