ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 4| April 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

ENGLISH CORPORA MAKING: HISTORICAL OVERVIEW

Teshabaeva Dilfuza Muminovna[1], Parpieva Shakhnoza Muratovna[2]
[1]DSc, Professor at Uzbekistan State World Languages University
[2]Teacher at Uzbekistan State World Languages University

ABSTRACT
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 corpus linguistics called pre-electronic can be distinguished. 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. Second-generation corpora are the products of the Internet and are distinguished by their large size. The third generation corpora are large and have many technological advantages. In this period, a number of new corpora were created, with a total volume of several billion words.
KEY WORDS: ***corpus, concordance, pre-electronic era, computer, generation, modern, megacorpus.***

Swedish writer and linguist J.Svartvik declares that in the history of corpora making there was the first so-called "Stone Age" or pre-computer period, when corpora were created by hand on paper[1]. These first paper corpora were essentially concordances, that is, alphabetical lists of words in their contextual surroundings. The creation of such paper corpus concordances was time-consuming and required strenuous analysis, which was done by hand. Paper corpora played a significant role in linguistic projects such as the concordance of the Bible and literary works as well as the writing of grammars and dictionaries.

The first concordance was compiled in the thirteenth century by Friar Anthony of Padua for the fifth-century Latin version of the "Vulgate" Bible. This concordance was called "Concordantiae Morales".

The first important work among English-language concordances was "A Concordance to Shakespeare: Suited to All the Editions" by A. Beckett (1787) and is an important work in the development of corpus linguistics. It contains, information about the place of use of a particular word (play, act and action) and passage of a work in which a certain word is used[2].

The tradition of composing concordances by hand for fiction works was preserved until 1995 and was implemented in the following works: "The Concordance to Conrad's The Secret Agent by Conrad" (Bender, 1979), "A Concordance to Henry James's Daisy Miller by Henry James" (Bender, 1987), the "A Concordance to the Complete Poems and Plays by T. S. Elliot" (Dawson, 1995)[3].

In addition to concordances, large samples of texts were also used to create early grammar books. Early grammar of the English language was also based on the classical tradition to use quotes from real texts, for instance, "A Short Introduction to English Gramma" by Robert Lowes (1762). One of the most famous grammars of this period was the seven-volume work of Jespersen (1909-1949), "A Modern English Grammar on Historical Principles" and was also built exclusively on examples selected from a huge number of texts. Otto Jespersen belonged to that type of linguists who were convinced, that the linguistic description should be based not on fictional, but on real examples from real speech texts. The tendency to cite literary works as examples with grammatical rules continued in grammars of the late 19th and mid-20th centuries by such authors as J. Ruhl, H. Poutsma, and Ch. Fries. However, not all grammarians followed this tradition.

Studies of large texts have also been carried out with a view to dictionaries. Starting with Samuel Johnson's Dictionary (1755), lexicographers used quotations from texts by famous writers to illustrate the meanings and usage of words. The lexicographer collected 150,000 illustrative quotations for the 40000 words dictionary. The Oxford English Dictionary (OED), which was created under the direction of James Murray (1880), was based on a corpus of five million card quotes[4].

[1] Svartvik, J. Corpus linguistics 25+ years on / J.Svartvik. – Amsterdam, NY 2007. – P. 11-27
[2] Beckett, A. A Concordance to Shakespeare: Suited to all the Editions. Printed for G.G.J. and J. Robinson, 1787. 167-183 p.
[3] Tribble C. What are concordances and how are they used // The Routledge handbook of corpus linguistics / ed. by A. O'Keeffe, M. McCarthy. 2010. P. 167–183.
[4] O'Keeffe, A. & McCarthy, M.'Historical perspective: What are corpora and how have they evolved?'. The Routledge

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 4| April 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

At the turn of the 19th and 20th centuries, several projects were organized to collect empirical material for lexicographic purposes. On their basis, the dictionary of the American version of the English edited by Noah Webster (Noah Webster's an American English Dictionary) (1828) and The Oxford English Dictionary (OED) (1884) were compiled. To create the research base of the Oxford Dictionary, two thousand volunteer readers collected about five million citations totaling approximately 50 million word uses in order to illustrate the meanings and usage of 414,825 words in the dictionary. Based on the collected texts of English dialect speech, J.Wright compiled a dictionary of English dialects "The English Dialect Dictionary" (1898-1905).

The most important pre-electronic corpus is considered to be The Survey of English Usage, created by Randolph Quirk in 1959. The corpus was a large database on cardboard cards containing samples of speech (both written and spoken) of ordinary citizens. This project was a transitional stage in the development of corpus linguistics. R. Quirk called the collected research material "source material" or "texts". This corpus proved to be the most well-structured and systematic corpus of the pre-electronic era. The spoken and written forms of speech were represented by texts of different genres, with both formal and informal communication as sources. The corpus consisted of 200 text fragments, each with a volume of 5,000 word uses. This corpus marked the transition from the pre-electronic to the electronic era.

THE FIRST GENERATION CORPORA

The idea of creating a corpus (already in the modern sense of the word) emerged in the 1960s, heavily influenced by empirical research. By the end of the 1960s, there were several small corpora created on different principles. It was advances in computer technology, rather than in linguistics, that gave rise to the first electronic corpora. J. Svartvik claims that in 1960 the term "corpus" was hardly ever used and there was a long debate about the plural form of the word "corpus" (corpuses, corpora or even corpi) at the conference[5].

Computers were just coming into general use in the mid-twentieth century. They were the first primitive machines which were difficult to work with, but their huge potential was immediately recognised and attracted to linguistic research. The computerisation of texts started with Father Busa's Index Thomisticus before 1950 (completed in 1978). The first linguistic corpora of machine-readable texts appeared in the 1960s. They were very small by modern standards, but were characterised by an elaborate organisation.

THE BROWN CORPUS

In the early 1960s, two projects emerged in Scotland and in the USA in order to create corpora in electronic format. The University of Edinburgh in Scotland was creating a spoken corpus that included transcribed versions of everyday conversations of British English speakers. This corpus is small in size at 300,000 words.The reason was the time-consuming process of collecting and transcribing spoken language and the absence of a computer at the university.

At the same time at Brown University (USA), Henry Kucera and W. Nelson Francis started creation of a one-million-word corpus, which was named the Brown Corpus. The purpose of the corpus was to investigate the linguistic features of the American English. It contained 500 text passages of 2,000 words each, for a total of about 1 million words. The texts were selected from the fifteen largest genres: newspaper articles (reports, editorials), religious literature, professional literature, popular science literature, fiction, samples of business prose (including government documents), scientific literature, prose fiction, detectives and science fiction, adventure and westerns, romance, humorous stories and novels.

The compilers took into account such characteristics as:

1. The origin and composition of the text (the author had to be a native speaker of American English);
2. Time period (all the texts selected for the corpus were first published in 1961);
3. Balanced representation of different genres;
4. Accessibility to computer processing (special markings to convey graphic features of the text).

The emergence of the Brown corpus sparked interest in the academic community. The corpus quickly became a popular object of linguistic research. Gradually, in the process of its use, scholars came to the realisation that it was possible to make certain comparisons and identify specific patterns only by analysing significant size arrays of texts according to certain rules.

Thus, new studies of language began to be carried out at a higher and more reliable level within the framework of a new trend in linguistics, which is corpus linguistics.

The Brown Corpus has become the standard for corpus compilation, both in terms of volume and the range of writing styles and genres represented in it. With the publication of the Brown Corpus, similar corpora began to appear, first in the UK and then in other countries. For example, in 1976, The Lancaster-Oslo-Bergen corpus (LOB) (1961-1978) was published[6]. In the early 1990s, similar corpora with a volume of at least one million words, consisting of 500 texts of fifteen different genres of writing, began to be created. At the same time, each text had to contain at least 2,000 words. These include, for example, The Australian Corpus of English, ACE (1986), The Wellington Written English, WWE (1986), The Freiburg-Brown Corpus of American English, (1991-1992), The Freiburg London-Oslo / Bergen corpus, F-LOB, (1991-1992), The Kolhapur Corpus Indian English (1978)[7]. These corpora were collectively called

Handbook of Corpus Linguistics. London: Routledge, 2010. 3-13 p.

[5] Svartvik J. Corpus linguistics 25+ years // Corpus Linguistics 25 Years On / ed. by R. Faccinetti, 2007. 11–27 p.

[6] The LOB Corpus. URL: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html

[7] Baker P., Hardie A., McEnery T. Glossary of Corpus Linguistics. Edinburgh University Press, 2006. 192 p.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 4| April 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

the Brown family of corpora[8]. The only difference between these corpora was that the corpora contained texts of one of the variants of written English: American, British, Australian, New Zealand and Indian.

The Lancaster-Oslo-Bergen Corpus

The Lancaster-Oslo-Bergen Corpus created on the Brown Corpus model following the same principles: 15 genres (registers), 500 texts of 2000 words (word uses). It included 1 million words of British English and was called The Lancaster-Oslo-Bergen Corpus (from the names of the British and two Norwegian universities, or LOB for short). Balanced corpora such as created on Brown corpus model are very important for researchers whose interests lie in the field of linguistics and who wish to use the corpus for purposes of linguistic description and analysis.

The London-Lund Corpus

In 1975 the London-Lund Corpus (LLC) - a corpus of spoken English was completed. The project was a collaboration, funded by IBM, between the Unit for Computer Research on the English Language (UCREL) at the University of Lancaster and the IBM Scientific Centre in Winchester. It contained about 500,000 words with spelling, phonetic and prosodic transcriptions. This huge work was first done on paper by staff at University College London and then transferred to computer form by linguists from the Swedish city of Lund. The LLC corpus consists of 100 transcribed texts of spoken monological and dialogical speech of 5000 words each. Dialogical speech is recorded in texts of conversational style between friends and colleagues, in talks and telephone conversations. Monological speech is represented by spontaneous (comments and stories) as well as prepared speech.

The Second Generation Corpora

Second-generation corpora are products of the Internet and are characterised by significant volume. For example, in the late 1980s the first mega-corpus was created in the UK, setting a new standard for corpus - the British National Corpus. This corpus is characterised by a volume of 100 million words, with the availability of mark-up and access via the Internet. The corpus is distinguished 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.).

The Bank of English

Many European language corpora have been created according to the standards set by the British National Corpus. The Bank of English project began to develop in the 1980s. In 1989 it had 20 million words and in 2012 it had 650 million words.

The distinctive feature of this corpus is a comprehensive reflection of modern English, it covers the English language in general, in proportion to all its variants. 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. 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 COLLINS BIRMINGHAM UNIVERSITY INTERNATIONAL LANGUAGE DATABASE

Then the Collins Birmingham University International Language Database (hereafter-COBUILD) came into existence. This corpus became the basis for the dictionaries and a number of English grammar books. The corpus was created by a team of scholars led by John Sinclair. The project uses the so-called Birmingham Collection of Texts (The Birmingham Collection of Texts), which includes 20 million uses of written and spoken texts. The main corpus contains 7.3 million words, and the so-called "reserve corpus" 13 million. The corpus consists of 75% of the written texts, 25% of the spoken texts. The COBUILD corpus contains texts published between the 1960s and 1982. The written speech consists mainly of prose fiction texts. The corpus captures oral codified speech, which uses only common non-special vocabulary. 75% of the spoken speech is the speech of men over 16 years old, 25% is the speech of women. 20% of the corpus consists of texts of the American English. According to Johansson, the COBUILD project was a breakthrough for its time for a number of reasons:

1) the corpus exceeded 20 million word uses;
2) the sources were full texts rather than short fragments;
3) it was the most representative and included spoken and written texts of various genres. COBUILD became the most voluminous corpus of its time and formed the basis of the Collins English Dictionary, the Collins COBUILD Dictionary of English (1987)[9].

The Longman Corpus Network

Another megacorpus that was created in the late 1980s by a group led by D. Summers in the Longman Publishing House, is the Longman Corpus Network. This corpus network is now a commercial database consisting of five main corpora:

1) The Longman Corpus of Learners' English (10 million word uses);

2) The Longman Written American Corpus (100 million word uses);

[8] Xiao R. Well-known and influential corpora // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. 383–457 p.

[9] The history of COBUILD. URL: https://www.collinsdictionary.com/cobuild/

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 4| April 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

3) The Longman Spoken American Corpus (5 million word uses);

4) The Longman / Lancaster English Language Corpus (30 million word uses)

5) The Spoken British Corpus (10 million word uses)[10].

Kennedy writes, that although each part of the Longman Corpus Network was set up for a specific purpose, the combined corpus became a powerful tool, recording a large variety of texts of different genres and speech produced by native and non-native speakers of English.

This type of corpus has been used to create dictionaries and textbooks on communicative English grammar. Later, the spoken English corpus was included into the spoken part of the British National Corpus[11].

The International Corpus of English

The International Corpus of English (ICE) was developed at University College London under the direction of Sidney Greenbaum in 1996. The aim of the project was to collect texts of regional variants of English. The sub-corpora include spoken and written texts of regional variants of English: Britain (ICE-GB), East Africa, India, New Zealand, Singapore, Canada, Hong Kong, Jamaica, Philippines, USA, Cameroon, Fiji, Ireland, Kenya, Malta, Malaysia, Pakistan, Sierra Leone, Sri Lanka, Trinidad and Tobago.

All sub-corpora contain 60% written texts and 40% spoken transcripts. The dialogic speech subcorpus includes the following genres of spoken language: private conversations (face-to-face and telephone conversations) and public conversations (lessons, radio and television talks, TV and radio interviews, parliamentary debates, business discussions, face to face meetings.

The sub-corpus of monological speech is divided into two parts. The first includes statements of spontaneous speech (commentary, speech at demonstrations and in court). The second part contains prepared read-alouds (TV and radio news, TV and radio talks (talk shows).

Each subcorpus includes written texts of different types and recordings of oral speech. Currently, the British component of the corpus (ICE-GB) is fully prepared and its texts are provided with morphological and syntactic markup. The volume of each national subcorpus is 1 million words. The British component of the corpus is distributed on disk on a fee basis and a small fragment of it (20 thousand copies) is freely available.

The Michigan Corpus of Academic Spoken English

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 native speakers' speech moreover, information about the speaker is given in the transcription name. All transcriptions are written in spelling correct form and do not contain markings. Corpus is publicly available and allows transcriptions of individual records to be searched for by transcription and speaker parameters. 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). In addition, it is possible to search for specific words and collocations by selecting the parameters.

The Third Generation Corpora

The trend towards compiling larger corpora continued even after the 2000s. A. Mauranen, S. Kubler and H. Zinsmeister describe this generation by the slogan "the bigger the corpus, the better"[12], and L. Flowerdew is the first to call this period the generation of giant corpus[13]. At this time, a number of new corpora emerged (COCA, Google Books Ngram), with the volume reaching several billion words. The large volume of corpora made it possible to carry out frequency studies on a larger scale and to study collocations consisting of three, four or more words.

The early 2010s were marked by the emergence of great technical possibilities: the fourth generation of BNCweb (2009), CQPweb (2012), SketchEngine (2013), Wmatrix (2013), functionally similar to the third generation of concordancers, were developed. Fourth-generation concordancers have been developed to address the following issues: limited PC power, incompatibility of PC operating systems and legal restrictions on the distribution of enclosures. To solve the legal issues and simplify the access procedure, the enclosures moved to online versions, which increased the speed of processing requests and increased the number of users.

Direct access became available through a web browser equipped with an Online search engine. The fourth generation of concordancers works online and allows a contrasting analysis of a small private corpus with BNC corpus or texts from Internet. M. Davies calls fourth-generation concordancers hybrid corpora, as their interface is

[10] Leech G. A brief users' guide to the grammatical tagging of the British National Corpus. URL: http://www.natcorp.ox.ac.uk/docs/gramtag.html

[11] The British National Corpus. URL: http://www.natcorp.ox.ac.uk

[12] Kuebler, S. & Zinsmeister, H. Corpus Linguistics and Linguistically Annotated Corpora. London: Bloomsbury Publishing, 2005.

[13] . Flowerdew, L. The argument for using English specialized corpora to understand academic and professional language. In: Connor, U. & Upton, T. (eds) Discourse in the Professions: Perspectives from Corpus Linguistics. Amsterdam: Benjamins, 2004. 11–33 p.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 4| April 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

a kind of common field for corpus creation and frequency analysis on morphemic, lexical, syntactic and phrase levels[14].

REFERENCES

1. *Svartvik, J. Corpus linguistics 25+ years on / J.Svartvik. – Amsterdam, NY 2007. – P. 11-27*
2. *Beckett, A. A Concordance to Shakespeare: Suited to all the Editions. Printed for G.G.J. and J. Robinson, 1787. 167-183 p.*
3. *Tribble C. What are concordances and how are they used // The Routledge handbook of corpus linguistics / ed. by A. O'Keeffe, M. McCarthy. 2010. P. 167–183.*
4. *O'Keeffe, A. & McCarthy, M.'Historical perspective: What are corpora and how have they evolved?'. The Routledge Handbook of Corpus Linguistics. London: Routledge, 2010. 3-13 p.*
5. *Svartvik J. Corpus linguistics 25+ years // Corpus Linguistics 25 Years On / ed. by R. Faccinetti, 2007. 11–27 p.*
6. *The LOB Corpus. URL: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html*
7. *Baker P., Hardie A., McEnery T. Glossary of Corpus Linguistics. Edinburgh University Press, 2006. 192 p.*
8. *Xiao R. Well-known and influential corpora // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. 383–457 p.*
9. *The history of COBUILD. URL: https://www.collinsdictionary.com/cobuild/*
10. *Leech G. A brief users' guide to the grammatical tagging of the British National Corpus. URL: http://www.natcorp.ox.ac.uk/docs/gramtag.html*
11. *The British National Corpus. URL:*
12. *http://www.natcorp.ox.ac.uk*
13. *Kuebler, S. & Zinsmeister, H. Corpus Linguistics and Linguistically Annotated Corpora. London: Bloomsbury Publishing, 2005.*
14. *Flowerdew, L. The argument for using English specialized corpora to understand academic and professional language. In: Connor, U. & Upton, T. (eds) Discourse in the Professions: Perspectives from Corpus Linguistics. Amsterdam: Benjamins, 2004. 11–33 p.*
15. *Davies, M. Corpora: an introduction. In: Biber, D. & Reppen, R. (eds) The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press. 2015. pp. 11–31.*

[14] Davies, M. Corpora: an introduction. In: Biber, D. & Reppen, R. (eds) The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press. 2015. pp. 11–31.