



## CORPUS OF SOURCE TEXTS AS AN EMPIRICAL BASIS FOR LINGUISTIC RESEARCH

**Olimzoda Parvina**

*Assistant,*

*Department of Languages,*

*Samarkand State Medical University*

**Abstract:** This article examines the theoretical and methodological foundations of corpus linguistics and source text corpora. It emphasizes the importance of empirical data in modern linguistic research and presents corpus linguistics as an objective, statistical approach to language analysis. The study outlines corpus structure, formation, annotation, lemmatization, statistical analysis, and visualization, while highlighting applications in language teaching, lexicography, translation studies, and applied linguistics. The findings confirm that corpus linguistics is a key tool for ensuring scientific accuracy and objectivity in linguistic research.

**Key words:** corpus linguistics, corpus of source texts, annotation, lemmatization, statistical analysis.

**Introduction.** In modern linguistics, research based on empirical data has become one of the priority directions. As a result of the rapid development of information and communication technologies, it has become possible to study linguistic phenomena on the basis of large volumes of text. While traditional linguistic research relied more on individual observation and theoretical conclusions, corpus linguistics serves to conduct objective, systematic, and statistical analysis based on real speech material. Corpus linguistics is formed on the basis of a collection of electronic texts that reflect the real use of linguistic units. A corpus of source texts consists of a collection of texts selected according to specific criteria, converted into electronic form, and linguistically marked, which makes it possible to empirically study the lexical, grammatical, and stylistic features of a language. The purpose of this article is to elucidate the theoretical foundations of the corpus of source texts, to describe the methods of its formation and processing, and to scientifically substantiate its place in linguistic research. At the same time, the significance of corpus linguistics as a methodological approach and its fields of practical application are analyzed.

In modern linguistics, research based on empirical data occupies an important place. Corpus linguistics opens up new methods and research techniques for describing the lexical layer of a language more fully, while at the



same time making it possible to achieve new results. “Over the last decade, new information technologies have significantly expanded the possibilities of language study. Now linguistics is carried out not only through traditional ‘language data repositories,’ such as dictionaries, works of fiction, works of classical writers, or written texts regarded as representative samples of the main collection, but it has also become possible to enter this data into a computer and process large volumes of texts - text corpora.” A corpus reflects the use of linguistic phenomena in real speech and makes it possible to conduct research in various fields of linguistics on a scientific basis. A corpus, as a collection of texts, is a systematic set of the object and subject under study. The automatic analysis of texts in the form of a collection became an important reality for the development of corpus linguistics. A text corpus is understood as a collection of texts selected according to specific criteria, compiled in electronic form, and linguistically marked.

The theoretical roots of corpus linguistics are connected with the generative linguistics put forward by Noam Chomsky. In this regard, V. Zakharov rightly observes that a corpus is a reduced model of language. Indeed, language is realized across various discourses in the form of natural speech. Corpus linguistics makes it possible to observe, analyze, and study all expressions of speech.

In his article “What is Corpus Linguistics?” Ch. Taylor examines how the term “corpus linguistics” is defined in various scholarly works. He does not attempt to give a “final answer”; rather, by comparing different definitions and approaches, he aims to stimulate scholarly discussion on the topic. In the article, the author traces the history of the term corpus linguistics, its methodological foundations, and its place in modern linguistics, as well as the differences in approaches used in research in this field. He provides a general scholarly understanding of the theoretical foundations of corpus linguistics - that is, of its terms and methodologies - and considers the field not merely as a set of technical tools but as a methodological approach within linguistics itself. The assessment of a corpus not as a language model but as a methodological approach can also be found in the views of Douglas Biber. In discussing the history of corpora in his article, Ch. Taylor dwells on the “corpus of source texts” and cites the Brown Corpus and other early corpora as examples. One such source is the corpus of source texts, which makes it possible to study linguistic phenomena in an objective, systematic, and reliable manner. The corpus of source texts reflects the real use of linguistic units in speech. This, in turn, is of great importance in analyzing many problems in linguistics, including lexical meaning, polysemy, synonymy, terminological precision, and stylistic features. Research conducted on the basis of a corpus makes it possible



to study linguistic phenomena not through subjective observations but on the basis of statistical and quantitative indicators.

In her work, Susan Hunston, analyzing the possibilities of using corpora, notes that they are widely applied in seven main fields: language teaching and learning, lexicography, critical linguistics, comparative and translation studies, stylistics and literary analysis, forensic linguistics, as well as the development of software tools to assist writers. According to the author, data-based learning helps students actively participate as “language detectives,” independently discovering facts about the language they are studying based on authentic examples.

S. Hunston gives an example of organizing mutual learning, through the use of parallel corpora, between a French-speaking learner of English and an English-speaking learner of French. In this case, grammatical structures in French (for example, the pronoun *on*) and their translation equivalents in English are extracted from parallel corpora. By analyzing concordance lines, learners teach each other the language.

A corpus of source texts also plays an important role in identifying the differences between language and speech. While language, as a system, is an abstract phenomenon, speech is its practical manifestation. Corpus linguistics reveals the laws of language precisely through texts that are the product of speech activity. For this reason, a corpus is an important tool in determining the frequency, distribution, and contextual features of linguistic units. Furthermore, the corpus of source texts constitutes the main basis for research in statistical and computational linguistics. Processes of automated analysis, modeling, forecasting, and extrapolation are carried out precisely on the basis of corpus data. This, in turn, increases the accuracy and scientific validity of linguistic research. Processing a corpus of source texts provides linguistic research with a number of advantages: it strengthens the empirical basis, harmonizes statistical and qualitative analysis, and makes it possible to determine the distribution of terms and lexical units.

The corpus of source texts and the methods for its processing serve to conduct linguistic research in a systematic, objective, and scientifically grounded manner. The methods of corpus processing consist of the following main stages:

**1. Selection and collection of texts** - texts suited to the purpose of the research are selected and converted into electronic form. At this stage, the genre, topic, and volume of the texts are taken into account.

**2. Annotation and marking** - linguistic units, morphemes, word forms, terms, and their grammatical features are marked. This process is carried out using automated programs or manually.



**3. Lemmatization** -the base forms of words are determined. For example, various verb forms or plural forms of nouns are combined under a single lemma.

**4. Statistical analysis and modeling** - the frequency, distribution, and correlation of words and units are calculated and, if necessary, analyzed using distributions and models.

**5. Visualization and reporting** - the results are presented through graphs, tables, and diagrams, which makes the conclusions more understandable.

Thus, the corpus of source texts serves as the main empirical source, an analytical tool, and a guarantee of scientific reliability in solving problems in linguistics.

**Conclusion.** Corpus linguistics is an effective methodological approach in modern linguistics that serves to study linguistic phenomena on an empirical, objective, and statistical basis. Since the corpus of source texts reflects the real use of linguistic units in speech, it is of great importance in determining the relationship between language and speech. The processes of forming and processing a corpus - the stages of selecting texts, annotation, lemmatization, statistical analysis, and visualization - ensure the accuracy and scientific validity of the research results. The application of statistical and computational linguistics methods makes it possible to analyze linguistic phenomena deeply and comprehensively. Thus, the corpus of source texts serves as a reliable empirical basis in linguistic research and contributes to the development of theoretical and applied linguistics. In the future, improving corpus technologies and expanding national language corpora will further increase the effectiveness of scientific research.

## REFERENCES

1. Abdurakhmonova N.Z. Computer models of the Uzbek language electronic corpus: DSc dissertation in Philology. -Tashkent, 2021. - 250 p.
2. Biber D., Conrad S., Reppen R. Corpus linguistics: Investigating language structure and use. - Cambridge: Cambridge University Press, 1998. - 300 p.
3. Biber D. Corpus-based and corpus-driven analysis of language variation and use // The Oxford Handbook of Linguistic Analysis / Ed. by B. Heine, H. Narrog. — Oxford: Oxford University Press, 2015. - P. 159-191.
4. Cheung M. L. L. Merging corpus linguistics and collaborative knowledge construction: PhD thesis. - Birmingham: The University of Birmingham, 2009. - 438 p.
6. Hunston S. Corpora in applied linguistics. - Cambridge: Cambridge University Press, 2002. -233 p.



7. Hunston S., Francis G. Pattern grammar: A corpus-driven approach to the lexical grammar of English. - Amsterdam: John Benjamins, 2000. -300 p.
8. McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. -Cambridge: Cambridge University Press, 2012. - 294 p.
9. Sinclair J. Corpus, concordance, collocation. - Oxford: Oxford University Press, 1991. - 179 p.
10. Taylor C. What is corpus linguistics? What the data says // ICAME Journal. - 2008. -№ 32. - P. 179-200.
11. Magomedova A.N. Corpus linguistics and the contextual resolution of lexical polysemy: Author's abstract of PhD dissertation in Philology. -Nalchik, 2004. -21 p.
12. Marchuk Yu.N. Fundamentals of computational linguistics. - Moscow: Nauka, 1999. - 236 p.
13. Zakharov V.P. Corpus linguistics: a study guide. - St. Petersburg: St. Petersburg State University, 2005. -48 p.