ISSN 2181-3187

# CORPUS LINGUISTICS: THEORETICAL FOUNDATIONS AND STAGES OF DEVELOPMENT

**Shakhzoda Normamatova**
1st-year student,
Department of Foreign Languages and Literature (English),
National University of Uzbekistan

**Abstract:** This article thoroughly examines the theoretical foundations of corpus linguistics and its stages of development. It also provides a detailed discussion of its methodology, key concepts, technology used, and important scientists who helped this field grow.

1. **Key words:** corpora, concordance, Roberto Busa, Bible, parallel corpus, synchronic and diachronic corpora, specialized corpus

Corpus linguistics is a modern method of studying language by using large collections of real texts, called *corpora* [1]. These texts are used to find patterns and meanings in how people actually use language in everyday life. With the help of computers, corpus linguistics allows researchers to look at thousands or even millions of words to understand grammar, vocabulary, and communication more clearly. The main ideas behind corpus linguistics are based on real examples, careful observation, and the study of how often certain words or structures appear. These ideas help researchers study language in a more objective and scientific way.

Let's talk about **why this area is crucial**. Corpus linguistics is important because it helps us understand how language is really used in everyday life. Instead of only looking at grammar rules in books or made-up examples, it uses real texts—like conversations, newspapers, websites, and books—to study how people speak and write. This method gives researchers and language learners a more accurate picture of language. For example, it shows which words are most common, how grammar is used

ISSN 2181-3187

in real situations, and how language changes over time. It also helps create better dictionaries, language-learning materials, and translation tools. Corpus linguistics is also useful in many areas, such as education, translation, language teaching, and even in technology like speech recognition and machine translation. Because it is based on real data, it helps make decisions about language more scientific and reliable.

**Stages of development:** The development of software tools for corpus analysis can be traced back to 1951, when Roberto Busa initiated one of the earliest projects involving machine-readable texts. He created the first electronic corpora and performed the earliest computerized concordances (a concordance is a system or a list that shows where and in what context each word is used within a text or a corpus) [1] . While Busa did not invent the idea of concordances — since manually created concordances had existed for centuries — his work marked a major shift in how they were produced.

What made Busa's contribution revolutionary was his demonstration that concordances could be generated efficiently using computers, turning what had been a slow, manual task into a much faster and more scalable process. His efforts transformed concordancing from something reserved for a few culturally significant texts, like the Bible or Shakespeare's works, into a method that could be applied to any text. Busa's work thus laid the foundation for what we now refer to as first-generation concordancing tools.

The development of corpus linguistics covers several important stages over many years. In the first stage, scholars created concordances **by hand**. For example, in 1230, Hugh of St. Cher made a concordance for the Latin Bible [1], which is the holy book of Christians. This method was very slow and required a lot of effort. At that time, the idea of a corpus did not exist yet. In the next stage, as mentioned earlier, Roberto Busa developed the first machine-readable corpus. This was an important step forward. By the 1980s and 1990s, corpus linguistics started to become a separate field of study.

ISSN 2181-3187

During this time, John Sinclair supported an empirical approach to language based on real usage in corpora. From the 2000s to today, we can call this the globalization stage. Nowadays, corpora are becoming multilingual, and we also have parallel corpora and lexicographic corpora for use in different areas of linguistics.

In addition, corpus linguistics also studies which forms of a word are used more in different language skills [2]. For example, research has shown that the word "really" is more common in speaking, "quite" is used more in writing, and "very" is one of the most frequent words in both speaking and writing.

**Frequency results per million of adverbs of degree in COCA**

| Word | Speak | Write | Total |
|---|---|---|---|
| very | 2,543 | 673 | 3,216 |
| really | 1,637 | 392 | 2,029 |
| exactly | 271 | 93 | 364 |
| quite | 267 | 150 | 417 |
| completely | 87 | 78 | 165 |
| too | 656 | 699 | 1,355 |
| thoroughly | 7 | 18 | 25 |
| Total | 5,468 | 2,103 | 7,571 |

Source: Corpus of Contemporary American English

There are different types of corpora, and each of them covers information related to a specific type:[3]

2. **General-purpose corpus**: A general-purpose corpus is a large collection of texts that includes many styles and topics, such as literature, news, science, official documents, fiction, spoken and written language. This type of corpus is used to study

ISSN 2181-3187

how the language is used in general. Example: The British National Corpus (BNC) – a multi-topic corpus of the English language.

3. **Specialized corpus**: A specialized corpus is a group of texts from one specific topic or field. It helps to study how people use language in a special area, like law, medicine, or technology. These corpora are useful for translators, professionals, and researchers.

4. **Parallel corpus**: A parallel corpus has the same texts in two or more languages. It shows the original text and its translations side by side. This helps compare grammar, vocabulary, and style between the languages.

**5. Synchronic and Diachronic corpora**

a) Synchronic corpus: This corpus includes texts from only one period of time. It shows how the language was used during that time.

b) Diachronic corpus: This corpus includes texts from different time periods. It helps study how the language has changed over time.

As can be seen from the above, corpus linguistics has become one of the most important and active areas in modern linguistics. Its main idea is based on studying real examples of how people use language in everyday life. Instead of only using theories or personal opinion, corpus linguistics uses real texts—called corpora—to study language in a more scientific way. These texts are collected and organized in large databases, which help researchers find patterns in grammar, vocabulary, and meaning. This method gives more accurate and reliable results compared to traditional ways of studying language. Its stages of development show that corpus-based methods are becoming more advanced and multifunctional. As language continues to change and technologies for data collection and analysis improve, corpus linguistics will remain one of the leading fields in linguistic research. Because of this, corpus linguistics has changed how we understand and research language today.

ISSN 2181-3187

## LIST OF REFERENCES:

1. Tony McEnery, Andrew Hardie. “Corpus Linguistics: Method, Theory and Practice. Cambridge University Press. 2012.

2. Kennedy.G. An Introduction to Corpus Linguistics. 1998.

3. John Sinclair. Corpus, Concordance, Collocation. Cambridge University Press. 1991.