

Corpus of Parallel Texts in Foreign Language Teaching

Yusufova Liliya¹

The rapid development of new information and communication technologies has contributed to the emergence of a number of methodological works in which the authors developed methods for teaching the types of speech activity and the formation of students' language skills based on modern ICT [1]. The linguistic corpus is one of the technologies on the basis of which it is possible to form the lexical and grammatical skills of students' speech.

Note that the history of the use of the linguistic corpus in teaching and research goes back to the 60s. XX century, when researchers W. Francis and G. Kuchera created the world's first linguistic corpus at Brown University (USA). Their linguistic corpus included 500 printed works, each of which was no more than 2000 words in volume. All texts belonged to the most common genres of printed prose written and published in the United States. The texts selected for the corpus had to represent 15 genres, from which 6 to 80 samples were made:

1. press: reportage;
2. press: editorial;
3. press: reviews;
4. religious texts;
5. skills, activities, hobbies;
6. popular science literature;
7. fiction, biographies, essays;
8. Miscellaneous (government documents, business reports, industry reports, college directories);
9. scientific essays;
10. fiction;
11. mysticism and detective stories;
12. scientific prose;
13. adventure literature and westerns;
14. love stories;
15. humorous works.

It was in the 60s. 20th century W. Francis was the first to use the term "corpus", which meant "the totality of the most representative texts for linguistic analysis". As P.V. Sysoev, the appearance of the first linguistic corpus instantly attracted the attention of the scientific community to it, which was reflected in a completely scientific discussion [5]. One of the main aspects of the discussion was the principles of text selection and the list of research tasks to be solved on the basis of this linguistic corpus. According to W. Francis and G. Kucher, when creating their linguistic corpus, a number of criteria, which include, guided them:

1. the origin and composition of the text (the author had to be a native speaker of American English, the dialogue had to take up less than half of the text);
2. synchronization (texts first published in 1961 were included);
3. a well-thought-out ratio of the numerical representation of various genres and selection individual texts using a special probabilistic procedure;
4. accessibility for computer processing (special marks for the transfer of graphic features of the text, etc.) [5].

The main task of the authors of any linguistic corpus is to collect as many texts as possible related to the subset of the language for which the corpus is being created. In this regard, we can conclude that the corpus is a reduced model of a particular language. The main characteristic feature of the corpus is its representativeness, i.e. the obligatory proportional representation in the corpus of texts of various periods, genres, styles, authors, etc. Thus, a linguistic corpus is a system of texts located on a machine source and combined according to various parameters (language, genre, style, industry). One of the characteristics of the case is its placement on an electronic carrier for computer data processing. Concordance is one of the programs that allows you to process and analyze large arrays of texts, as well as to identify linguistic patterns in them.

Concordance searches for a specific word or phrase in a corpus. After the appropriate command, the program issues a specified number of text fragments containing the requested units. Based on the information obtained, it is possible to draw a conclusion about the functions, use and application of the unit we have specified in a particular language space [5].

The management of textual and linguistic data in a corpus is often referred to as a corpus manager. A corpus manager is a specialized search system that includes software tools for searching data in a corpus, obtaining statistical information and providing results to the user in a convenient form. [1].

¹ Language Teacher, Department of English Language Practical Course, Jizzakh State Pedagogical University

The hull manager allows you to do the following:

- build both KWIC (Key Word in Context) and full concordance lists;
- search not only for individual words, but also phrases;
- search by templates (complex queries);
- sort lists according to several criteria selected by the user;
- to give the opportunity to display the found word forms in an unlimited context;
- give statistical information on individual elements of the body;
- display lemmas, morphological characteristics of word forms and metadata (bibliographic, typological), which depends on the degree of corpus markup;
- save and print the results;
- work both with individual files and with corpus unlimited in size;
- quickly process requests and issue results;
- support various text data formats (txt, doc, rtf, html, xml, etc.);
- be easy (intuitive) to use for both experienced and novice users [1].

Linguistic corpora can be used to obtain a variety of references and statistics about language and speech units. In particular, on the basis of corpora, one can obtain data on the frequency of word forms, lexemes, grammatical categories, track the change in frequencies and contexts in different periods of time, obtain data on the joint occurrence of lexical units, etc. A representative array of language data for a certain period allows one to study the dynamics processes of changing the lexical composition of the language, to analyze the lexical and grammatical characteristics in different genres and by different authors. Corpora are also intended to serve as a source and tool for multifaceted lexicographic work on the preparation of various historical and modern dictionaries [2]. Corpus data can be used to build and refine grammars and for language learning purposes. One of the most famous and widely used is the British National Corpus. It was he who formed the basis for the development of modern dictionaries in the English language.

The national linguistic corpus has two important features. First, the corpus is characterized by representativeness, or a balanced composition of texts.

In other words, it contains all types of written and oral texts presented in a given language (fiction of various genres, journalistic, educational, scientific, business, colloquial, dialect, etc.). In addition, all these types of texts are included in the corpus, if possible, in proportion to their share in the language of the corresponding period. Good representativeness is achieved only with a significant volume of the corpus (tens and hundreds of millions of word usages). Secondly, the corpus has markup that allows you to track specific words or phrases in the text. In the scientific literature, one can find various typologies of linguistic corpora, each of which is based on certain typological features. The focus of our study is the corpus of parallel texts, so it seems appropriate to dwell on such a typological feature as "parallelism".

It should be noted that most of the methodological research was devoted to the study of the formation of grammatical and lexical speech skills based on monolingual monolingual corpora [5]. However, in addition to monolingual linguistic corpora, there are also bilingual and multilingual corpora, i.e., corpora that contain texts in two or more languages. Multilingual corpora can be roughly divided into several types:

1. a corpus containing the original text and its translation;
2. monolingual corpus, which contains the original text and its interpretation in the same language;
3. corpus, which contains the text of the original and its interpretation in another language. It should be noted that the researchers have not come to a consensus regarding the terminology within the framework of this typological feature.

Some scholars consider the first type of corpus to be the corpus of translations, and the second type of corpus to be the corpus of parallel texts [3]. Others, on the contrary, the first type of corpus is a corpus of parallel texts, and the second is a comparative corpus [4]. For many other authors, both the first and second types of corpora are corpora of parallel texts [5]. In our study, under the corpus of parallel texts, we mean a type of linguistic corpus consisting of a source text in one language and its translation into another or other languages.

Summing up, we note that a corpus of parallel texts is a type of linguistic corpus consisting of a source text in one language and its translation into another or other languages. The characteristics of the corpus of parallel texts, on the basis of which it is possible to teach students professional vocabulary, include the following:

- A. the corpus uses professional written texts;
- B. the corpus uses source texts in one language and their translations in another;
- C. in the corpus it is possible to translate both from the studied language into the native language, and vice versa;
- D. the corpus contains texts of exclusively professional orientation of students;
- E. the corpus contains texts of the scientific genre;
- F. access to corpus resources can be open or limited (when using commercial products);
- G. in the learning process, "illustrative" corpora should be used, clearly showing the features of the use of professional vocabulary in the language being studied;
- H. the corpus should include modern professional texts;
- I. in the learning process, both full-text and fragment-text corpora can be used. The corpus of parallel texts has distinctive didactic properties and methodological functions.

Didactic properties include:

- a. multilingualism;
- b. contextuality of search results;
- c. the possibility of limiting the scope of the use of the language. These didactic properties of the corpus of parallel texts allow students to form lexical speech skills through translation when studying a foreign language for special purposes, as well as more accurately determine the meanings of new words and show the features of word usage in real language situations.

Reference

1. Baker M. The role of corpora in investigating the linguistic behavior of professional translators // *International Journal of Corpus Linguistics*. 1999. No. 4. P. 281-298.
2. Chernyakova T.A. Methodology for the formation of lexical skills of students on the basis of the linguistic corpus: author. dis. ... cand. ped. Sciences. M., 2012.
3. Johansson S. On the role of corpora in crosslinguistic research // *Corpora and cross-linguistic research*. Amsterdam, 1999. P. 3-24.
4. Ryazanova E.A. Methodology for the formation of grammatical skills of students' speech based on the linguistic corpus: author. dis. ... cand. ped. Sciences. M., 2012.
5. Sysoev P.V. Linguistic corpus in the methodology of teaching foreign languages // *Language and Culture*. 2010. No. 1 (9). pp. 99-111.
6. Zakharov V.P. *Corpus linguistics*. SPb., 2005.
7. Bakirova H.B. Teaching foreign language terminology at non-language universities. *International journal of discourse on innovation. Integration and education*. Volume: 01 Issue: 01. 2020 <http://summusjournals.uz/index.php/ijdiie>
8. Bakirova H.B. Terminological competence of the specialist in training vocabulary of specialty/ *Web of scientist: International scientific research journal*. ISSN 2776-0979 Vol. 2, Issue 5, 2021, Indonesia.
9. Bakirova H. Typology of methodological and linguistic difficulties in the formation of lexical competence. *ACTA NUUZ*. 1/5/1 2021. 44p. <http://science.nuu.uz/uzmu.php>
10. Bakirova H.B. The role of terms of specialty in professionally oriented education. *Journal of Hunan University (Natural sciences)* Vol 48. No.11.2021. 1430p.
11. Bakirova H.B. The content of teaching foreign languages. *Eurasian Journal of Learning and Academic Teaching*. Vol.2 www.geniusjournals.org. ISSN: 2795-739X. Belgium. 10-14p.