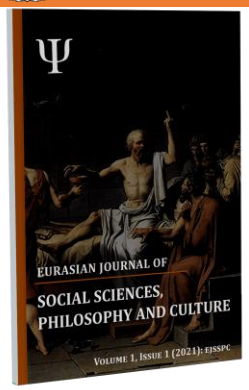
EURASIAN JOURNAL OF SOCIAL SCIENCES, PHILOSOPHY AND CULTURE
Innovative Academy Research Support Center
UIF = 8.2 | SJIF = 6.051 www.in-academy.uz




# EXPLORING THE POTENTIAL OF CORPUS LINGUISTICS IN LANGUAGE LEARNING AND TEACHING

**[1]Lolakhon Khamidovna Nigmatova**
Bukhara State University, nigmatovalolaxon@gmail.com,
**[2]Комышкова Анна Дмитриевна**
Нижегородский государственный педагогический университет имени Козьмы Минина, filcomanka@mail.ru.
https://www.doi.org/10.5281/zenodo.7732385

**ARTICLE INFO**

Received: 03rd March 2023
Accepted: 13th March 2023
Online: 14th March 2023

**KEY WORDS**

*Linguistic corpora, web corpora, corpus linguistics, corpus research, text markup, linguistic analysis, educational literature, foreign language teaching, search engine, linguistic parameters.*

**ABSTRACT**

*The objective of this study is to explore the potential utilization of linguistic corpora in the investigation of foreign languages. The subject under investigation is the conception of establishing an English language corpus that reflects the regional and dialectical peculiarities linked to globalization. The authors delve into the process of constructing language corpora, their various types and structures, and the current trends that influence the study of the English language. Corpus linguistics provides a framework for examining the communicative use of the English language, which is predicated on extensive volumes of speech. Corpus-based research can serve as a foundation for designing curricula and composing textbooks in English, delivering lectures and specialized courses on intercultural communication, linguistic and cultural studies, cognitive linguistics, and foreign language teaching methodologies.*

## Introduction

In contemporary times, a key domain of scholarly exploration within the field of English language acquisition is “corpus linguistics”. This area of study is concerned with the evolution, construction, and utilization of text corpora, and the term itself came into existence during the 1960s, concomitant with the establishment of the practice of creating corpora, which was bolstered by the advent of computer technology from the 1980s onward. The need for such investigations in the context of modern English studies is prompted by both linguistic and historical factors. The latter encompasses, among other things, the need to enhance the practice of translation, as well as proficiency levels in foreign languages. The movement towards internationalization and societal integration in European countries has engendered the prerequisites for a resurgence of interest in comparative research in the setting of greater demands for proficiency in foreign languages, as stipulated by the international community.

## Results and Discussion

In the field of linguistics, a corpus refers to a specific collection of texts that has been selected and processed according to certain criteria and rules for the purpose of studying a language. Corpora are commonly used for statistical analysis, as well as for testing and confirming linguistic rules and hypotheses within a given language. The term “corpus of the first order” may be used to refer to any collection of texts that share a common feature, such as language, genre, authorship, or time period. According to the definition put forth by V.P. Zakharov and S.Yu. Bogdanova, a linguistic corpus is a large, structured, and labeled array of machine-readable linguistic data that is designed to address specific linguistic problems. The main characteristics of a modern corpus include its machine-readable format, representativeness, and the inclusion of metalinguistic information.

The selection of texts for a corpus is done through a specific procedure to ensure its representativeness. The creation of text corpora is deemed advantageous due to the provision of linguistic data in an actual context, as well as the substantial representativeness of data resulting from the extensive volume of the corpus. Moreover, a corpus created once can be utilized for various linguistic purposes, including graphematic and lexical-grammatical analyses of texts, among others. [1]

There are numerous definitions of a corpus, but its fundamental characteristics include: being in electronic form, possessing representativeness by accurately modeling its object, being marked in contrast to a mere collection of texts, and having a pragmatic orientation designed for a specific purpose.

English is a language of significant global importance, serving as a primary means of international communication. Its reach is unparalleled, with its use prevalent in various regions around the world. The United Kingdom, the United States, and Australia are among the primary English-speaking countries. English is also recognized as a secondary official language in education and administration in countries such as India, Kenya, Nigeria, and Singapore. Additionally, many countries that use English as a foreign language offer significant opportunities for further expansion of the “English-speaking” population [2].

While there is no unanimity among scholars about the exact proportion of native English speakers to non-native English speakers, the majority tend to agree that the latter outnumber the former by a ratio of 2:1. Such a situation has undoubtedly had an impact on the general approach taken to studying and comprehending English, with its diverse functions and variations.

The concept of “non-native English” or English as a foreign language (EFL) has gained importance in linguistic research and teaching practice. The dominance of English as a native language is being challenged by EFL, with some experts predicting that the latter will soon surpass the former. Therefore, the preservation of a single linguistic culture depends on the quality of EFL. The need to generalize the diverse manifestations of English, including dialects and registers, led to the creation of a corpus of the English language. The reliability of comparative analysis data for revealing the peculiarities of dialects was a challenge, and a uniform collection and consideration of material was necessary to ensure reliable results.

The International Corpus of English language was created to ensure the comparability of text samples for the purpose of the project. The project manager, S. Greenbaum, followed the British descriptive linguistics tradition by including new materials that reflect regional and

dialectal characteristics in the corpus. In today's world, where countries that have English as their second official language are striving for linguistic independence, it is essential to maintain some consistency in the language. The creation of an international corpus aimed to preserve the identity of at least the written form of English language.

Although the availability of electronic texts has made it easier to create large and representative language corpora containing tens and hundreds of millions of words, there are still several challenges that need to be addressed, such as collecting and processing a large number of texts, resolving copyright issues, and ensuring uniformity across the corpus. Additionally, creating a balanced corpus that includes a range of themes and genres requires significant effort and time. Nevertheless, there are currently representative corpora in existence or being developed for various languages, including German, Polish, Czech, Slovenian, Finnish, Modern Greek, Armenian, Chinese, Japanese, Bulgarian, and others.

The Brown Corpus (BC) is recognized as the earliest significant computer corpus, which was developed in 1960 at Brown University. The corpus comprised 500 text samples, each containing 2,000 words, published in the United States in 1961. Its size, which consisted of 1 million tokens, established a benchmark for constructing representative corpora in other languages.

In the 1970s, a frequency dictionary of the Russian language was created based on a corpus of texts containing 1 million words, using a model similar to that of the Brown Corpus. The corpus comprised roughly equal proportions of socio-political texts, fiction, scientific and popular science texts, and drama from various fields. A Russian corpus was also developed in the 1980s at Uppsala University in Sweden, following a similar model. However, a corpus of one million words is sufficient only for the lexicographic description of the most common words, since words and grammatical constructions of average frequency occur only a few times per million words. From a statistical perspective, a language is composed of a vast collection of infrequent events.

One of the prominent ongoing projects in the field of corpus linguistics is the British Corpus of the English language, which comprises over 100 million words. This corpus is carefully designed with a balanced collection of over 4,000 texts that represent a range of genres and language varieties, including spoken English, newspaper articles, and complete novels. It can be used as a valuable resource for both English language teaching and research purposes, as it provides authentic examples of contemporary language usage and can help identify emerging language trends.

As common words such as “polite” or “sunshine” appear only 7 times in the Brown Corpus, and phrases like “polite conversation”, “polite smile”, and “polite request” never occur, efforts were made in the 1980s to create larger corpora due to the limitations of the Brown Corpus and advancements in computer technology capable of handling larger amounts of textual data.

The Bank of English, created by Cobuild Collins, is the most ambitious and widely known project in corpus linguistics. It currently comprises 200 million words, including 15 million spoken texts, and provides a vast amount of data for scientific and educational publications, such as the Cobuild English Language Dictionary. Other research centers have also contributed to the development of corpus linguistics in various areas, including the creation of

corpora of translations into English from European languages, studies of oral communication in English between speakers from different countries, and investigations into Euro-English, the official English used by the European Commission.

The creation of large-scale international databases containing extensive corpora of language use has facilitated the development of various areas of British linguistics, particularly text organization, pragmatics, and discourse. As a result, several works have emerged that emphasize the importance of considering the act of communication and real-life speech data rather than solely focusing on internal semantics to establish a comprehensive model of language. This approach emphasizes the significance of contextual factors in the interpretation of language use. [3]

As a result, the speaker's speech activity became the center of attention, which depended not only on his or her “linguistic competence”, but also on the pragmatic ability to construct speech in a particular communicative context. The notion of context, as well as the frequency of the use of a linguistic unit in determining the meaning of a word and its position in the structure of the language's vocabulary, became particularly important.

The Collins Cobuild English Language Dictionary is a pioneering work that is noteworthy for being one of the earliest English dictionaries to have been developed using a computerized corpus. Its unique approach is to prioritize the meanings of words based on their frequency of usage in established contexts, rather than providing the literal meaning of a word as the primary entry. For instance, the noun “way” is first described in Cobuild Collins as “a way of doing something or a way to do it” (way, method, manner), which is one of its figurative meanings. In contrast, most other dictionaries list the direct meaning of “way” (“the way” – “the way to a particular place”) as the primary entry, which is relegated to the end in Cobuild Collins.

In this instance, the primary focus is not on a single meaning within the internal semantic structure of a word, but on the meaning that is realized within common and repetitive contexts. This particular meaning holds a significant social significance because of its high frequency of usage within discourse. While the nominative meaning is typically considered to be the primary meaning in the development of a word's semantics, in the realm of actual speech, priorities can shift depending on the needs of the speakers.

To understand the practical use of words in speech, it is necessary to take a broader perspective and consider the “grammar of the word” and its syntactic functions. This approach leads to the emergence of a textological study of lexical units that seeks to establish the correlation between vocabulary and syntagmatics, rather than treating them as separate entities. A particular focus is placed on the word within the context of a text, which constitutes a single speech complex characterized by the unity of its semantic connection and integrity.

Corpus linguistics presents various opportunities for the development of a new direction in linguistic research. First and foremost, it allows for the study of the communicative use of language through large datasets of actual speech, as opposed to regulated data found in normative publications that primarily serve to illustrate grammatical rules. Discourse, as a form of “speech immersed in life”, necessitates the adoption of a new methodology that takes into account both the grammatical-syntactic and semantic-pragmatic aspects of text, while also providing contextually appropriate interpretations of each

statement. Linguists view the figure of the speaking subject as playing a crucial role in the speech process. Such an analysis provides significant advantages in identifying the particularities of registers and functional styles. Corpus studies of English have identified significant discrepancies between spoken and written forms of the language, revealing stable grammatical deviations from standard (written) speech that may require the creation of a new grammar for English that reflects its spoken form.

At present, corpus linguistics exerts a significant influence on the teaching of English as a foreign language. By comparing authentic English speech and texts, researchers can identify features associated with the influence of speakers' native languages on their speech activity. The “English as a Foreign Language” international corpus is an invaluable resource for conducting comparative or contrastive studies in two directions.

Having access to dependable comparative data enables researchers to explore the distinctive speech patterns of different groups of students, based on their first language or mother tongue. This allows for the identification of both general or invariant errors in English, common to all students, as well as particular phenomena arising from the influence of specific languages. In essence, it becomes feasible to compare different varieties of English as a foreign language, reflecting the speech activity of students from countries such as Germany, France, Russia, and others.

One example of how corpus linguistics can be used in teaching English as a foreign language is the identification of difficulties that arise for students whose native language is Russian, with respect to the use of the passive voice. This is because the passive voice is used much more frequently in English than in Russian. Computer-based comparisons of texts have revealed a relatively low frequency of passive voice usage in the Russian segment of the corpus, which supports this claim. However, this error should be considered a general or invariant difficulty, rather than being attributed solely to the influence of the Russian language, as a similar problem is observed in other sections of the corpus, such as the French language segment.

As corpus linguistics is applied in the field of language teaching, two main directions have emerged: 1) contrastive analysis, which helps students identify challenging elements of a foreign language that significantly differ from their native language, and 2) discourse analysis, which identifies the most frequent, typical, well-established, and therefore “socially significant” features of real speech use. The latter is particularly important, as opposed to systematized rules, since the corpus demonstrates what happens in live speech and what tendencies can manifest themselves in certain contexts.

Contemporary technological advancements have enabled the development of “web corpora”, which are formed by utilizing automated techniques to collect texts from Internet sources. These corpora are a distinct type of linguistic corpus.

Automated procedures can be used to create web corpora, which are linguistic corpora obtained by downloading texts from the internet. These procedures determine the language and encoding of individual web pages, remove unwanted elements such as boilerplate, and transform the documents into text format. The data obtained can then be processed using traditional tools of corpus linguistics and introduced into the search corpus system. Creating a web corpus is cheaper and can provide larger corpora than traditional methods. Linguists can

use a method called Googleology to work with the internet, which involves composing queries to a search engine and interpreting the results. However, this approach has limitations since text markup tools used on the web do not provide information about some linguistic features such as stress, grammatical classes, and phrase boundaries.

The second approach involves extracting a large number of web pages from the internet and treating them as a regular corpus. This method allows for the annotation and use of linguistic parameters in queries, making it possible to quickly create a representative corpus for any language that is well-represented on the internet. However, the genre and thematic diversity of the corpus will reflect the interests of internet users. Using Wikipedia as a corpus of texts is also becoming increasingly popular in the scientific community, as computer processing of the material allows for a systematic description of language aspects in discourse, while taking into account the peculiarities of text organization and speech pragmatics.

There are various criteria that can be used to classify corpora, such as their purpose, the type of linguistic data they contain, their literary or genre focus, their level of dynamism, the type of markup used, and the volume of texts included. One criterion for classification is parallelism, which can divide corpora into monolingual, bilingual, and multilingual categories. Within the multilingual and bilingual categories, there are two types: 1) parallel corpora, which consist of many texts and their translations into one or more languages; and 2) comparable (or pseudo-parallel) corpora, which include original texts in two or more languages.

Marking up of texts involves assigning special tags to various components of texts, including linguistic and extralinguistic elements. The types of linguistic markup may include semantic, morphological, syntactic, anaphoric, prosodic, discourse, among others. Some corpora may be analyzed at deeper structural levels, such as full syntactic markup in small corpora, which are referred to as syntactic or deeply annotated corpora. Manual annotation of texts is a labor-intensive and costly process, but there are several publicly available software tools for text markup [4]. These tools can be broadly categorized as standalone or web-based.

Recently, developers have been increasingly emphasizing web applications. These systems have several benefits, including the ability to allow several people to mark the same document simultaneously, no need to install additional software aside from the browser, flexible differentiation of access rights, display of the current progress of the marking process, and the ability to modify the markup body.

**Conclusion**

In corpus linguistics, a corpus refers to a collection of texts that are gathered and marked up according to a particular standard and made accessible through a specialized search engine. Corpus research can be used as a foundation for developing English language textbooks and assist specialists in identifying linguistic phenomena that require particular attention. By analyzing a corpus of students' language use, typical errors can be identified and used to inform the structure of language courses. This includes correcting both individual grammatical and syntactic constructions, as well as more general issues related to lack of idiomatic expression. The future of corpus linguistics holds promise for improving foreign

language education, and it is possible that an international database may serve as a basis for a more student-focused generation of educational materials.

## References:

1. Sobirovich A. S. Development of a Parallel Corpus of the Uzbek and Russian Languages //Vital Annex: International Journal of Novel Research in Advanced Sciences. – 2022. – T. 1. – №. 5. – C. 152-155.
2. Khamidovna N. L. Expression of the Harmony of Language and Culture in World and Uzbek Lexicography //resmilitaris. – 2023. – T. 13. – №. 1. – C. 233-244.
3. Nigmatova L. K. Language and cultural issues in uzbek vocabulary //Scientific reports of Bukhara State University. – 2021. – T. 5. – №. 1. – C. 30-49.
4. Avezov S. S. MACHINE TRANSLATION TO ALIGN PARALLEL TEXTS //International Scientific and Current Research Conferences. – 2022. – C. 64-66.