DOI: https://doi.org/10.32839/2304-5809/2020-9-85-13
UDC 81/33

Kubrak Diana-Olha
Lviv Polytechnic National University

CORPUS LINGUISTICS AND CORPUS ANALYSIS

Summary. The article addresses the features of the linguistic corpus and corpus analysis, as well as their classification, providing a more clear perspective of the corpus studies. Language corpora research has been gaining momentum in the last two decades thanks to the efforts of computational linguists, speech technologists, and linguists who work with authentic language resources. The research outputs stemming from the corpus and its characteristics on compiling and analyzing most known definitions of corpus linguistics and corpus analysis. / In this aspect we dwell upon the corpus-based technology that has been widely used in people's lives although it is still a strange concept for many. Theoretical problems of corpus linguistics are studied by such Ukrainian scientists: O. Demska (Kulchytska), V. Shyrokova, E. Karpilovska, N. Darchuk, V. Zhukovska, V. Balog, S. Buk et al.

Keywords: corpus, corpus linguistics, corpus analysis.

Кубрак Д.-О.Д.
Національний університет «Львівська політехніка»

КОРПУСНА ЛІНГВІСТИКА ТА КОРПУСНИЙ АНАЛІЗ

Анотація. У статті розглядаються особливості лінгвістичного корпусу та корпусного аналізу, а також їх класифікації, забезпечуючи більш чітку перспективу корпусних досліджень. Корпусний метод і його ресурси сьогодні активно використовують у різних лінгвістичних дослідженнях багатьох світових мов; мовні корпуси розрізняють за обсягом, типом, структурою, наповненням, призначенням тощо. Масиви даних писемного й усного мовлення дають змогу як досліджувати окремі мовні явища, так і з'ясовувати закономірності функціонування мовних одиниць різних рівнів. В ході дослідження з'ясовано, корпус мовлення – це спеціальна колекція ретельно відібраних уривків (слів, фраз, речень), вимовлених численними мовцями за різних акустичних умов. Дослідниками українського мовлення було розроблено такі корпуси: UkReco – українськомовний багатодикторний мовленнєвий корпус, що містить понад 30 000 реалізацій слів і тисячі речень; постійно доповнюється і вдосконалюється Акустичний корпус українського ефірного мовлення (АКУЕМ). Корпсуи іноземного мовлення: the London-Lund Corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), theCambridge and Nottingham Corpus of Discourse in English (CANCO DE), the Santa Barbara Corpus of Spoken AmericanEnglish (SBCSAE) та the Wellington Corpus of Spoken New Zealand English (WSC). Таким чином, використання ресурсів корпусної лінгвістики дає змогу поглибити іншомовну компетенцію студентів факультетів іноземної філології, удосконалити спеціалізовану лінгвістичну підготовку, а також створити ефективні умови для організації науково-дослідної роботи студентів-лінгвістів та реалізації їхнього творчого потенціалу. Перспективи дослідження полягають у подальшому аналізі дидактичних можливостей використання ресурсів корпусної лінгвістики, зокрема корпусу мовлення у виробленні методичних рекомендацій застосування корпусів англійської мови в навчальному процесі. Теоретичні проблеми корпусної лінгвістики досліджуються такими українськими вченими: О. Демської (Кульчицької), В. Широкова, Є. Карпіловської, Н. Дарчук, В. Жуковської, В. Балог, С. Бук та ін.

Ключові слова: корпус, копусна лінгвістика, корпусний аналіз.

Problem statement. With the development of corpus technology, a new problem has appeared: on the one hand, many corpora have been established, and much money and time have been put into their technology; on the other hand, these corpora are difficult to share among different affiliations. The main reason for this problem is the lack of general specifications for corpus collection, annotation, and distribution.

Recent research and publications analysis. Corpus linguistics is one of the fastest growing methodologies in modern linguistics. Corpus linguistics is an applied linguistic approach that has become one of the dominant methods used today for language analysis.

O. Selivanova proposes the description of corpus linguistics with four main features [2, p. 667–669]:

1) it is an empirical (experimental) approach, analyzing the patterns of language use, which are observed in the texts of real language (spoken and written);

2) it uses a representative sample of the target language, which is stored as an electronic database (corpus), as a basis for analysis;

3) it relies on computer software to calculate linguistic data as part of the analysis;

4) for the interpretation of conclusions, it depends on both quantitative and qualitative analytical techniques.

John Sinclair is a well-known British linguist who made a great contribution into the development of corpus linguistics and English-language corpora.

J. Sinclair was the first leader in colloquial research, colloquial language research, and computational linguistics [10, p. 78].

According to J. Sinclair *"A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research"* [10, p. 79].

Concerning the aspects of Corpus linguistics represented by famous methodologists, it is worth of attention that the key point in Corpus Linguistics is a corpus. In fact, the word corpus has always been used by linguists to indicate *"a collection of naturally occurring examples of language, consist-*

ing of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study" [12, p. 74].

Allocation of previously unsolved parts of the overall problem. More and more linguists become interested in corpora, especially in various kinds of speech-based documentation of culture and local history, originating outside of the field of linguistics, suggesting the possibility of greater collaboration between linguists and non-linguists in this area.

The article objectives setting. The purpose of this article is to examine the features of the corpus linguistics and corpus analysis, further as their classification.

Presentation of the main research material. A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language.

There are many ways to define a corpus but there is an increasing consensus that a corpus is a collection of:

– machine readable;

– authentic texts (including transcripts of spoken data) which is

– sampled to be

– representative of a particular language or language variety [8].

With the appearance of computers and the development of modern corpus linguistics, the word *corpus* has acquired a more specialized meaning as a collection of electronic texts selected and collected for a specific purpose according to certain criteria.

A corpus can be defined as a systematic collection of naturally occurring texts (of both written and spoken language) [9, p. 156].

Corpus can be regarded as a powerful tool in lexicographical studies, because the use of web-based corpora is the basic source for gathering lexicographical data.

The following list describes the four main characteristics of the modern corpus [8, p. 68]:

– Sampling and representativeness

– Finite size

– Machine-readable form

– A standard reference

Corpuses can be classified in a number of different ways. For example [3; 4; 6]:

By medium:

– Printed

– Electronic text

– Digitalised speech

– Video

– Mixed

By design method:

– Balanced

– Pyramidal

– Opportunistic

By size:

– Fixed size

– Monitor [5].

There are many types of corpora that can be used, for example, for different types of analyses [3, p. 198]:

– general/reference corpora (vs.specialized corpora) – (e.g. BNC = British National Corpus): aim at representing a language or variety as a full (contain both spoken and written language, different text types etc.);

– historical corpora (vs. corpora of present-day language) – (e.g. Helsinki Corpus, ARCHER) aim at presenting an earlier stage or previous stages of speech;

– regional corpora (vs.corpora containing more than one variety) – (e.g. WCNZE = Wellington Corpus of Written New Zealand English) aim at presenting one regional variety of a language;

– learner corpora (vs. native speaker corpora) – (e.g. ICLE = International Corpus of Learner English) aim at representation of the language as such, which is produced by learners who study this language;

– multilingual corpora (vs. one-language corpora) represents several, at least two, different languages, often with the same text types (for contrastive analyses);

– spoken (vs. written vs. mixed corpora) (e.g. LLC = London-Lund Corpus of Spoken English) represents spoken language;

– monitor corpora – The monitor corpus is one that is regularly "updated" with new texts. This is done in such a way that "...the proportion of text types remains the same...", meaning that each new version of the corpus is comparable to all previous versions [11, p. 123].

Corpus linguistics, therefore, is the study of a language that occurs in nature, based on computerized corpora. The analysis is usually performed using a computer, ie using specialized software, and takes into account the frequency of the studied phenomena [2, p. 667–668].

The use of corpora as a tool in the creation of dictionaries began in the 1980s with the COBUILD project. Subsequently, the publication of Collins COBUILD English Language Dictionary in 1987 was declared the first corpus English dictionary.

The lexicographers might use corpora to [7, p. 38]:

– collect evidence that can either supplement or refute their intuitions;

– find new words in the language;

– identify how existing words changing their meanings;

– recognise how existing words balance their use by genre, texts, etc.;

– name all the examples of specific words to justify their contextual variations;

– review existing dictionaries;

– present more complete and accurate definitions of different language subjects;

– provide updated information about any changes or losses of any word in the language;

– organize examples taken from corpora into more meaningful groups for analysis;

– categorize certain words according to different research parameters;

– single out word combinations to investigate the existence of any inherent mutual relations that guarantee their joint occurrence, etc.;

– treat phrases and collocations more systematically because phrases and collocations may provide important clues about the specific meaning of a word;

– link the use of certain words or phrases as typical for certain regional varieties, genres, etc.,

by studying corpora rich in textual and non-textual information (e.g., regional variety, author, date, sex, genre, part-of-speech tags, etc.).

Anyway, we can sum up the use corpora in the following way [1, p. 153]:

– Corpora will become natural and indispensable resource in the field of general language study, description, and teaching.

– Corpora will become the most reliable treasure-house for creation of various dictionaries and reference books (both monolingual and bilingual).

– Corpora will become indispensable in the development of various language processing tools, systems, and software.

– Corpora, for their easy availability and fast accessibility in machine-readable form, will become ready-made source for multi-purpose (mostly non-linguistic) use by the end-users.

– Corpora can be customised for studying some particular area of interest.

Research findings and prospects. To conclude, Corpus Linguistics is becoming one of the dominant approaches used in linguistic research and it is increasingly being used in lexicography. The success of the approach is inextricably linked to the tools used to access, analyze, and display search results in the corpus. We hope that this scientific work will provide a new perspective on the corpus analysis of IT neologisms, which will lead to further growth of corpus linguistics and the industry.

There is no need to look beyond well-known corpora such as the British National Corpus (BNC), the American National Corpus (ANC) and the International Corpus of English (ICE) to understand how widespread spoken and written differences become a feature of corpus design.

References:

1. Zhukovska, V.V. (2013). Resursy korpusnoi linhvistyky u doslidzhenni istorychnoi dynamiky movy. Materialy mizhnarodnoi naukovoi konferentsii *«Slovo i rechennia: syntaktyka, semantyka, prahmatyka»*. Kyiv: Kyiv. un-t im. B. Hrinchenka, pp. 151–156.
2. Selivanova, O.O. (2008). Korpusna linhvistyka. Suchasna linhvistyka: napriamy ta problemy: pidruchnyk. Poltava: Dovkillia-K, pp. 667–669.
3. Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Language and Politics*, no. 4(2), pp. 197–226
4. Crystal, D. (2002). Language and the Internet. Cambridge: Cambridge University Press, 272 p.
5. Crystal, D. (2003). English as a Global Language. Camridge: Cambridge University Press, 212 p.
6. Leech, G. (2007). New resources, or just better old ones? *Corpus Linguistics and the Web*. Amsterdam: Rodopi, pp. 134–149.
7. McEnery, T., & Gabrielatos, C. (2006). English Corpus Linguistics. The Handbook of English Linguistics: McMahon-Blackwell Publishing, pp. 33–72.
8. McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based Language Studies: an Advanced Resource Book. London: Routledge, 386 p.
9. Meyer, C.F. (2002). English corpus linguistics. Cambridge: Cambridge University Press, 241 p.
10. Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press, 170 p.
11. Stefan, Th. (2019). Gries What is Corpus Linguistics / Language and Linguistics Compass 3. Blackwell Publishing Ltd., 17 p.
12. Zhukovska, V.V. (2011). Corpus-based approach to teaching vocabulary and grammar // XVI TESOL-Ukraine International Conference Current Studies in English «Linguistics and methodology perspectives». Zhytomyr, Kamianets-Podilsky, p. 171.