ЖАС ҒАЛЫМДАР
МОЛОДЫЕ АВТОРЫ

UDC 81'374

D. Myrzakhat
Magistrant of a I course of Kazakh National University of the name al-Farabi,
Almaty, Kazakhstan
Scientific leader – D. f. n. professor A. H. Azamatova
e-mail: dana_2408@mail.ru

Electronic corpora as a basis of modern English dictionaries

The article is devoted to analysis of the various aspects of correlation between two linguistic disciplines: corpus linguistics and lexicography. It describes the current state of electronic corpora, provides a brief overview of the development stages of English lexicography. The article details the process of compiling major English corpora and defines their main functions. Particular attention is paid to determine the value of electronic corpora when creating dictionaries, the positive and negative impact of corpora on the development of lexicography.
Key words: electronic corpus, lexicography, corpus linguistics, text, dictionary.

Электронды корпус заманауи ағылшын сөздіктерінің негізі ретінде

Мақаланың негізгі мақсаты – корпустық лингвистика мен лексикография салаларының арақатынасының барлық аспектілерін талқылау. Бұл мақалада электронды корпус және ағылшын лексикографиясы салаларына сипаттама беріліп, олардың тарихы мен даму кезеңдеріне қысқаша шолу жасалынады. Ағылшын тілінің ірі корпустарының құрастырылу үдерісі қарастырылып, корпустың негізгі қызметтері айқындалады. Сондай-ақ электронды корпустардың сөздік жасауға қандай пайдасы бар екендігі анықталады. Корпустың лексикография саласына тигізетін әсерінің артықшылықтары мен кемшіліктері мысалдар арқылы нақты дәлелденеді.
Түйін сөздер: электрондық корпус, лексикография, корпустық лингвистика, мәтін, сөздік.

Д. Ы. Мырзахат
Электронный корпус как основа словарей современного английского языка

Данная статья посвящена анализу различных аспектов соотношения двух лингвистических дисциплин: корпусной лингвистики и лексикографии. В ней описывается современное состояние электронных корпусов, дается краткий обзор этапов развития лексикографии английского языка. В статье подробно рассматривается процесс создания крупных английских корпусов и выделяются их основные функции. Особое внимание уделяется определению значения электронных корпусов при создании словарей, положительное и негативное влияние корпусов на развитие лексикографии.
Ключевые слова: электронный корпус, лексикография, корпусная лингвистика, текст, словарь.

Corpus linguistics plays an important role in compiling, writing and revising dictionaries, as within a few seconds the linguist can get examples of words or phrases from millions of spoken and written texts. And since corpora continue to grow and are constantly being expanded with new texts, lexicographers have an instant access to up-to-date information.

The practice of dictionary-making began in 1600s when Robert Cawdrey involved words that were considered difficult as they were borrowed from another language into his version of the dictionary [1, 49]. The words which came from Latin-English dictionaries and other available

texts of the time were given brief definitions, synonym and a fixed form. It was Samuel Johnson who distinctly introduced the methods or steps that were taken to create his dictionary in the 1700s and some of the methods were then followed by the committee entrusted to create "A New Dictionary" or currently known as the Oxford English Dictionary in the 1800s.

A corpus is a collection of samples of authentic spoken and written text which are used for analysis of words, meanings, grammar and usage [2, 74]. In Saussurian terminology, the text is akin to that of *parole*, while the corpus provides the evidence of *langue*. The term corpus linguistics is used when a corpus is specifically used to study a language. Lindquist distinguishes the term with other branches of linguistics such as sociolinguistics (the study of language and society), or psycholinguistics (the study of language and the mind) in that corpus linguistics is a specific method used in language study, the "how to" rather than the "what". In other words, corpus linguistics is an approach rather than a specific field of language study [3, 38].

In 1950s, there was a growing dissatisfaction of how language could not reason out the many 'ungrammatical' patterns found in English. There was a strong call for empirical, real language data [4, 46]. It was then that corpus was invented. The first corpus was made out of a survey of English usage conducted by two universities, University of London and the Brown University Corpus in Providence. In the 1960s, both compiled its million word corpus of *written text* from 500 reading passages, which was named Brown Corpus. This American corpus was a landmark in corpus linguistics since it was the first corpus to employ a computer in its making. In 1982, the British version of the corpus, named the LOB corpus was compiled by Hofland and Johansson. LOB is an abbreviation from The Lancaster-Oslo-and Bergen, and as its name suggests it is a collaborative attempt between the three universities: the University of Lancaster, the University of Oslo, and the University of Norwegian Computing Centre of the Humanities.

However, both the Brown corpus and LOB corpus were deemed to be inadequate to sample English vocabulary. This gave birth to John Sinclair's English Lexical Studies which specifically aimed to investigate vocabulary using an electronic text of spoken and written language. The study gave prominence to collocation - words that naturally co-occur together. Aimed to represent varieties of English where it is used as a first or second language, Sidney Greenbaum compiled one-million-word corpora called The International Corpus of English in 1988. The unique feature of this corpus is that it samples more spoken language (60%) than its written counterpart (40%).

In the early 1990s, major universities and companies together compiled British National Corpus (BNC) containing 100 million words from 1980 up to 1993. The compilers were Oxford University Press, Longman, Chambers, the British Library, Oxford University and Lancaster University. The aim of the corpus is to provide a balanced corpus that represents British English. The corpus includes 10% spoken language and 90% written language, which comprises of 25% fiction and 75% non-fiction. One big distinction between BNC and Brown is that the former took samples from a longer piece of text between 40,000 and 50,000 words. This gives BNC an added advantage of being representative since text contains a different use of words at the beginning, in the middle, and at the end. Due to its sheer size, representativeness, and care, most British publishers prefer to make use of this corpus as their source of lexicographic information.

Typically, any corpora will need to go through a three-step process in its making. Before going through these three steps, however the writer needs to determine the basic outlines of a corpus such as the size of the corpus, the genre of the corpus, whether it will specifically look into written, spoken language, or both. Sinclair points out that the principles underlying corpus creation should be as large as possible including samples from a broad range of material in order to accomplish one way of representativeness to be anticipated with the technology of the time. The corpus should also be classified into different genres and even size. Once this basic outlines is determined, the three-step process may begin. It starts with collecting the data, spoken and written. It entails gathering a large mass of speech, written texts, obtaining permission, and doing a careful and organized record-keeping. The next step is computerization which entails converting raw spoken or written text into a digital format in a computer. Recording of speech may be painstaking since it needs to be transcribed manually. Another concern with spoken text is the issue of naturalness of the speech; it needs to be recorded in a natural, casual way that resembles how people speak every day in real life, not in a stilted way. Though written records seem to be less painstaking, it also has its problem, mainly the

copyright issue. Still some texts that come from books, magazines, and other written sources need to be retyped since scanning device software that detect and scan words automatically usually contain errors, so many that it's best to avoid using them altogether. The last step is annotating, which involves assigning information such as parts of speech, etymology, for each data. It should be noted that the three mentioned steps need not to be seen as a separate process; they are all closely connected. For example, after gathering recording of speech, it may be best to transcribe it there and then.

Corpus may have given a lot of contributions in language study, but its impact to lexicography did not start until 1989. Together with the advance of computer software, both have since contributed significantly to the development of lexicography. Since everything is automated and recorded in a digital format, lexicographers can now save their time and the tremendous amount of work needed in compiling a dictionary. Typically, a dictionary usually has information on the part of speech, usage, meaning, pronunciation, etymology of a word. Before the advent of corpora, all this information had to be gathered manually; lexicographers needed to do the hard labor of collecting slips of paper containing text that they intend to include in the dictionary. For this reason, it took roughly 50 years to complete *Oxford English Dictionary*, which was later known as *New English Dictionary* [5,124]. With corpora, dictionary makers can now use a large sample of authentic spoken and written text as a source to illustrate how each word in their list is used in real life. The citation used in dictionary comes from real-life discourse. Real contexts also provide accurate, well-defined lexical meanings in the definition of a word in dictionary, which is a huge improvement over the previous dictionary practice where words were defined using an unscientific manner. One huge improvement in dictionary making is the rich information available for words that have many invariant meanings such as *take, go,* and *time,* which tend to be overlooked in the previous dictionary practice [6,86].

Another huge advantage of using corpora in lexicography is that information on word frequency can also be obtained. This way, lexicographers can assign whether a word is among the first 500 most common words, the next 500 and so on. Meyer notes that the most frequent words are functional words such as *the, an, a, and,* and *of* which carry little lexical meaning and the least frequent words are content words such as proper nouns. Gries mentions two kinds of frequency information that lexicographers can obtain from a corpus: frequencies of occurrence of linguistic elements in the so-called *frequency list*, and frequencies of co-occurrence of these linguistic elements in *concordances*. Lindquist defines concordance as "a list of all the contexts in which a word occurs in a particular text". Using a Key Word in Context concordance, words can be retrieved within their surrounding text, and be presented vertically on the screen. Since the information is presented in contexts, lexicographers can easily assign the collocations of each word in their dictionary.

Since corpus is discourse-based, it means that the word appears in haphazard, arbitrary collection of occurrences. Dictionary makers need to check for some contradictions with 'real' meaning. It is thus dangerous to solely depend on corpus. One way to check the word in context is to expand the text by retrieving its original source.

The huge amount of data in the corpus also allows lexicographers to look for new words that occur for the first time in spoken or written text. However, the corpus has to be large enough to glean information on vocabulary items. A small corpus such as LOB corpus which stores roughly one million word items could not give lexicographers enough information on the range of vocabulary items. A monitor corpus is also needed, in which large data of language is pooled from time to time, rather than fixed only in one particular time period. This way, the corpus is frequently updated with new words and meanings in today's growing language.

The first dictionary to be founded wholly on corpus is *Collins COBUILD series of English Language Dictionary* compiled in 1987, guided by John Sinclair. The dictionary has its citation taken from real life discourse, and each word is defined from these authentic texts, instead of relying on previous dictionary. This entails using a very large corpus so that it may be able to include all lemmas including their word senses. However, this presents problem in that there tends to be an exclusion of rare words such as *apothegm.* Besides being the first corpus-based dictionary, COBUILD is innovative in that the definitions are related to a classroom teacher explaining the words. For example in describing the word *junk*, it says: "You can use *junk* to refer to old and second-hand goods that people buy and collect".

In the practice of dictionary-making, one crucial distinction has to be made between corpus-

based dictionary and corpus-driven dictionary. Dictionaries such as *Collins COBUILD series of English Language dictionaries* are said to be corpus-driven if the corpus itself is used to validate information presented in the dictionary. However, if the corpus is used to extract the information used in the dictionary, it is called corpus-driven. Teubert suggests that dictionary should follow corpus-driven approach so that it may complement standard linguistics and not just extend it.

The role of the computer has a clerical role in lexicography which reducing the labor of sorting and filing and examining very large amounts of English in a short time [7, 39]. From simple tools, it has evolved to a substantial progress together with crucial, profound and basic linguistic generalizations. By these kinds of developed tools, they have revealed many topics for inquiry which have not been well explored by traditional linguistic methods.

In the modern era, the word has been reserved for collections of texts that are stored and accessed electronically. Electronic corpora are usually larger than the paper-based collections which are basically small, previously used to study the aspect of language [8,120]. This is due to the capacity of computers that can store and process large amount of information compared to the previous time.

Together with the technological advance in computer, corpus provided a significant contribution to the development of dictionary making. Corpus linguistics made such a huge impact in dictionary-making: it significantly reduces the time and the heavy work it needs to compile a dictionary since everything is automated and computerized; each dictionary now resembles how language is used in real world; frequency of each word in the list can be assigned; much more information can be given to words with a lot of variant meanings such as *go*, and *take;* it makes it easy to include collocation because words appear in its surrounding text; it can quickly take 'new' everyday words into the system.

References

1 Siemens R. G., Cawdrey R. A Table Alphabetical of Hard Usual English Words. – 3-rd pub. – Toronto: University of Toronto Library, 1994. – 342 p.

2 David C. An Encyclopedic Dictionary of Language and Languages. – Oxford: Oxford University Press, 1996. – 433 p.

3 Gries S.T. What is Corpus Linguistics? // Language and Linguistics Compass. – 2009. – № 3. – P. 1-14.

4 Teubert W. Language and corpus linguistics. Lexicology and Corpus Linguistics. – London: Continuum, 2004. – 276 p.

5 Meyer C.F. English Corpus Linguistics. – Cambridge: Cambridge University Press, 2002. – 257 p.

6 Lindquist H. Corpus Linguistics and the Description of English. – 6th ed. – Edinburgh: Edinburgh University Press, 2009. – 318 p.

7 Sinclair J. Corpus,Concordance,Collocation. – Oxford: Oxford University Press, 1991. – 245 p.

8 Hunston S. Corpora in Applied Linguistics. – UK : Cambridge University Press, 2002. – 373 p.