Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 03, March, 2025
brightmindpublishing.com
ISSN (E): 3061-6964

THE SOCIAL NECESSITY AND MAIN FEATURES OF CORPUS LINGUISTICS

Ashirbayeva Madina Nuraliyevna
Chirchik State Pedagogical University, Teacher of the "Linguistics and English Teaching Methodology" Department,
Mail: ashirbaevamadina5@gmail.com
Tel:+998998570691

Abstract:
Many prominent scholars (Sinclair, Leech, Bieber, Francis, Johnson, Conrad, Hunston and McCarthy, etc.) have been involved in corpus linguistics and have played a major role in its development. Of course, we have considered the work of only a few of them. These scholars have made significant contributions to the field of corpus linguistics, both in the past and in the present, and their conclusions on the characteristics of corpuses, or important elements that distinguish a corpus from the Internet, have been considered in our article. Many corpus scholars consider John Sinclair to be one of the most influential scholars of modern corpus linguists. Sinclair established that a word does not have meaning by itself, but that this meaning is often carried out through the sequence of words that accompany it [Sinclair, 1991]. This idea served as the basis of corpus linguistics.

Keywords:Corpus, corpus linguistics/ linguistics, authenticity, representativeness, corpus size, balance, Uzbek National Corpus.

KORPUS TILSHUNOSLIGINING IJTIMOIY ZARURIYATI VA ASOSIY XUSUSIYATLARI

Annotatsiya:
ko'pgina taniqli olimlar (Sinclair, Leech, Biber, Francis, Johnson, Conrad, Hunston va McCarthy va boshqalar) korpus lingvistikasi bilan shug'ullanib, uning rivojida ulkan rol ijro etishgan. Albatta, ulardan faqat bir nechtasininggina ishlarini ko'rib chiqdik. Bu olimlar o'tmishda ham, hozirgi kunda ham korpus

lingvistikasi sohasiga katta hissa qo‘shib kelmoqda va ularning korpus xususiyatlari, yoki korpusni Internetdan ajratib turadigan muhim elementlari bo‘yicha bergan xulosalarini maqolamizda ko‘rib chiqildi. Ko‘plab korpusshunos olimlar Jon Sinklerni zamonaviy korpus tilshunoslarining eng nufuzli olimlaridan biri deb bilishadi. Sinkler so‘z o‘z-o‘zidan ma’noga ega emasligini aniqladi, lekin bu ma’no ko‘pincha u bilan birga keladigan so‘zlar ketma-ketligi orqali amalga oshiriladi degan g‘oyani berdi [Sinclair,1991]. Bu esa korpus lingvistikasining asosini tashkil etuvchi g‘oya bo‘lib xizmat qildi.

Kalit so‘zlar: korpus, korpus lingvistikasi/tilshunosligi, autentiklik, reprezentativlik, korpus o‘lchami, muvozanat, O‘zbek Milliy Korpusi.

СОЦИАЛЬНАЯ НЕОБХОДИМОСТЬ И ОСНОВНЫЕ ЧЕРТЫ КОРПУСНОЙ ЛИНГВИСТИКИ

Аннотация:

Многие выдающиеся ученые (Синклер, Лич, Бибер, Фрэнсис, Джонсон, Конрад, Ханстон и Маккарти и другие) занимались корпусной лингвистикой и сыграли важную роль в ее развитии. Конечно, мы рассмотрели работы лишь некоторых из них. Эти ученые внесли значительный вклад в область корпусной лингвистики как в прошлом, так и в настоящем, и в нашей статье рассматриваются их выводы относительно особенностей корпусов или важных элементов, отличающих корпус от Интернета. Многие исследователи корпусной лингвистики считают Джона Синклера одним из самых влиятельных ученых среди современных корпусных лингвистов. Синклер обнаружил, что слово само по себе не имеет значения, но это значение часто передается через последовательность слов, которые его сопровождают [Синклер, 1991]. Эта идея легла в основу корпусной лингвистики.

Ключевые слова: корпус, корпусная лингвистика, аутентичность, репрезентативность, размер корпуса, баланс, Узбекский национальный корпус

Introduction

In recent years, the concept of "corpus of texts" has increasingly entered the scientific language circle of linguists, on the basis of which corpus linguistics is developing. This concept exists in the scientific literature along with the concepts of "collection of texts", "full-text database", "electronic archive", "electronic library", and is often combined. In a narrow sense, a corpus of texts is usually understood as a unified, systematized and defined collection of linguistic (speech) data in electronic form, which has a certain philological and broader meaning.

It is necessary to consider the concept of corpus linguistics in conjunction with the corpus, since it is the academic field that originally dealt with corpus. In order to ensure a complete and progressive understanding of the topic and to further develop the topic, it is necessary to present the basic principles of corpus linguistics. As noted above, "... it is a mistake to think that the analysis of corpora, in particular, has nothing to offer to generative theory or to the theory of language in general" [Meyer, 2002]. The most popular approaches in corpus linguistics are the Firthian and neo-Firthian approaches, which focus on language not only in its social context. The goal of corpus linguistics is to describe all empirical aspects of language use, and to this end Aarts (1993) developed a number of requirements for a descriptive model of language, which are as follows:

- The model should allow for the combination of quantitative and qualitative descriptions of data;
- The model should establish a relationship between phenomena external to the language system and phenomena internal to the system;
- The model should be able to describe the full range of types, from non-self-edited language (usually spoken) to self-edited language use (usually written or printed);
- The model should be able to describe syntactic, lexical, and discourse features in a holistic manner. [Aarts in Olohan, 2013,424].

Corpus linguistics seeks to find linguistic answers through qualitative and quantitative analysis. The use of the Internet in corpus linguistics adds new possibilities to research, which "...allows us to make generalizations about language use, emphasizing the need to focus attention and interest not on what usually happens, but on what might happen" [Kennedy, 1998].

Although there are definitions that describe corpora as collections of texts, we need to look at them more broadly. If we were to define a corpus as simply a collection of texts, "almost any collection of several texts ... could be called a corpus" [McEnery & Wilson, 2001], and this would be very difficult for researchers and linguists.

To limit the scope of corpus studies, linguists have proposed definitions of corpuses that "…contain essential criteria and standards" (both explicit and implicit) [Gatto, 2013], i.e. the definitions given above.

There are many standard definitions across corpus studies, but there is a fundamental agreement on the requirements: "corpus texts must be authentic, expressive, and machine-readable" [McEnery & Xiao, 2006]. Other important elements that a corpus should consider are balance, sampling, and size, as well as content.

B.B. Rykov gives the following definitions of corpus properties: "the definition of a corpus adopted in corpus linguistics includes 4 main properties: the arrangement of the corpus by machine programs, the standard appearance of oral materials, representativeness as a result of special selection, qualitative, quantitative dimensions are the most important features in scientific literature" [B. Rykov, 2004]

Another Russian scholar, corpus scholar V. Plungyan, who made a significant contribution to the development of the corpus, also defined a language corpus as "a collection of texts from a specific language, entered in electronic form and processed using scientific techniques" [Plungyan]. According to Y. Kratova, a corpus is "a collection of texts covering a specific field, used or written by a single type of user in specific situations" [E.B. Kratova, 2013].

The property of authenticity, as defined by many scholars, is that the corpus should consist of real spoken and written texts. Data obtained from experimental conditions and artificial situations are not valid for any research, because it introduces artificial or modified data into the other natural results of the research. This quality is closely related to representativeness. A real text or non-artificial text is an adequate representation of the language. For example, "television interview texts may appear natural, but they are deliberately placed in artificial conditions to elicit responses that will attract the attention of the audience" [Dash, 2015].

Another important element that a corpus should consider is representativeness. The academic tradition of corpora has included representativeness as an important property of a corpus [Francis, 1992; Biber et al., 1998]. Francis defined a corpus as "a collection of texts that are representative of a particular language, dialect, or other subset of a language and are used for linguistic analysis" [Francis, 1992]. Similarly, Biber states: "A corpus is not simply a collection of texts; rather, it seeks to represent a language or a part of a language. The appropriate design for a corpus therefore depends on what it is intended to represent. The representativeness of a corpus, in turn, determines the types of research questions that can be addressed and the generalizability of the research results" [Biber, 1998].

It is important to note that "representativeness" is the only common characteristic that is mentioned in both definitions of what a corpus is. It is very common for the concept of "representativeness" to come up when trying to define what a corpus is, because it is relevant to the ultimate goal of a corpus and is to represent a language or part of it for the purpose of drawing a final conclusion. The goal of corpus linguistics is to find trends and word forms in "individual performance-level language samples" [Gatto, 2013], and therefore the authenticity and representativeness of the language reported in a corpus make it such a valuable and useful tool. In other words, the results of corpus linguistics and its evolution studies are based on these two values, and therefore they are very important in a corpus.

When it comes to corpus size, there is no agreed upon standard size. Each corpus has its own purpose and data, so it is very difficult to determine what the optimal size is. Obviously, a corpus created by a student in 30 minutes may not be the same size as a corpus designed for research purposes. A corpus created for research purposes attempts to provide evidence of the variety of linguistic forms of objects, so it must be large enough and consistent enough to provide important information as evidence. In contrast, a corpus created for a single use is much smaller. This disadvantage has increased exponentially over time and as technology has become more sophisticated and scalable.

For example, the first computerized corpus, the Brown Corpus, was about one million words long, and modern corpora such as the Collins Cobuild Bank of English can contain 650 million words and WebCorpora 1 billion words. This billion-word corpus is likely to become obsolete over time. The truth about corpora is that the larger the corpus, the more information it contains, but not all of it may be useful for the subject and research purpose.

One of the main characteristics of a language corpus, which distinguishes it from the Internet, is the form (or balance) of the language data it contains, which is understood as the proportion of texts from different periods and genres in the corpus, which in turn provides the necessary completeness and uniqueness.

The formation of corpus and corpus linguistics in Uzbekistan began with theoretical studies in 2018, and began to develop practically with the creation of the Uzbek language education corpus in 2021. This is indicated in the textbook of M. Abjalova on the development of Uzbek corpus linguistics. According to her, the development of corpus linguistics in Uzbek linguistics has accelerated, practical developments have been presented to users, and this field has actually developed on the basis of computer linguistics. So, the initial ideas on working with large-scale texts in Uzbek corpus linguistics, numerous educational studies of special electronic text collections were put forward by A. Polatov and A. Rahimov. Later, corpora and their types were studied by B. Mengliyev, Sh. Khamroyeva, M. Abjalova, A. Eshmuminov, D. Akhmedova, O`. Kholiyorov, N. Abdurakhmonova, G. Toirova, D. Orinboyeva, B. Elov and A. Rakhmanova [M. Abjalova]. As a result of the studies, the following authorial studies were studied:

- Linguistic foundations of corpus construction [Sh.Khamroyeva]
- Natural language processing [M.Abjalova]
- Semantic tagging of synonymous words [A.Eshmuminov]
- Creation of an educational corpus of the Uzbek language [L.Raupov, B.Elov, M.Abjalova, R.Alayev]
- Lexical-semantic tagging of atov units [D.Akhmedova]
- Electronic corpora models
- Computer methods of corpus construction [A.Rahmanova]

As a result of these theoretical studies, the scientific team at the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi created the “Educational Corpus of the Uzbek Language”, thereby laying the foundation for the Uzbek National Corpus in Uzbek corpus studies. Currently, the text base in this corpus is being enriched. [M. Abjalova, 2021] Agreeing with the opinion that broader definitions are needed on this topic, this will be analyzed and studied in more detail in our future articles.

CONCLUSION

This article has reviewed and analyzed the theoretical aspects of corpus and corpus linguistics. In order to determine what corpus and corpus linguistics are and to see how difficult it is, definitions and explanations given by various scholars have been presented. With all the explanations provided in the article, the reader can draw his own conclusions on the topic, or vice versa, because there may not be a single, agreed-upon answer to this question, even by scholars. One of the reasons why scholars cannot agree on a universal definition can be based on the fact that corpora always depend on the goals and data that create them, and they are always changing. And again, this topic is always developing and changing, as noted. And summarizing the opinions of many scholars on the elements of the corpus considered in the article, it can be noted that. The most important of them are authenticity, representativeness, accurate size and balance.

References

1. Aarts, J. (1999). 'The description of language use'. H. Hasselgård & S. Oksefjell (eds) (1999). Out of Corpora: Studies in honour of Stig Johansson. Amsterdam and Atlanta;
2. Abjalova M. Korpus lingvistikasi. Uslubiy qo`llanma/ M.A Abjalova.—Toshkent: Bookmany print, 2022.—103 b.
3. Biber, D., Conrad, S. and Reppen, R. (1998), 'Corpus Linguistics. Investigating Language Structures and Use'. -Cambridge: Cambridge University Press., 1998.-286 p.
4. Dash, N.S., "Corpus based language teaching: A new method"//Proceedings of the Orientation-Cum-Workshop Programme on recent trends in language teaching, Institute of advanced studies in education, Kalkutta, India, 2010,-vol 13.1. 200-207 p.
5. Francis, N.W. "Language corpora B.C.", in J. Svartvik (ed.), 17 p.
6. Gatto, M. (2013). 'The Web as Corpus: Theory and Practice Cambridge: Cambridge University Press
7. Meyer Ch.F. English corpus linguistics: An Introduction. Cambridge University Press, 2002.-p.68.
8. McEnery, A. M., & Wilson, A.(2001). Corpus linguistics: An Introduction. Edinburgh: Edinburgh University Press

BRIGHT MIND
PUBLISHING

Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 03, March, 2025
brightmindpublishing.com
ISSN (E): 3061-6964

9. Sinclair, J. Corpus, concordance, collocation.—Oxford, UK: Oxford University Press., 1991-pp. 13-18.
10. Кратова Е. Б. Корпусная фразеография на материале немецкого языка.—дисс.канд.фил.наук.- Москва.: 2013.-296 с.
11. Плунгян В.А. Корпус как инструмент и как идеология о некоторых уроках современной корпусной лингвистики// Русский язык в научном освещении.—Москва, 2008.-№ 02 (16).-с. 7-20.
12. Рыков В.В. Корпус текстов как онтология речевой деятельности и труды международного семинара Диалог. -М.: Наука, 2004-59-61с.