THE HISTORY OF CORPORA MAKING

Parpieva Shahnoza Muratovna,
Uzbekistan state world languages university, second year master's student

Abstract

The development of corpus linguistics as well as the construction of corpora is one of the current issues of modern linguistics. Today, the use of corpora plays a leading role in most linguistic research. Although back in the first half of the twentieth century, corpus making was only possible by hand. It was time-consuming, labor-intensive and expensive. Therefore the use of corpora was minimal and only when a large number of researchers were interested in it. But thanks to the development of electronic technologies the use of corpora is now possible everywhere.

Keywords:

History of corpus linguistics, text corpora, corpus linguistics, pre-electronic corpora, earliest electronic corpora.

The term corpus linguistics was first used in 1977, which makes corpus linguistics a very new scientific field. However, during this short time, corpus linguistics as one of the approaches of computational linguistics has managed to become one of the leading areas of modern linguistics in general. The emergence of corpus linguistics was preceded by a centuries old period of the use corpus methods and the creation of text corpora. In connection with the non-electronic form of storage of these corpora, as well as non-automatic methods of data processing, a special period in the history of the development of corpus linguistics called pre-electronic can be distinguished.

Many pre-electronic corpora were associated with the sacred writings of various religions and the most researched of them was the corpus of Biblical texts. The Bible-based word lists with verses are called concordances (symphonies). After the publication of the first edition of the concordance by A.Cruden in 1737, concordances to the works of great writers began to be compiled according to the this principle. Thus, an important work for the development of corpus linguistics was the "Concordance to the works of W. Shakespeare in all editions" (1787) by E. Beckett. There are also known concordances to the works of W. Shakespeare, compiled by M. Cowden-Clarke (1847) and S. Ayscough (1790). The corpora of the XVIII-XIX centuries were mostly dictionaries and the period itself was characterized by the development of lexicography.

In the modern view the pre-electronic era corpora are collection of texts or an archive, they do not have a single system for collecting texts, their volume and sources vary greatly. The same features are characteristic of the concordances of that time. In the pre-electronic era, the foundations and principles of forming concordances were laid. By the end of the pre-electronic era, the terms "concordance", "keywords in context" and "lemmatization" already existed.

At the end of the XIX century and at the beginning of the XX century, the corpora began to be created for the purpose of conducting linguistic research or – more often – for solving practical problems (for example, for calculating the frequency of language units).

With the invention and widespread use of computers, a new stage of development corpus linguistics begins – the created corpora differ from the old ones not only in the storage format, but also in volume. The first computerized corpus – The Brown Corpus – included 500 texts from American books, newspapers and magazines, was first published in the United States in 1961. The Brown corpus was named after the US University The Brown University, located in Rhode Island. Its name officially included the term "corpus". A group of scientists led by H. Kuchera and W. Francis worked on the creation of the corpus in the period from 1961 to 1964. Each text in the Brown Corpus has a length of 2000 words and the entire collection includes 1 million words (500 texts of 2000 words each). The authors Francis W. and H. Kucera accompanied it with a large number of materials of primary statistical processing: a frequency and alphabetical-frequency dictionary, a variety of statistical distributions. The corpus contains the following fifteen genres of

written texts of the American English: newspaper articles, scientific works, ads, hobby books, religious literature, biography, essays, fiction (detective stories, adventures and westerns, popular science literature, romance novels, feuilletons). The purpose of the Brown Corpus is to provide a systematic study of individual genres of written English and a comparison of genres. Its appearance aroused general interest and lively discussions.

Later, European researchers compiled a corpus of texts first published in the UK in 1961, following the same principles: 15 genres(registers), 500 texts of 2000 words (word usage). It included 1 million words of the British English and it was called the Lancaster-Oslo-Bergen Corpus, after the names of the British and Norwegian Universities or LOB in short. Balanced Brownian-type corpora was very important for researchers whose interests lie in the field of linguistics and using the corpus for the purposes of linguistic description and analysis.

Therefore, the two earliest large corpora are the written corpora of the American and British English. Both corpora remain useful today and numerous studies of the English language are based on them. In the decades since the creation of these cases, computers have become cheaper and much more powerful, in addition, inexpensive and reliable scanners have made it optional to type texts on a computer using a keyboard. These inventions have made it easier to create corpora and the latest ones already contain billions of words (word usage).

References

1. McEnery T., Wilson A. Corpus Linguistics. Edinburgh, 1997.
2. Meyer, Ch.F. Pre-electronic corpora. Corpus Linguistics: An International Handbook. Walter de Gruyter, 2008.
3. McCarthy, M. & O'Keeffe, A. Historical perspective: What are corpora and how have they evolved? In: O'Keeffe, A. & McCarthy, M. (eds) The Routledge Handbook of Corpus Linguistics. Routledge, 2010.
4. Kennedy, G. An Introduction to Corpus linguistics. Addison Wesley Longman limited, 1998.