World Bulletin of Social Sciences (WBSS)
Available Online at: https://www.scholarexpress.net
Vol. 32, March 2024
ISSN: 2749-361X

WHAT IS CORPUS LINGUISTICS?

Nortojiyeva Dildora Saloxiddinovna,
Student of UzSWLU

Article history:
Received: 28th December 2023
Accepted: 23rd February 2024

Abstract:
The goal of this article is to give explicit information on definitions of corpus linguistics, electronic linguistic corpus, and corpus technologies. A brief historical background on the development of corpora and corpus linguistics was given. Morover, the types and advantages of corpora were discussed.

Keywords: corpus linguistics, corpora, text, representativeness, case technologies,concordance, computer

Corpus technologies originated in corpus linguistics, a branch of linguistics that designs, creates, and uses various linguistic text corpora. Corpus linguistics suggests that more precise analysis of language is more feasible with corpora collected in natural contexts, as well as with minimal experimental intervention.

Corpus linguistics is a kind of tool for other branches of linguistics, allowing the necessary research to be carried out in a more convenient environment, providing more accurate results. In addition to linguistic research, the collected corpora are used to compile dictionaries and grammar reference books.

Corpus linguistics distinguishes a set of research methods that try to determine how corpus linguistics has come from experiments to theoretical foundation, trace the path from data to theory. Linguists Wallis and Nelson proposed the so-called three-A perspective in 2001. This refers to the three capital letters of the following concepts: abstract, abstraction and analysis (Wallis SA, Nelson G., 2001).[1]

The first of these terms, annotation (otherwise known as markup), is the application of a specific scheme to texts in which individual elements of the text are marked in a special way. The next stage, abstraction, involves a certain transition from a terminological scheme to a theoretical model. Abstraction often involves linguistic research. Analysis is the stage where statistical methods are used and data is processed.

Linguistic corpora have a number of advantages. In 1992, J. Svartvik[2] highlighted the following advantages of data obtained from a linguistic corpus:

– greater objectivity than data based on self-analysis;

– can be easily verified by other researchers; they can also share the same data instead of always compiling their own;

– necessary for studying differences between dialects, registers and styles;

– provide the frequency of occurrence of linguistic phenomena;

– not only serve as illustrative examples, but also as a theoretical resource;

provide important information for a number of application areas such as language teaching and linguistic technologies (machine translation, speech synthesis, etc.);

– corporations provide the opportunity for full responsibility for linguistic features: the analyst must take into account everything in the data, not just selected functions;

– computerized corporations provide researchers worldwide with access to data;

– ideal for non-native speakers (Svartvik J., 1991).[3]

However, there are several problems in corpus linguistics that scholars are currently seeking solutions to. Among them, it is worth highlighting the problem of representativeness, the problem of labeling and the problem of presenting results. As for the problem of representativeness, the reasons for its occurrence can be seen due to Zipf's law. According to this law, in any linguistic corpus the frequency of any word usage is inversely proportional to its rank in the frequency table. This means that there are many more rarely used words in the language than frequently used ones. Accordingly, the size of the corpus, as well as the selection of texts, are important for an adequate representation of a particular language or sublanguage.

[1] 1. Wallis SA, Nelson G. Knowledge discovery in grammatically analyzed corpora. Data Mining and Knowledge Discovery. 2001. Vol. 5(4). Pp 305- 336.

[2] 2. Svartvik J. (ed.). Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991. Berlin: Mouton de Gruyter, 1992. pp. 487.

[3] 1. Svartvik J. (ed.). Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991. Berlin: Mouton de Gruyter, 1992. pp. 487.

P.V. Sysoev proposed a definition of the linguistic corpus. The author interprets this concept as "an array of texts collected into a single system according to certain characteristics (language, genre, time of creation of the text, author, etc.) and equipped with a search system" (Sysoev P.V., 2010, p. 99) .[4] In other words, a linguistic corpus is a collection of texts in electronic form, selected according to certain characteristics and available for qualitative and quantitative analysis.

*Case technologies*is a set of tools and methods for processing and analyzing data from electronic linguistic corpora.

*Concordance*is one of the main tools in corpus linguistics, which involves using corpus software to search for the occurrence of a particular word or phrase. This tool will be discussed in more detail below.

It is believed that corpus linguistics as a branch of linguistics appeared in the 60s of the 20th century. However, it should be noted that corpus linguistics has a long history and prerequisites for its emergence. Of course, before the invention of computers, the existence of electronic corpora was impossible, but large volumes of texts existed in paper form long before the end of the second millennium AD. The pre-digital period in the development of corpus linguistics dates back several centuries. Initially, non-digital corpora were closely associated with religious texts. An example would be symphonies (or concordances), which are essentially lists of words accompanied by verses from the Bible. Later, large dictionaries appeared, created on the basis of card indexes. This stage in the history of the development of corpus linguistics can be called lexicographic. Already at the turn of the 20th century, collections of texts for linguistic analysis began to appear.

The digital stage in the development of corpus linguistics begins with the advent of first-generation electronic corpora. Its very first representative was the Brown Corps (full original name -"Brown University Standard Corpus of Present-Day American English"), the authors of which are considered to be N. Francis and G. Kuchera. It was this corpus that became the founder of modern corpus linguistics and laid the basic principles for creating electronic text corpora. The corpus consists of a million words of American English texts printed in 1961. To make the corpus balanced, texts were selected in different proportions from 15 different text categories: journalism, everyday texts, religious texts, scientific texts, various types of fiction, etc.

The compilers of the Brown corpus were guided by four criteria for selecting its content:

– *origin of the author and composition of the text*(representatives of American English only);

– *the ability to process data using a computer;*

– *synchronization*(the creation of the corpus began in 1961, so texts were selected from this year of publication);

– *numerical ratio of genre diversity of texts*.

Today this building is considered small and already outdated. However, the housing is still in use. Much of its usefulness lies in the fact that the structure of Brown's corpus has been copied by other corpus compilers. The LOB (Lancaster-Oslo-Bergen; language is British English) corpus and the Kolhapur (Indian English) corpus are two examples of corpora made in accordance with the Brown corpus. Both consist of 1 million words of written language (500 texts of 2000 words each), which were selected in the same 15 categories as the Brown corpus.

The availability of corpora that are so similar in structure is a valuable resource, for example, for researchers interested in comparing different varieties of a language. For a long time, the Brown corpus and the LOB corpus were the only ones available for computer processing. Therefore, many studies in the field of corpus linguistics have been based on these corpora.

London-Lund Corpus (LLC) also belongs to the first generation of corps. Although also small in volume, this corpus was the first corpus of spoken language. It consists of 100 oral texts of about 5000 words each. Texts are divided into different categories such as spontaneous conversation, spontaneous commentary, spontaneous and prepared speech, etc. The texts have a transcription, and they also contain a detailed prosodic analysis of speech.

As for the history of Russian corpus linguistics, there was an attempt to create the so-called Machine Fund of the Russian language. Work on creating the corpus began in 1985 at the Institute of Russian Language of the USSR Academy of Sciences. Unfortunately, the creation of the corpus could not be completed due to funding problems in the early 1990s. However, the Russian language corpus was still created at that time. True, this happened in Sweden in the city of Uppsala. The Uppsala Corpus of the Russian language, created at the Institute of Slavic Studies in Uppsala, contained 1 million word usages and about 600 texts.

The second generation of electronic enclosures comes in the 1990s as computer technology becomes

[3] Sysoev P.V. Linguistic corpus in the methodology of teaching foreign languages // Language and culture. 2010. No. 1. P. 99–111.

more advanced. Among the first representatives of this period is the corpus of The Cobuild Project / The Bank of English (BoE). This is a British monitoring corpus, constantly expanding the volume of word usage. 25 percent of the corpus is spoken language and 75 percent is written language.

One of the most significant corpora that is often used by researchers is the British National Corpus (often used abbreviation in English - BNC). The corpus consists of 100 million words. Like the British Monitoring Corpus, the corpus contains both written and spoken material, but unlike the BoE, the British National Corpus is finite, meaning no more text is added once it is completed. Texts from the British National Corpus have been selected according to carefully defined criteria to ensure a balanced corpus. The texts were coded with markup providing information about the texts, authors, and speakers.

Later, other national corpora were created, including the American National Corpus and the National Corpus of the Russian Language, containing millions of word usages.

Modern linguistic corpora must meet a number of criteria, or parameters. The first of these criteria is representativeness. This means that the corpus must be reliably representative. This is achieved due to the required volume and genre variety of texts. Another important criterion is balance. A balanced corpus has an even distribution of texts from different categories. Knowing the exact volume of the case is also important. This is essential for researchers conducting quantitative research. Electronic form is the fourth parameter. The electronic format for presenting text has greatly facilitated the processing of information. Finally, the fifth and final criterion is markup (another common name is annotation). By markup we mean

"automatically or manually entered linguistic or metatextual information about all selected units of the corpus: text, sentence, text form, morpheme, sound, etc." (Kopotev M.V., 2014, p. 30).[5]

Over time, different types of housings have emerged for different purposes. There are different types of hull classifications. Let's start with the fact that corpora can be distinguished by the language of use. Here, the parameters of monolinguality are taken into account, that is, all the texts of the corpus belong to one language, and multilingualism, where the texts of the corpus are written in two or more languages. Multilingual corpora are usually divided into mixed and parallel. The first include texts that are not translations of each other. Parallel corpora contain original texts and translation texts. The parallel corpus also has the property of alignment, which means that the texts and their translations are connected in sentences and paragraphs. A special type of corpus in this classification is a comparative corpus, in which, in addition to the original text, there are several translation texts. An example of such a corpus is the corpus of Bible translations being developed at the University of Maryland in the USA, numbering several thousand translations.

Based on the type of language data, three types of linguistic corpora can be distinguished. First of all, these are oral, which contain recordings and transcripts of oral speech. The largest number of buildings are classified as written. There are also mixed cases that include both written and oral speech, and, as a rule, a larger share is allocated to written speech than oral speech.

Corpora can be either labeled or unannotated (that is, without labeling). The nature of the marking is also a classifying parameter. Markup can be (1) metatextual, containing a passport of the text (information about the author and text), (2) linguistic, which can be syntactic, semantic, morphological, etc., and (3) extralinguistic, which contains information about gestures and other accompanying nonverbal signs.

Linguistic corpora can also vary in genre diversity. They can cover the whole range of genres - from conversational to scientific. Also, the corpuses differ in volume into representative (otherwise national), monitoring and illustrative. Based on the type of access, linguistic corpora should be divided into open ones and those where access is limited. As a rule, the latter are paid.

According to the representation of linguistic material, linguistic corpora are divided into full-text and n-gram, or fragmented. The essence of the latter is that the text in such corpora is divided into small sections, called grams, for the purpose of ease of working with them.

Some of the tools provided by corpora are widely used in foreign language teaching practice. The most common corpus analysis software is concordance. This tool extracts examples of words or tags (or sequences of words/tags) and presents them to the user.

Concordance is a basic tool in corpus linguistics, which involves using corpus software to find the occurrence of a particular word or phrase. This idea is not new, and many scholars over the years have made, for example, concordances of the Christian Bible by hand, painstakingly finding and recording each example of certain words (symphonies). By using With a

[5] 4. Kopotev M.V. Introduction to Corpus Linguistics. Electronic textbook for students of philological and linguistic specialties at universities. Praha: Animedia, 2014.

computer we can now search millions of words in a matter of seconds. The search word or phrase is often called a "node", and concordance lines are usually represented by a node word/phrase in the center of a line with seven or eight words on either side. These are known as Key-Word-In-Context (KWIC) displays. Concordance lines are usually scanned vertically, that is, viewed from top to bottom or bottom to top, focusing on the key word (or phrase), which is located in the center. This may initially seem awkward because we are used to reading from left to right. Lines of concordance allow us to read in a whole new way, vertically or even from the center outwards in both directions.

In conclusion, the corpus is a representative array of unedited texts, presented in electronic form, as a rule, marked up for analysis for linguistic purposes, provided by a relatively easy–to-use search engine, representing as many "variants" of the language as possible.Today, we consciously limit ourselves to various limits when studying texts of a certain language, which calls into question the objectivity of this kind of research. With the advent of electronic corpora, the variety of forms of language existence has become more evident, and the possibilities of studying language data have expanded. The modern linguistic corpus contains hundreds of millions of word- uses, and the fact that with the help of an electronic corpus, the results of examples of word usage can be obtained, greatly simplifies a task for linguists.

REFERENCE:

1. Wallis SA, Nelson G. Knowledge discovery in grammatically analyzed corpora. Data Mining and Knowledge Discovery. 2001. Vol. 5(4). Pp 305- 336.
2. Svartvik J. (ed.). Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991. Berlin: Mouton de Gruyter, 1992. pp. 487.
3. Sysoev P.V. Linguistic corpus in the methodology of teaching foreign languages // Language and culture. 2010. No. 1. P. 99–111.
4. Kopotev M.V. Introduction to Corpus Linguistics. Electronic textbook for students of philological and linguistic specialties at universities. Praha: Animedia, 2014.
5. Common European Framework of Reference for languages: Learning, teaching, assessment. Cambridge University Press, Cambridge, 2001.