Intent Research Scientific Journal-(IRSJ)
ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

THE HISTORY OF DEVELOPMENT OF CORPUS LINGUISTICS AND ITS ROLE IN TEACHING FOREIGN LANGUAGES

Shayusupova Nargiza Bakhtiyorovna
Branch of Astrakhan State Technical University in Tashkent Region,
Republic of Uzbekistan, Assistant Teacher

Abstract
The article contains brief information on the history and development of corpus linguistics — one of the sections of modern linguistics, the purpose of which is the creation and use of language corpora in linguistic research. On the basis of the literature review by several scholars the author tries to show the advantages of using corpora in teaching foreign languages.

Key words: corpus linguistics, databases, corpora, corpus, COCA

Introduction

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

Corpus linguistics in its modern sense originated in the United States and Western Europe in the late 1960s. With the growth of the capabilities of modern computer technology, since the mid-1980s. corpus projects of various scales began to appear actively in different languages and for various purposes. The first corpus was created in 1963 in the USA (The Brown Standard Corpus of American English). The volume of this corpus, authored by W. Francis and H. Kucera, amounted to 1 million word usages (500 texts of 2 thousand word usages each).

It has become a popular subject of research and an example for creating similar databases. Among scientists, there was an understanding that a number of

ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

correct linguistic studies can be carried out only on a large speech material. All this stimulated the emergence of an approach that develops the rules for organizing texts into a corpus and methods for their analysis and corpus linguistics, thus, it was born as a methodology for such an approach.

V.P. Zakharov defined this discipline by saying that: "Corpus linguistics is a branch of computational linguistics that develops general principles for the construction and use of linguistic corpora (corpuses of texts) using computer technologies"[1] . However, due to the insufficient development of the subject of study, the issue of definitions in this case is still open. Scientists are arguing about the direction to which corpus linguistics should be attributed.

So, V. V. Mamontova considers the above definition to narrow the understanding of this discipline and limit it to the framework of computational linguistics, while "computational linguistics is usually understood as a wide area of using computer tools - programs, computer technologies for organizing and processing data - for modeling the functioning of a language in certain conditions, situations, problem areas, as well as the scope 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 integral parts of computer science.

Having studied the modern conditions for the functioning of corpus linguistics, comparing its various tools, goals, objectives and results of research carried out using corpora, in terminological aspect, we agree with the opinion of V.P. Zakharov, since it was global computerization that contributed to the emergence of this phenomenon, and also because that without machine collection and processing of material, research in this case is not possible, while even with the development of technology, many other progressive areas of linguistics do not depend on computers.

In modern conditions, the corpus is a powerful and effective tool for scientific research, including in the theory and practice of translation. The task of the corpus is to provide all kinds of references from different areas (vocabulary, grammar, accent ology, history of the language, etc.) for teaching a native or foreign language. One of the advantages of corpus research in lexicography is

ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

that a corpus can be used to show the many contexts in which a word is used. Then, from these contexts, you can extract different meanings associated with the word.

According to experts, the methodological apparatus of corpus linguistics is a promising tool in the 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 class, the corpus as a method can be used to critically evaluate existing teaching materials.

So in most of these studies, scientists 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 colloquial corpus. However, an analysis of thematically conditioned literature shows that, so far, corpora of written and oral texts are used in linguistic analysis much more successfully than in teaching a foreign language and in linguistic pedagogy.

In general, experts agree that the methods of corpus linguistics should be mandatory in the development and evaluation of the effectiveness of educational materials and teaching aids so that the most common uses receive priority attention, and peripheral uses take their appropriate place. It is important to note that corpus linguistics as a scientific methodology and branch of linguistics is very young and develops unevenly in different countries.

For some languages, such as English, German, Finnish or Japanese, the most extensive and representative annotated languages, while for other languages, including Russian, the process of creating full-fledged corpora is in its infancy. The situation with multilingual (parallel or comparative) corpora has revealed even more that the considerable prospect of corpus methods in the field of linguistic research, language technologies and teaching foreign languages is extremely great.

As a result, the terminological system related to it still lacks equivalents in Russian. Modern linguists have to master the concepts of "concordance", "tokenization", "lemmatization", "stemming", "parsing", "tagger" and many others. The opinions of specialists differ even regarding the Russian-language

ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

norm of the plural of the starting lexeme “corpus”: some linguists, along with the generally accepted norm “corpus”[3], introduce the unit “corpus” into the usus, which causes only a strong association with the common modern colloquial error “polis” - "poles".

Nevertheless, in recent decades, the vast corpora of texts used by researchers of foreign language teaching methods to assess the realities of the language in its natural state have greatly influenced the improvement of the quality level of the produced language manuals. Special mention deserves new dictionaries created using corpus linguistics techniques, such as Longman, Oxford, Collins, starting with the experience of critical rethinking of the postulates of the descriptive grammar of the English language.

One of the first in this series was the Longman Grammar of Spoken and Written English, published in 2000. According to T. N. Sergeeva, the main tasks of corpus linguistics can be reduced to the following:

1) development of the theoretical foundations of this direction;
2) analysis of the experience of creating and using hulls of various types;
3) formulation of general requirements for the hull;
4) creation of buildings for various research and educational tasks;
5) the formation of effective ways of using text corpora in various fields of linguistics[4]. As we see it, it is the latter task that is of practical value for the widest layers of researchers and teachers of a foreign language.

Corpus methods have also proven themselves in the world practice of teaching the language of professional communication, being a highly effective innovative addition to traditional educational technologies. These methods combine such aspects as interdisciplinarity, empirical adequacy, authenticity, flexibility and adaptation to specific tasks and target groups, the possibility of independent work of the student. A corpus search can return:

1) all uses of the selected word in the immediate context, on the basis of which the student will be able to decide on the choice of the exact equivalent in translation;
2) control of the verb, noun, etc. (words that most often stand next to the selected word);
3) intertextuality: the meaning of a word as the sum of its uses;

Intent Research Scientific Journal-(IRSJ)
ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

4) development of concepts in time. There are a number of different English-language corpora (British National Corpus, Corpus of Contemporary American English, etc.).

Acquaintance with these corpora demonstrates a wide range of opportunities not only for the researcher, but also for the student as part of independent work. On the basis of the Existing British Corpus (BNC), it is possible to conduct a comparative analysis, comparison of data with the data of the Corpus of Modern American English (COCA), not only lexical and grammatical features, but also the frequency of use of those other language units and their possible collocations.

The British National Corpus (BNC)[5] is a collection of 100 million written and spoken words from a wide range of sources, designed to represent a broad cross-section of British English since the late 20th century, both spoken and written.

The latest edition of the British National Corpus, the BNC XML Edition, was released in 2007. The written part of the BNC (90%) includes extracts from regional and national newspapers, specialized periodicals and magazines for all ages and interests, academic books and popular literature, published and unpublished letters and memorandums, school and university essays and many other types of text.

The oral part (10%) consists of scripted informal conversations (recorded by volunteers selected from different age, regional and social classes in a demographically balanced way) and other examples of spoken language collected in various contexts, ranging from formal business or government meetings to radio broadcasts and phone calls.

COCA[6] was created by Mark Davies, Professor of Corpus Linguistics at Brigham Young University. The corpus, released in 2008, is currently popular and is used by tens of thousands of users (linguists, teachers, translators and other scientific researchers). It is one of the few freely available corpora of texts in American English.

The corpus consists of more than 450 million words and more than 160 thousand texts. It covers the period from 1990 to 2012 and is updated with approximately 20 million words annually. The last update was made in the summer of 2012. Texts come from different sources, which are evenly distributed among five genres:

Intent Research Scientific Journal-(IRSJ)
ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

1) Spoken / Conversational genre (85 million words) - transcripts of conversations from almost 150 different television and radio programs;
2) Fiction / Fiction (81 million words) - short stories, plays, film scripts and some book chapters written from 1990 to the present;
3) Popular magazines / Popular magazines (86 million words) - about 100 different magazines from such areas as news, finance, religion, sports, health, home and gardening;
4) Newspapers / Newspapers (81 million words) - 10 American newspapers with sections such as local news, expert opinion, sports and finance;
5) Academic journals / Scientific journals (81 million words) - almost 100 different peer-reviewed scientific journals.

Many linguists use the corpus as a "bank of examples", i.e. try to find empirical support for their hypotheses, the principles and rules they work on. The corpus linguistics approach ensures the representativeness and balance of linguistic material, as well as a search tool that usually allows a good sample in a given corpus.

Over time, access to language corpora has become easier. Even 10 years ago, as a rule, special programs, servers, CDs were required. To date, these resources are available for free download or online.

According to Z. K. Gadzhiyeva, the following factors that determine its reliability and reliability speak in favor of this method: - the corpus method is considered to be empirical, its main goal is to analyze real word usage in a natural language environment; - uses a sufficiently large, representative selection of texts;

– active use of computers and special concordance programs for analysis in automatic and interactive modes of operation;

- relies on the methods of statistical and qualitative analysis of the text;

— is targeted, i.e. should be oriented towards real application and results [7].

In large corpora, a problem arises that is becoming more and more urgent: a query search can produce hundreds and even thousands of results (contexts of use) that are physically impossible to view in a limited period of time. To solve this problem, systems are being developed that allow grouping search results and automatically breaking them into subsets (search results clustering) or

ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

producing the most stable phrases with a statistical assessment of their significance.

Conclusion

However, it can still be concluded that, although at present the possibilities of corpus linguistics methods in our country have not yet been properly implemented in teaching native and foreign languages, the analysis of text corpora, methods and developments of corpus linguistics is a fairly promising area in the field of teaching foreign languages, not only to students of linguistic specialties, but also to students of non-linguistic universities.

With the development of computing technology capable of storing and processing a huge amount of linguistic data, and with the development of corpus linguistics, it became possible to base a linguistic judgment on something much more reliable and diverse than the personal experience or intuition of a single author of a foreign language textbook.

References

1. Zakharov V. P. Корпусная лингвистика : учеб.-метод. пособие. СПб., 2005. С. 3.
2. Baranov A. N. Компьютерная лингвистика // А. Н. Баранов. Введение в прикладную лингвистику : учебное пособие. М. : Эдиториал УРСС, 2003.
3. Ojogev S. I. Словарь русского языка. М. : Гос. изд-во иностр. и нац. словарей, 1963. 900 с
4. Sergeeva T. N. Тезаурусное моделирование предметной области «Корпусная лингвистика» на материале английского языка // Вектор науки ТГУ. 2009. № 1 (4). С. 87—91.
5. Gadjieva Z. K. Особенности использования корпусной лингвистики в обучении иностранному языку // Молодежь и наука: реальность и перспективы развития : материалы II Всероссийской научно-практической конференции с международным участием. Махачкала : ИП Дагерманов И. Д., 20

ISSN (E): 2980-4612
Volume 2, Issue 6, June -2023
Website: intentresearch.org/index.php/irsj/index

6. Baker M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research // Target. — № 7. — 1995. — P. 223—243.
7. Baker M. Corpus Linguistics and Translation Studies: Implications and Applications // Text and Technology. Honour of John Sinclair. — Amsterdam &Philadelphia : John Benjamins, 1993. — P. 233—250.
8. Baker P. A Glossary of Corpus Linguistics / P. Baker, A. Hardie, T. McEnery. — Edinburgh : Edinburgh University Press Ltd, 2006.
9. Biber, D., Conrad S., Reppen R. Corpus Linguistics : Investigating language structure and use. — Cambridge University Press, 2011.
10. Corpas Pastor G., Seghiri M. Specialized Corpora for Translators : A Quantitative Method to Determine Representativeness // Translation Journal. — 2007. — Vol. 11. — No. 3.
11. Fillmore Ch. J. 'Corpus linguistics' or 'Computer-aided armchair linguistics' // Directions in Corpus Linguistics. — Berlin : de Gruyter, 1992. — P. 35—60.
12. Martynenko I. A., Pikalova V. V. Teaching legal English in Russia: traditions and problems // International scholarly and scientific research and innovation : Book of Abstracts. — 2018
13. URL: http://www.natcorp.ox.ac.uk
14. URL: https://corpus.byu.edu/coca/