International Journal of Pedagogics

VOLUME Vol.05 Issue06 2025
PAGE NO. 199-202
DOI 10.37547/ijp/Volume05Issue06-55

Definition Of Basic Concepts and Typology Of Linguistic Corpusa

Akhmedjanova Sitora Djurakhodjayevna

Lecturer, Department of Russian Language and Literature Bukhara State University Bukhara, Uzbekistan

Received: 14 April 2025; **Accepted:** 15 May 2025; **Published:** 19 June 2025

Abstract: Corpus linguistics is considered one of the most promising and progressive areas in language study. The relevance of this article lies in the enormous potential of linguistic corpora, which has not yet been fully recognized by the scientific community, at least because the text – the main object of corpus linguistics – in its various forms of implementation is one of the main components of the language system and the mental and speech activity of a modern native speaker. The article reveals the concept of "corpus", provides a classification of text corpora, describes in detail each group of text corpora, provides criteria for linguistic corpora, explains the concept of "markup", and examines the basic concepts of corpus linguistics, methods and areas of its application. The advantage of text corpora in linguistic research is described. The article also analyzes the emergence and development of corpus linguistics, provides a typology of corpora, and describes each type of corpus separately.

Keywords: Corpus, linguistics, corpus linguistics, text, semantic, anaphoric, graphematic, representativeness, conceptual apparatus, variation.

Introduction

Each study conducted by a linguist should be aimed at at least at certain stages of activity:

1. Selection of provisions and a basis for categorization of the objects under study.
2. Division of objects into categories in accordance with this basis.
3. Understanding and interpreting the results of dividing objects into categories, interpreting the grounds for such division.

At the same time, the first stage of this activity presupposes the existence of objects under study, i.e. the acquisition of practical data for the creation of a theory at the final stage.

Nowadays, corpus linguistics in the preparation and analysis of empirical data is becoming widespread, thanks to the intensive growth of information technology.

Methods. Corpus linguistics first became known in the 1960s. Texts were formed mainly on the basis of the English language, but soon corpora began to appear (in corpus linguistics, the plural form "corpora" is used. See: Explanatory Dictionary of the Russian Language edited by D. N. Ushakov: Corpus, pl. corpora) based on material from other languages.

At the same time, at Brown University in the USA, scientists W.N. Francis and G. Kuchera compiled the first corpus of texts on an electronic medium, which consisted of 1 million word usages (500 texts with 2,000 words each). It also had appendices in the form of an index of word frequency in alphabetical order and certain statistical data.

A corpus is a collection of texts in one or more languages that are related by certain characteristics. In their work, L. Lemnitzer and H. Zinsmeister gave the following definition of a corpus: "A corpus is a collection of written or oral statements. Corpus data is usually digitized, i.e. often stored on computers and machine-readable." [2, 7].

At the same time, the constituent elements of the corpus – texts – are made up of materials, metadata that these materials represent, and linguistic generalizations that these materials organize.

As a particular section of linguistics, corpus linguistics was finally formed at the end of the 20th century.

Corpus linguistics as a separate section of linguistics was finally formed in the first half of the 90s of the 20th century. At the same time, the conceptual apparatus began to take shape [5, 37]. In particular, J. Sinclair defines the concept of "corpus" as follows: "a set of naturally occurring language texts selected to characterize the state of diversity of a language" [3, 171].

Here is one of the key provisions for selecting texts for compiling a corpus - we are talking about unfinished texts, in other words, the language sounds in the form in which it was expressed (oral or written speech). Moreover, the corpus does not offer real “templates” and “positions” for the correct organization of a message, but the maximum possible number of “variations” of the language, although some of them are not located in the center of the language system.

Further, the concept of “corpus” is increasingly clarified: “A corpus is a collection of texts intended for some purpose, usually educational or research. [...]A corpus is not something a speaker says or knows, but something created by a researcher. It is a record of the performance of, usually, many different users, intended to be studied so that we can draw conclusions about typical language use. Because it provides methods for observing patterns of the kind that have long been noticed by literary critics but not revealed empirically, computer-aided exploration of large corpora may perhaps offer a way out of the paradoxes of dualism" [4, 239-240]. (Our translation)

We assume that a more or less complete formulation of the concept of "corpus" can be found in the works of V.P. Zakharov. The linguist describes the corpus as a large, electronically represented, organized and planned, philologically impressive conglomerate of linguistic data, designed to resolve specific linguistic issues and tasks.

This formulation can be qualified as “activity-based”, which, by and large, explains the linguistic tendency of organized text arrays.

Results and discussion. As a result, in any of the described formulations of the concept of "corpus" the following is noted:

1) a large number of texts must be presented in electronic form (on the Internet or on any medium);

2) the language material must be distributed for consideration for linguistic purposes;

3) following the review, there should be a method for variously dividing the obtained language data (by topic, genre, year of creation, etc.). Considering the first, the possibility of constant access to texts in electronic form was noted. A huge number of text corpora can be classified into three significant groups:

1. Freely available;
2. Partially available;
3. Commercial.

The first group includes a relatively small number of text corpora available today. The National Corpus of the Russian Language, which contains more than 500 million words, is considered to be quite substantial.

The next group includes the majority of available corpora; however, for solving certain linguistic problems, this partial access is considered quite sufficient. For example, in the British National Corpus, the query results are only up to 50 arbitrary examples, and most of the functions of the search interface, which is only provided together with the full (and paid) version of the corpus, are missing.

There is, however, a non-commercial version of this corpus that is made available through a simple registration process. This version offers searchable texts from 1980-1993, with around 100 million words.

The third group includes, for example, the British National Corpus, which has a free one-month pre-subscription option to gain access to Collins Wordbanks Online, which contains about 533 million words, before you can buy the commercial version of the corpus.

Another significant criterion of a linguistic corpus of texts is the presence or absence of markup, because the presence of a simple conglomerate of texts is not enough to solve linguistic questions and problems.

Markup is the assignment of special marks to texts and their elements: external, extralinguistic, systemic and strictly linguistic, which describe various parameters of text elements. Metamarkup includes not only information about the text, but also data about the author. Let's study strictly linguistic types of markup. One can start with marking the parts of speech that are frequently encountered in existing corpora, but at the same time not only morphological indicators are taken into account, but also grammatical ones.

The marking of parts of speech is carried out with the participation of special programs of automated morphoanalysis. For example, in a small part of the National Corpus of the Russian language (6 million word usages) manual elimination of morphohomonymy and auxiliary correction of the results of the process of the program of automatic morphoanalysis were carried out [6, 86].

In the Mannheim Corpus of the German language, the marking of parts of speech is present mostly in the sub-corpora of journalistic texts. Among other types of markup, it is especially necessary to pay attention to syntactic markup, which is not presented in the entire conglomerate of the corpus, but only in a small part of it, because this type of markup, which involves determining the syntactic structure for any sentence, is done almost manually and requires significant time expenditures. Also, the corpus contains other types of markup, for example, semantic, prosodic, anaphoric, graphematic, etc. This largely helps to simplify the procedure of natural data collection by the researcher, taking into account the correctly specified search conditions.

However, in order for the developed corpus of texts to satisfy various kinds of linguistic tasks that a linguist faces, it must, in turn, have at least two more indicators.

First of all, we mean the representativeness of the text corpus. Kibrik A.E., Brykina M.M., Leontyev A.P. and Khitrov A.N. believe that representativeness can be assessed by transforming the "relative frequency" of the fact under study with the growth of the "sample". If the “relative frequency” of a fact does not change frequently with the increase in “each subsequent fragment of text”, then this means that “the corpus as a whole is representative”. At the same time, although it is observed that it is inadmissible to determine the links with statistics in this formulation of representativeness, it is emphasized that this requirement is mandatory, but still incomplete for establishing the representativeness of the text corpus.

Basically, the issue of establishing the representativeness of different text corpora is still considered relevant, but, admittedly, it has not been sufficiently developed. Only representativeness transforms the usual complex of different texts into a text corpus suitable for carrying out linguistic research. At the same time, human speech activity is so diverse that it is almost impossible to actually convey all the existing variations of language mentioned above. For this reason, the question of the representativeness of a text corpus is considered more of a question of the impartiality of any scientific research. In this case, it is advisable to rely on the common sense of the researcher himself, when we mean a user corpus (developed by the researcher himself in accordance with the goals of his research), or a group of researchers, when we mean the creation of a corpus that requires a large scale of linguistic phenomena, styles, genres, etc. (for example, a national corpus of a specific language).

An important condition when designating a case is also its ease of use, in other words, the case must be equipped with a specialized search system, which must be (ideally) quite understandable to a sufficient degree and easy to operate. The operation of the National Corpus of the Russian Language or the British National Corpus (English Language Bank) presents significant problems, which cannot be said about the search system of the Mannheim German CorpusWe believe that the corpus should not take up a lot of time, which is necessary to search for a certain phenomenon, and should not offer a tricky search methodology, since studying its basic points requires from the researcher in some cases purely technical and mathematical knowledge.

Corpus and its types

In some cases it is very difficult to navigate among the existing diversity of research corpora, because the goals and tasks set before the linguist are often identified in general, but in specific fields and areas they differ. The initial stage carried out by the researcher in studying the "objects" under study is the correct choice of the appropriate corpus. The entire diversity of existing corpora is determined by the diversity of "research and practical tasks for the solution of which they are created" [7, 12].

1. Oral, written, mixed.

Oral corpus is a systematized complex of speech fragments, equipped with software capabilities for accessing them [1, 71-72]. Oral corpora first began to function in the 80s of the 20th century on the basis of American English. Then special coordination centers appeared that collected, stored, distributed and created oral corpora. For example, LDC (Linguistic Data Consorcium), CSLU (Center for Spoken Language Understanding), ELRA (European Language Resources Association).

Most of the existing corpora are written or mixed (for example, the accessible part of the Mannheim Corpus of the German language), yet the part of linguistically marked oral texts even in mixed corpora is quite small relative to the entire conglomerate of the corpus (very often these are national corpora of a certain language, for example: Russian, English).

2. Monolingual – bilingual/multilingual.

There are two groups of monolingual corpora:

– corpora covering the entire language,

– corpora covering only the language for specific purposes.

For example, the Corpus of Early English Medical Writing (CEEM) is a corpus of medical texts in English from 1375 to 1750, the volume of which is approximately 1.5 million words. It contains theoretical works, reference books, and poetic texts on medical topics.

In bilingual and multilingual corpora, texts can be presented either comparable or parallel. For example, in 1992, the European Corpus Initiative (ECI) was established as an international organization that is engaged in compiling a large multilingual corpus for research purposes. The present comparable corpus contains not only texts of European languages, but also texts in Russian, Turkish, Chinese, Japanese and many others. Their volume is more than 98 million words. This type of corpus is considered commercial. Corpora of parallel texts are intended, first of all, for comparative analysis of texts in the direction of "original - translated" for teaching methods, techniques and methods of translation. For example, the European Parliament Proceedings Parallel Corpus 1996-2011, which presents parallel texts of the session of the European Parliament in different European languages with translation into English.

3. Synchronous – diachronous. Synchronous corpora provide a representation of text data for studying the systemic state of a language in a specific period of time. Thus, the non-commercial version of the British National Corpus only contains texts from the period from 1980 to 1993.

To study the historical development of a certain linguistic phenomenon or the entire linguistic system in general, there are diachronic corpora. For example, the Thesaurus Indogermanischer Text- und Sprachmaterialien, which presents Indo-Germanic texts from different eras.

4. Unmarked – marked.

An unmarked corpus is a conglomerate of texts containing a specific number of mentions of the required component. At the same time, the search results offered in unmarked corpora can be used in linguistic research, but only from a statistical point of view.

Annotated corpora (morphologically, syntactically, etc.) are considered multifunctional, as they provide many more opportunities for linguistic analysis.

CONCLUSIONS

So, a corpus is a representative conglomerate of unedited texts, presented in electronic form, usually marked up for linguistic analysis, equipped with a relatively easy-to-use search system, which represents as many language variants as possible.

During the years of the emergence of corpus linguistics, the problems of computerization in this area were not identified, and “researchers pointed to the possibility” of ignoring the variability of language, namely “territorial, social, age, gender,” etc. linguistic distinctions. Nowadays, by ignoring it, we deliberately limit ourselves with various frameworks when studying texts of a specific language, which calls into question the objectivity of this kind of research. With the advent of electronic corpora, the diversity of forms of language existence has become more indicative, the means and possibilities for studying language data have increased. The modern linguistic corpus contains hundreds of millions of word usages, and the fact that, thanks to the electronic corpus, the results of word usage examples can be obtained incredibly quickly makes the task of linguists much easier. The typology of corpora shown, without claiming to be large-scale, presents us with a real diversity of text corpora and allows us to navigate it for further scientific research.

REFERENCES

Kodirov A. Linguistic and cognitive representation of the concept" HOPE" IN RUSSIAN //International Bulletin of Engineering and Technology. – 2023. – Т. 3. – №. 5. – С. 106-108.

Lemnitzer L., Zinsmeister H. Korpuslinguistik: Eine Einführung. Tübingen, 2006. – С. 7.

Sinclair J. McH. Corpus, Concordance, Collocation. Describing English language. Oxford: Oxford University Press. – 1991. – С. 171.

Stubbs M. Words and phrases: corpus studies of lexical semantics. Oxford, 2001. – P. 239-240.

Ахмеджанова С. Д. РУССКИЙ И УЗБЕКСКИЙ ЯЗЫКИ В СОПОСТАВИТЕЛЬНОМ АСПЕКТЕ //Ответственный редактор. – 2023. – С. 37.

Ахмеджанова, С. . (2023). СПЕЦИФИКА ЛАКУНАРНОСТИ В РУССКОЙ ЛИНГВИСТИКЕ. Евразийский журнал социальных наук, философии и культуры, 3(3), 84–88. извлечено от https://in-academy.uz/index.php/ejsspc/article/view/11205

Коршунова А. В. Зоонимы Как Слова-Компоненты В Русских Фразеологизмах //Miasto Przyszłości. – 2024. – Т. 47. – С. 673-675.

Кривнова О. Ф. Области применения речевых корпусов и опыт их разработки // Тр. XVIII Сессии Российского акустического общества РАО. Таганрог, 2006.