CORPUS LINGUISTICS: PROBLEMS AND PROSPECTS

Ataboyev Nozimjon Bobojon o‘g‘li
Associate Professor of Bukhara State University, PhD

Astanova Gulnora Maqsudovna
Master student of Bukhara State University

ANNOTATION: *Corpora of written and oral texts are currently successfully used in linguistic pedagogy and in teaching a foreign language. World practice in the development of corpus linguistics and its application in teaching prove the effectiveness of corpus methods. Modern technologies not only change old linguistic tools (transforming, for example, traditional dictionaries into computer databases), but also create new ones. Such new linguistic resources include text corpora.*

KEY WORDS: *corpus linguistics, languages, databases, computational linguistics, computer science, information.*

INTRODUCTION

Corpus linguistics in its modern sense originated in the USA and Western Europe in the late 1960s. With the growth of the capabilities of modern computer technologies, since the mid-1980s[1]. Corpus projects of various sizes began to actively appear in different languages and for a variety of purposes. The first corpus was created in 1963 in the USA (The Brown Standard Corpus of American English). The volume of this corpus under the authorship of W. N. Francis (W. Francis) and G. Kucera (H. Kucera) amounted to 1 million words (500 texts) 2 thousand uses of words each). It has become a popular research subject and example for creating similar databases. Among scientists there is an understanding that a number of correct linguistic studies

LITERATURE ANALYSIS AND METHODOLOGY

It can only be carried out on a large speech material. All this stimulated the emergence of an approach that developed rules for organizing texts into a corpus and methods for their analysis, and corpus linguistics thus emerged as a methodology for such an approach. In our country today the most common definition is that given to this discipline by V.P. Zakharov: “Corpus linguistics is a section computational

[1] *Захаров В. П.* Корпусная лингвистика : учеб.-метод. пособие. СПб., 2005. С. 3.

linguistics, engaged in the development of general principles for the construction and use of linguistic corpora (text corpora) using computer technology[1]". However, due to insufficient development subject of study, the question of definitions in this case is still open. Scientists are arguing about the direction to which corpus linguistics should be classified. Thus, V.V. Mamontova considers the above definition of a deep understanding of this discipline and limiting it to the framework of computational linguistics, while "computational linguistics is usually understood as a broad area of using computer tools - programs, computer technologies for organizing and processing data - to model the functioning of language in certain conditions, situations, problem areas , as well as the scope of application of computer language models not only in linguistics, but also in related disciplines[2]". V.V. Mamontova believes that corpus linguistics uses computers precisely as a tool, and without them; of course, it would not be able to perform its functions. However, in her opinion, this can be attributed to almost any branch of modern knowledge, which does not make them components of computer science. In modern conditions, a corpus is a powerful and effective tool for scientific research, including the theory and practice of translation. The purpose of the corpus is to provide all kinds of information from different areas (vocabulary, grammar, accentology, language history, etc.) for teaching native or foreign language. One of the advantages of corpus research in lexicography is that a corpus can be used to show the variety of contexts in which a word is used. Then from these contexts it is possible to identify different meanings associated with the word.

DISCUSSION AND RESULTS

According to experts, the methodological apparatus of corpus linguistics is a promising tool in theoretical and practical teaching of a foreign language. The results of corpus searches (concordances) in printed form can be easily incorporated into handouts, teaching aids, etc. and used in the process of traditional teaching. In addition to direct application in a foreign language classroom, the corpus as a method can be used to critically evaluate existing teaching materials. Thus, in most of these studies, researchers have found that there are significant discrepancies between what is prescribed in English grammar textbooks and how the language is actually used by native speakers, as evidenced by the spoken language corpus.

Currently, there are a significant number of corpora, which are arrays of text data in different languages, but the ability of a researcher to use them is limited due to a number of serious problems:

[1] *Захаров В. П.* Корпусная лингвистика : учеб.-метод. пособие. СПб., 2005. С. 3.
[2] *Баранов А. Н.* Компьютерная лингвистика // А. Н. Баранов. Введение в прикладную лингвистику : учебное пособие. М. : Эдиториал УРСС, 2003. С. 13.

1.1. Corpora available on the Internet are focused on the analysis of lexical or grammatical phenomena, and units of communication that do not have standard methods of expression, discursive phenomena, pragmatic features of texts, speech acts, etc. are not marked in them, since the creation of such automatic marking, at least At least, at this stage of development of computational linguistics, it is impossible. In this regard, some researchers deny the usefulness of corpora in the study of text and discourse.

1.2. The next problem is limited access or lack of access to some buildings. In particular, the largest corpus, including works of French literature from the 20th century to the 21st century Frantext is available for limited online searches only, subscription on behalf of an academic/educational institution is required. Many English language corpora are paid. Full access to the British, American National Corpuses, International English Corpus and some other corpora is not available online. But there are a number of corpora available - these are the corpora created by Mark Davis and the corpora available from the University of Leeds website. 1.3. Available corpora do not always allow us to obtain the context of use and access to the full text of the work, which reduces the possibility of adequate interpretation of the phenomenon under study in the text. Most often, the researcher has access to one line, one sentence, or a certain number of characters (the limit depends on the corpus) in which the requested word or expression is used.

1.4. The main problem of using corpora in comparative studies is the different volume, structure of corpus arrays, creation time and types of text documents. In particular, when comparing French, English and Russian corpora, all of the above (1–3) problems are revealed. For greater clarity, let us compare the available corpora of these languages. The French Corpus of the University of Leipzig is a database of 700 million words. It contains texts from the French press (about 350 million words), web pages and Wikipedia materials. This corpus allows you to get a list of examples in the size of one sentence per query word with a link to the source, but without a date, as well as a list of right-handed and left-handed collocates, i.e. words that are often used with the requested word. The Frantext Corpus , created by the University of Nancy, contains marked-up texts of French literary works from the 10th to the 21st centuries. 286 million words long and has limited access. The Canadian French Corpus is publicly available

Lexiqum (about 229 million words), created by the University of Montreal to study the joint lexical occurrence and syntactic compatibility of various expressions. The number of examples per request is limited to 500 sentences of no more than 200 characters. The corpus uses only metatext markup. At the moment, there is no national

corpus of the French language comparable in structure, layout, functions, for example, with the British National Corpus (100 million words), the COCA Corpus of Contemporary American English (445 million words) or National Corpus of the Russian Language (more than 500 million words). NKRYA balanced body with six types of markings. It includes works of fiction, scientific texts and journalism from Ser XVIII to the beginning XXI century There is also a computer corpus of newspaper texts in the Russian language of the late twentieth century (1994–1997) with metatextual, morphological, syntactic, and semantic markings. As we can see, the presented corpora are not equivalent in volume, content, structure, chronological framework, access capabilities, etc., which prevents the use of these corpora in comparative studies, since the results obtained from data analysis reflect different text arrays collected from various principles and criteria.

There are a number of different English-language corpora (British National Corpus, Corpus of Contemporary American English, etc.). Getting to know these corpora demonstrates a wide range of possibilities not only for the researcher, but also for the student as part of independent work. Thus, teaching English at a law school is traditionally based on the British and American national varieties. Based on the existing British National Corpus (BNC), it is possible to conduct a comparative analysis, comparing the data with the data of the Corpus of Contemporary American English (COCA), not only lexical and grammatical features, but also the frequency of use of certain language units and their possible collocations. The British National Corpus (BNC) 8 is a collection of 100 million words of written and spoken language from a wide range of sources, intended to represent a broad cross-section of British English since the late twentieth century, both spoken and written. The latest edition of the British National Corpus, the BNC XML Edition, was released in 2007. The written portion of the BNC (90%) includes extracts from regional and national newspapers, specialist periodicals and magazines for all ages and interests, academic books and popular literature, published and unpublished letters and memoranda, school and university essays and many other types of text.

CONCLUSION

We can highlight the prospects of corpus linguistics and the problems arising from them. Firstly, the huge amount of information makes it difficult to find, which requires the creation of special corpora and additional search tools, designed specifically for the purposes of this building. Secondly, conceptual search, designed to make a user query more effective, actually comes down to searching for synonyms and systematizing them according to semantic nests, i.e., it belongs more to the field of semantics than logic, and is difficult to formalize.

Thirdly, corpora are built with the goal of providing objective information, so the creators of the corpus are faced with the task of reducing the factor of subjectivity when selecting texts for the corpus and developing strict and clear selection criteria. All this does not reduce the significance and prospects of research in the field of corpus linguistics and points to new directions for applied research.

REFERENCES:

1. Захаров В. П. Корпусная лингвистика: учеб.-метод. пособие. СПб., 2005.
2. Баранов А. Н. Компьютерная лингвистика // А. Н. Баранов. Введение в прикладную лингвистику: учебное пособие. М.: Эдиториал УРСС, 2003.
3. French building of the University of Leipzig [Electronic resource]. Access mode: http://wortschatz.uni-leipzig.de/ws_fra/.
4. Frantext Corpus [Electronic resource]. Access mode: http://www.frantext.fr.
5. Corpus of the French language Lexiqum [Electronic resource]. Access mode: http://retour.iro.umontreal.ca/cgi-bin/lexiqum.
6 British National Corps [Electronic resource]. Access mode: http://www.natcorp.ox.ac.uk/.
7. Corpus of Contemporary American English COCA [Electronic resource]. Mode Access: http://corpus.byu.edu/coca/.
8. National Corpus of the Russian Language [Electronic resource]. Access mode: http://www.ruscorpora.ru/.
9. Computer corpus of newspaper texts in the Russian language at the end of the twentieth century [Electronic resource]. Access mode: http://www.philol.msu.ru/~lex/corpus/.