International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 28 No. 2 September 2021, pp.358-359

Capabilities And Distinctive Features Of The Coca Corpus

Nozimjon Bobojon o'gli Ataboev

PhD on Philology, Head of the department of English literature and stylistics at Bukhara State University

Abstract – The article analyzes the compliance of COCA with functional corpus principles and its advantages. Attention was paid to the capability of COCA corpus, a second-generation representative of the modern English corpus, was selected is the object of study, to meet the above requirements. It also deals with the corpus results and statistics related to the language use as well as the conclusions made relying on them.

Keywords – Capabilities,Distinctive Features, Coca Corpus.

Currently, new information technologies are used in almost all spheres of linguistics – both theoretical and applied. Actively improving computer programs and projects for working with text, search and recognition of information influenced on the emerge of the corpus linguistics. One of the types of such projects, allowing both a linguist and any user interested in a language, to access huge volumes of data in the shortest possible time, are language corpus.

It need to be noted that it is wrong to accept the notion of CL as a field such as semantics, syntax, sociolinguistics and so on, because it is not a section of linguistics but rather a methodology that does not require interpretation and description. That is to say that CL should be regarded as a research methodology that does not limit itself to only one area of linguistics but has research methods that can be used in any linguistic scientific works. It is reasonable to conclude that CL was considered as a research method in the early stages of its development and was considered as part of a network of computer science-based disciplines, but has now become a methodology with its goals and research methods and the wider scientific community. However, in the future, it is worthwhile to consider the high-tech CL as a linguistic research paradigm and / or an individual science that can solve problems in all areas, in accordance to our high-level scientific hypothesis

The stages of development of Corpus linguistics (CL) are inextricably linked with the theoretical principles of the corpus. Although research on CL in linguistics and the organization of teaching began in the eighteenth century, the first electronic linguistic corpus of texts appeared in the 1960s. As a result of a project that began in 1961, in 1964, scholars at Brown University (USA) created large volumes of texts on a computer (Brown Corpus). U. Francis and G. Kuchera created five hundred and two thousand prose copies of the American edition in English. Texts belonging to the fifteen major genres of English prose were formed on the basis of the materials of the United States Press Corps and its basic statistical processing. In 1961, the first results on the corpus were published in the form of a frequency and alpha-frequency dictionary. According to X. Ling[1], such voluminous corpora were applied only in the study of specific aspects of phonetics.

As for us, the sphere of corpus linguistics is not only a language theory, but also the source of the exact statistical data concerning the language.

[1]Ling X. Cihui Yu. he Jisuan Yu. [Lexical Semantics And Computational Linguistics]. BeijingLingua Publishing house, 1999. – P. 240.

The 450+ million-word corpus, which emerged as a result of a project led by Mark Devies, is the only large-scale balanced corpus of American English. It is used by more than 40,000 individual users every month, making it the most widely used online corpus today. Thanks to its design, this linguistic corpus is the only major corpus that can be used to review existing changes in the English language[2]. That is, now that it has accumulated nearly thirty years of language data, it is possible to track changes in the lexical layer of the language during this period, and the project continues to add data to the database as it is not yet complete. Access to the COCA corpus is through *English-Corpora.org*, where one can register for free use of existing corpora. This has made it the most visited site today. For example, from October 1 to October 31, 2019, the number of users of the site reached 130 thousand. Attracting more than a hundred thousand users a month, the best-designed language corpus on this site is COCA.

To date, according to the database, which was last entered in March 2020, the COCA volume contains more than 1 billion and a million words consisting of 485 202 texts, and this division was formed between 1990 and 2020 as a result of 30 million entries per year. Over the years, the corpus materials have been divided into eight areas, such as spoken, fiction, popular magazines, newsapers, academic journals, blogs, Web pages and TV/Movies subtitles, and information has been added on a regular basis[3].

Corpus analyses show that the properties of the collocation units provided are important. As an example, in order to analyze the semantic properties of the verbs *cause* and *result in*, which have a similar meaning, words that are collocated with them in the COCA were searched. In doing so, the concordance of the corpus was used to search for nouns that came as an object of the verbs *cause* and *result in*. In the concordance lines, it was found that the verb *cause* was used 71078 times and the verb *result in* was used 84001 times. The results of searches in the order *Cause + nn* * and *result in + nn* * showed that at 4394 and 2907, respectively. When attention was paid to the objects of the verb *Cause*, it was clarified that almost all of them were nouns representing a negative meaning. For example, *problems, cancer, trouble, harm, damage, disease, pain, illness, injury, stress, infection, discomfort,* and more others. The verb "*Result in*" was found to be associated with negative concepts such as *death, loss, pressure, discrimination, rape, war, trauma,* as well as positive concepts such as *change, chance, negotiations, improvements, power, quality* and others.

J. Sinclair[4], a scholar who was the first in the history of CL to use phrases and individual words using linguistic corpora, analyzed two different phrases. They are: *naked eye* and *true feeling*. Examples of the use of phrases in his work are taken from the BNC corpus, and while 148 examples were found for the *naked eye*, for *true feeling* this number was 53. When these collocations were searched in COCA, it was observed that the number of the phrases increased significantly. That is, 654 and 175 results were obtained, respectively. This, in turn, leads to a broader discussion of the phraseological units sought.

In English, there are affixes as *–ous* and *–ness* to form adjectives and abstract nouns. The question of whether they can be in one word, and if so, in what order, can be answered on the basis of search results in COCA. The results of the search showed that the combination of morphemes in the order * *ness + ous* does not exist in the corpus, so it can be concluded that it does not exist in the language. Results corresponding to the * *ous + ness* order were observed in 112 cases, which is the total number of more than 10 word-types used. For example, *consciousness, seriousness, nervousness, righteousness, graciousness, dangerousness* and others.

Moreover, in the work, the possibilities that existing monolingual electronic dictionaries of modern English can provide to the user and the capacity of the COCA as a dictionary corpus has been compared. The optional word was randomly selected and it was stated that the analysis of the results obtained from the search in English dictionaries was not as sufficient as the ones COCA provided.

REFERENCE

[1] Ling X. Cihui Yu. Jisuan Yu. [Lexical Semantics And Computational Linguistics]. BeijingLingua Publishing house, 1999. – P. 240.
[2] Mark Davies, Professor of Linguistics, Brigham Young University. Available at URL:http://davies-linguistics.byu.edu/personal/
[3] Corpus of Contemporary American English//English-corpora.org https://www.english-corpora.org/coca/
[4] Sinclair J. Trust the text: language, corpus and discourse. London: – Routledge. 2004. – 30-36 pp.

[2] Mark Davies, Professor of Linguistics, Brigham Young University. Available at URL:http://davies-linguistics.byu.edu/personal/
[3] Corpus of Contemporary American English//English-corpora.org https://www.english-corpora.org/coca/
[4] Sinclair J. Trust the text: language, corpus and discourse. London: – Routledge. 2004. – 30-36 pp.