THEORETICAL APPROACHES TO LINGUISTIC CORPORA

https://zenodo.org/records/18622675

Fayzullayevna Y. Kh.
1st year student of Oriental university
Scientific advisor: ***Boboeva M. Kh. (PhD),***
Oriental university, Tashkent, Uzbekistan
yulduzxasanova@gmail.com

Abstract. This article provides a comprehensive analysis of the concept of linguistic corpora, their theoretical foundations, and their significance in modern linguistics. The study examines the main types of linguistic corpora, their structural organization, and the principles underlying their compilation and development. Particular attention is paid to corpus linguistics as one of the most dynamically developing branches of contemporary linguistic science. The article also explores the advantages of corpus-based research, highlighting its empirical nature and objectivity in linguistic analysis. Furthermore, various types of linguistic annotation are discussed, emphasizing their role in enhancing the accuracy and depth of corpus analysis. Special consideration is given to the importance of electronic linguistic resources in language learning, linguistic research, and language teaching practices.

Keywords: *Linguistic corpus, corpus linguistics, text annotation, electronic linguistic resources, language analysis, empirical linguistics*

Annotatsiya. Mazkur maqolada lingvistik korpuslar tushunchasi, ularning nazariy asoslari hamda zamonaviy tilshunoslikdagi ahamiyati har tomonlama tahlil qilinadi. Tadqiqotda lingvistik korpuslarning asosiy turlari, ularning strukturaviy tashkil etilishi, shuningdek ularni tuzish va rivojlantirish tamoyillari ko‘rib chiqiladi. Zamonaviy tilshunoslik fanining eng jadal rivojlanayotgan yo‘nalishlaridan biri sifatida korpus lingvistikaga alohida e’tibor qaratiladi. Shuningdek, maqolada korpusga asoslangan tadqiqotlarning afzalliklari tahlil qilinib, ularning empirik xususiyati va lingvistik tahlildagi obyektivligi yoritib beriladi. Bundan tashqari, lingvistik annotatsiyaning turli ko‘rinishlari muhokama qilinib, ularning korpus tahlilining aniqligi va chuqurligini oshirishdagi roli ta’kidlanadi. Elektron lingvistik resurslarning til o‘rganish, lingvistik tadqiqotlar va til o‘qitish amaliyotidagi ahamiyatiga ham alohida e’tibor qaratiladi.

Kalit so‘zlar: *lingvistik korpus, korpus lingvistika, matn annotatsiyasi, elektron lingvistik resurslar, til tahlili, empirik lingvistika.*

Аннотация. В данной статье представлен всесторонний анализ понятия лингвистических корпусов, их теоретических основ и значения в современной лингвистике. В исследовании рассматриваются основные типы лингвистических корпусов, их структурная организация, а также принципы их создания и развития. Особое внимание уделяется корпусной лингвистике как одной из наиболее динамично развивающихся отраслей современной лингвистической науки. В статье также анализируются преимущества корпусных исследований, подчёркивается их эмпирический характер и

объективность в лингвистическом анализе. Кроме того, рассматриваются различные виды лингвистической аннотации, акцентируется их роль в повышении точности и глубины корпусного анализа. Отдельное внимание уделяется значению электронных лингвистических ресурсов в изучении языка, лингвистических исследованиях и практике преподавания языков.

Ключевые слова: *лингвистический корпус, корпусная лингвистика, аннотация текста, электронные лингвистические ресурсы, анализ языка, эмпирическая лингвистика.*

INTRODUCTION: The development of modern linguistics is directly intertwined with the evolution of information technologies. As a result of the widespread application of digital technologies, new methods of studying, analyzing, and describing language have emerged. One of the most advanced directions in this field is corpus linguistics. While traditional linguistics often interpreted linguistic phenomena based on a linguist's personal observations and intuitive views, corpus linguistics necessitates conducting empirical research based on authentic speech materials. Linguistic corpora provide the opportunity to determine how language units are utilized within their natural environment [1].

A linguistic corpus is a collection of texts stored in electronic form, structured according to specific selection criteria, and linguistically annotated. Corpora can encompass both written and spoken forms of language, being firmly rooted in real-life speech samples.

Theoretically, corpus linguistics develops by relying on an empirical approach. According to this perspective, scientific conclusions about language must be drawn from real-world data[2]. Corpora enable the determination of the frequency of language units, their contextual usage, as well as their grammatical and semantical characteristics. Scholars such as J. Sinclair and D. Biber established the theoretical foundation of corpus linguistics; they emphasized the vital importance of large-scale text collections in language study. In their view, the true meaning and function of linguistic units are revealed only within a real context.

The scientific value of linguistic corpora lies not only in their large size but also in the principles on which they are compiled. The process of creating a corpus is complex and multi-stage, requiring a combination of linguistic, technical, and methodological approaches. Therefore, an in-depth study of the structure and principles of corpus formation is one of the important issues in corpus linguistics.

Linguistic corpora are compiled according to certain scientific requirements. These requirements ensure the reliability of the corpus and its effective use in research. The main principles followed when compiling a corpus are as follows:

Principle of Authenticity – Texts included in the corpus must be produced in real communication. Artificial or specially constructed texts reduce the scientific value of the corpus. Authentic texts reflect the natural use of language units.

Principle of Representativeness – A corpus should adequately represent the general characteristics of a particular language or language layer. This principle ensures that conclusions drawn from language units are generalizable and reliable.

Principle of Balance – Texts in the corpus should be selected in a balanced manner according to genre, style, topic, and communicative situation. For example, a corpus consisting solely of literary texts cannot provide a complete picture of a language.

Principle of Size – Since modern linguistic research is often based on statistical analysis, the corpus must have a sufficiently large volume. The larger the corpus, the more accurate the results.

Linguistic corpora have a complex structure and consist of several main components, which enable effective use of the corpus.

Text Database – This forms the core of the corpus. The database includes written and spoken language samples from various sources. Texts are usually converted into electronic format and stored in a unified system.

Metadata – Metadata is an important component of a corpus. It includes additional information about the text, such as author, date of creation, genre, topic, and style. This information allows analysis of the corpus according to different criteria.

Search and Analysis Tools – These tools play an essential role in corpus use. Modern corpora are equipped with specialized software and interfaces, allowing users to quickly and efficiently search for language units.

The process of creating a linguistic corpus usually involves several consecutive stages:

Defining the Purpose and Scope – The target language or language layer that the corpus will cover is determined.

Selection and Digitization of Texts – Texts are selected and digitized, with special attention to their quality and reliability.

Linguistic Annotation – Texts are linguistically annotated, which expands the possibilities for using the corpus in scientific research.

Testing and Correction – The corpus is tested, errors are corrected, and only then is it introduced into scientific circulation.

The development of information technologies has significantly influenced the structure of linguistic corpora. Today, corpora are being integrated with artificial intelligence, machine learning, and automated analysis systems. This increases the speed and accuracy of corpus-based research. Modern corpora are also made available through web platforms for open access, further popularizing linguistic research.

Linguistic corpora are classified into several types depending on their purpose, area of application, and structural characteristics. Each type of corpus serves a specific scientific function and provides different opportunities for studying language. Below are the most important and widely used types of linguistic corpora. General or national corpora are designed to fully reflect the overall state of a language, including its lexical, grammatical, and stylistic features. These corpora include texts from various genres and styles, such as literary works, journalism, scientific texts, mass media, and examples of everyday speech. The main goal of national corpora is to study the actual usage of a language on a scientific basis. For example, the British National Corpus (BNC) is one of the large corpora representing the state of English at the end of the 20th century. Such corpora allow linguists to determine word frequency, collocations, and grammatical

constructions. National corpora also serve as important resources for establishing normative language standards, compiling modern dictionaries, and developing language policy. Specialized corpora consist of texts related to a specific field, topic, or communicative situation. They are intended not to study the general state of the language but to investigate features in a particular domain. Examples include corpora of medical, legal, technical, economic, or academic texts. These corpora are valuable for terminological research, analyzing professional discourse, and creating domain-specific dictionaries. Specialized corpora help identify the contextual meaning, usage frequency, and stylistic characteristics of terms used in a particular field. They are also widely used in translation practice to ensure accurate and precise translation of domain-specific terminology [3].

Parallel corpora consist of texts with the same content in two or more languages. They are usually based on translated texts and are used to study cross-linguistic equivalence. In parallel corpora, texts are aligned sentence by sentence or paragraph by paragraph. They are very important for translation theory and practice, as they allow the analysis of translation strategies, equivalence issues, and grammatical differences. Parallel corpora also serve as key resources in developing machine translation systems. Multilingual corpora include texts in several languages and enable cross-linguistic comparative research. Learner corpora are compiled from written and spoken language samples of individuals learning a foreign language. Their main purpose is to identify learners' errors, their causes, and the dynamics of language development. Learner corpora help teachers and researchers identify typical errors of language learners and develop effective teaching methodologies, thereby improving the quality of language education. In modern linguistics, spoken corpora and web corpora are also of particular importance. Spoken corpora are compiled from conversations, interviews, and speech transcripts. Web corpora are created from texts available on the Internet. These types of corpora are important for studying the rapidly changing characteristics of modern language and the emergence of new words and expressions.

CONCLUSION: This article provides a comprehensive analysis of linguistic corpora, their theoretical foundations, and their role in modern linguistics. It highlights the empirical nature of corpus linguistics and the ability of corpora to determine the frequency, contextual usage, and grammatical-semantic features of language units. The main types of linguistic corpora — general (national), specialized (domain-specific), parallel, multilingual, learner, spoken, and web corpora — are described in detail. The process of linguistic annotation and its types (morphological, syntactic, semantic, pragmatic) are discussed, emphasizing their role in facilitating effective scientific research and simplifying automatic analysis. The article also underscores the importance of electronic linguistic resources in language learning, analysis, and teaching.

References:

1. Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

2. Biber, D., Conrad, S., & Reppen , R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

3. British National Corpus (BNC). (2007). *BNC Documentation*. Oxford: Oxford University Press.

4. Anthony, L. (2013). *A Critical Introduction to Corpus Linguistics*. London: Bloomsbury Academic.

5. McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

6. Gries, S. Th. (2009). *Statistics for Linguistics with R: A Practical Introduction*. Berlin: De Gruyter Mouton.

7. Johansson, S., & Hofland, K. (1994). *Word Frequency and Lexical Analysis*. Oslo: University of Oslo.