ENGLISH CONTEMPORARY CORPORA

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

A linguistic corpus is a collection of texts created in accordance with certain principles (machine-readable, unified, structured, marked up, philologically competent) and provided with a specialized search engine. The linguistic corpus can include both written texts (newspaper, magazine articles and literary works) and spoken transcripts of radio and television programs. Depending on the purpose of its creation, the corpus may include texts of one or more authors of different literary genres, written in a certain historical period. The entire array of texts in the corpus is systematized. The corpus records the order of each word in the sentence in relation to other words and also takes into account the frequency of its use in this corpus.

The largest number of corpora is based on the English language, which can be explained by its prevalence, as well as the high level of development of corpus linguistics in the United States and in Great Britain. Among the first national corpora was the British National Corpus (BNC), which is considered as a reference today, since most modern linguistic corpora were created on its model. This corpus was developed at the Oxford University with the participation of the Lancaster University, the British Library, the Oxford University Press, Longman and W. & R. Chambers from 1991 to 1994. The volume of the corpus is more than 100 million words, 90 percent of which correspond to written texts, 10 percent – to spoken ones [1]. BNC includes both metatext and morphological markup and is a general-purpose synchronous corpus. This corpus shows the state of British English in the late 20th and early 21st centuries. The Corpus is characterized by the use of full texts moreover, including a wide variety of texts by genre, style and subject (newspaper articles, magazine texts, letters, school essays and etc.).

Corpus of Contemporary American English (COCA) is the largest (560 million words) corpus of American English that includes a wide variety of texts of various genres. The corpus was created by Mark Davis, Professor of Corpus linguistics at Brigham Young University from 2000 to 2003. [2] COCA is a mixed-type corpus, since it contains both written texts (fiction, popular magazines, newspapers, scientific literature and etc.) and spoken language. The corpus contains 560 million words and includes texts from 1990 to the present. The search interface provides a wide range of features: search for words, phrases, lemmas, grammatical forms, synonymous series and etc. The corpus is updated twice a year and is convenient for tracking the dynamics of linguistic changes. This is the most widely used structured text corpus, with approximately 10,000 users every month.

The Oxford English Corpus is the largest corpus ever created. It contains over 2 billion words and reflects the state of the modern English in the entire area of its distribution. The body contains the texts which are created since 2000, the bulk of the materials are placed on the World Wide Web. The written texts are taken from literary novels, academic journals, daily newspapers and magazines, moreover, email letters, texts from social media and web blogs. The corpus is available only to Oxford University Press researchers and utilized for compiling dictionaries.

The Bank of English is an integral part of one of the largest language databases - Collins Corpus, which is used to create modern dictionaries. This corpus contains over 650 million words, 65-70% of which correspond to the British English, 25-30% – to the American English. The corpus consists of various types of written texts and spoken language. The corpus includes metatext markup, as well as partial markup. The Bank of the English presents a unique in its kind monitor corpus of the English language. Regular updating the corpus with new texts gives the ability to track all changes of English lexical systems, such as the emergence of new words, changing the value of existing lexemes, frequency of use and grammatical structures in speech and etc. The distinctive feature of this corpus is a comprehensive reflection of modern English, i.e. the Bank of English covers the English language in general, in proportion to all its variants. Access to the full hull version is chargeable. A free trial is available a one-month subscription to Collins Wordbanks Online for access (550 million words).

The American National Corpus (ANC) is a corpus of American English consisting 22 million words from written and spoken sources published since 1990. The corpus is modeled on the British National corpus (BNC) and intended to reflect the American version of modern English. The corpus creates the most complete picture of American English for further research in linguistics, education, lexicography and

technology. Currently, the corpus is constantly being updated and when completed, this figure will increase to 100 million words. Words are arranged by frequency of occurrence in speech-from the most frequent to the rarest. The project is implemented by an organization called the Linguistic Data Consortium. This consortium, founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), includes universities, companies and government research laboratories. The organization is led by the University of Pennsylvania. The corpus is available for non-commercial use for a fee and part of it, 15 million words - available for free download from the website.

The Michigan Corpus of Academic Spoken English (MICASE) contains approximately 1.8 million words of transcribed speech, obtained from various sources (lectures, discussions, seminars, interviews, student presentations, thesis defense). The corpus includes English speech from and information about the speaker is given in the transcription name. All transcriptions are written in spelling form and do not contain markings. The characteristics of the speaker include: academic role (teacher, graduate, student, doctor, researcher, etc.), native language (English - native language, English - non-native language, American English, other variants of English), native language (with non-native English). Transcription attributes include: type of event (consultation, colloquium, thesis, interviews, etc.), university unit (humanities and arts, biology and health, etc.), academic discipline, academic level of the participant, level of interactivity of the event (monologue, discussion).

References

1. British National Corpus [Electronic resource] / BNC. — Mode of access: https://www.english-corpora.org/bnc/.
2. Kauhanen, Henri The Corpus of Contemporary American English: Background and history. http://www.helsinki.fi/varieng/CoRD/corpora/COCA/index.html