# INTRODUCTION TO CORPUS LINGUISTICS AND ITS HISTORY

***Ibrohimova Mohichehra***
*Bukhara State University, Faculty of Foreign Languages*

***Key words:*** *Methods, corpus, linguistics, language, data, patterns, collocation, analysis*

## Abstract

Corpus linguistics is a field within linguistics that utilizes extensive sets of authentic language data, referred to as corpora, to investigate language patterns. These corpora can vary in size, ranging from niche collections to colossal databases housing billions of words. The roots of corpus linguistics can be identified in the early 1900s when academics started utilizing text collections for language analysis. Nonetheless, it was with the introduction of computers in the 1960s and 1970s that corpus linguistics evolved into a defined area of academic research.

Corpus linguistics is a method that uses computers to study language by analyzing large collections of natural spoken and written texts known as corpora. Research using this approach has demonstrated that relying solely on speakers' intuitions may not always fully capture the complexities of language, especially when exploring less common linguistic patterns like word combinations, grammar variations, meanings, idioms, and metaphors.

## Introduction

Corpus linguistics encompasses the compilation and analysis of collections of spoken and written texts as the source of evidence for describing the nature, structure, and use of languages. This work typically brings a quantitative dimension to the description of languages by including information on the probability with which linguistic items or processes occur in particular contexts.[1]

Corpus linguistics is a branch of linguistics that involves the study of language using large, structured collections of texts known as corpora. These corpora serve as databases that linguists analyze to investigate language patterns, usage, and structure.

Corpora come in different sizes and formats, but modern ones are largely electronic and come with specialized software for analysis. The field of corpus linguistics has evolved from a history of using texts for linguistic research, examining various aspects of language such as phonology, vocabulary, syntax, and

[1] G. Kennedy, in International Encyclopedia of the Social & Behavioral Sciences, 2001

communication. Many corpora are annotated to indicate grammatical categories and uses. A corpus language is a language that has no living speakers but for which numerous records produced by its native speakers survive. Examples of corpus languages are Ancient Greek, Latin, the Egyptian Language, Old English and Elamite.[2]

Some languages have a large corpus, like Ancient Greek and Latin, making it possible to fully reconstruct them despite some uncertainties in pronunciation. These languages, such as Sanskrit and Latin, remain relevant even today. On the other hand, languages like Ugaritic and Gothic have limited corpus, missing important words like pronouns. Some languages, known as "rubble languages," have very few surviving words and phrases, making it challenging to reconstruct them fully or determine their relationships with other languages, like Lombardic and Dadanitic. While corpus languages are studied using corpus linguistics methods, this approach can also be applied to records of living languages. However, not all extinct languages are corpus languages, as many lack surviving writings or records. Corpus linguistics is based on the idea that the best way to study language is to examine how it is actually used by real people in real-world situations. By studying corpora, linguists can gain insights into the structure, meaning, and use of language that would not be possible with traditional methods of language study.

Important Concepts in Corpus Linguistics:

- Corpus: A compilation of authentic language data from real-world sources.
- Concordance: A comprehensive list showing all instances of a specific word or phrase within a corpus.
- Collocation: Words that tend to appear together frequently in a sequence.
- Frequency: The number of times a specific word or phrase appears within a corpus.
- Distribution: The manner in which a word or phrase is utilized in various contexts.

## *History*

The history of corpus linguistics can be traced back to the early 20th century, when scholars began to use collections of texts to study language. However, it was not until the advent of computers in the 1960s and 1970s [3]that corpus linguistics began to develop into a distinct field of study.

One of the pioneers of corpus linguistics was John Sinclair, who developed the notion of the "corpus" as a representative sample of a language. Sinclair argued that corpora could be used to study language in a more objective and scientific way than had been possible with traditional methods.

A landmark in modern corpus linguistics was the publication of Computational Analysis of Present-Day American English in 1967. Written by

[2] https://en.m.wikipedia.org/wiki/Corpus_language
[3] www.studysmarter.co.uk

Henry Kučera and W. Nelson Francis, the work was based on an analysis of the Brown Corpus, which was a contemporary compilation of about a million American English words, carefully selected from a wide variety of sources.[4]
Corpus linguistics today is often understood as being a relatively new approach in linguistics that has to do with the empirical study of "real life" language use with the help of computers and electronic corpora. In the first instance, a "corpus" is simply any collection of written or spoken texts. However, when the term is employed with reference to modern linguistics, it tends to bear a number of connotations, among them machinereadable form, sampling and representativeness, finite size, and the idea that a corpus constitutes a standard reference for the language variety it represents. While linguistics divides up into many research areas depending on complexes of research questions, corpus linguistics in essence behaves diametrically: it offers a set of methods that can be used in the investigation of a large number of different research questions. For a number of reasons, we think that the time is right for a handbook on this approach: we now have access to large corpora and rather sophisticated tools to retrieve data from them. [5]Over the past few decades, corpus linguists have gained a great deal of experience in dealing with both theoretical and practical problems in their research. In other words, we are now much wiser about the ways in which legitimate claims can be made about language use on the basis of corpora. There is also a new focus on empirical data in theoretical linguistics, with growing interest in the techniques and procedures practised within the corpus linguistic approach. Our handbook is intended to sketch the history of corpus linguistics, and describe various methods of collecting, annotating and searching corpora as well as processing corpus data. It also reports on a number of case studies that illustrate the wide range of linguistic research questions discussed within the framework.

Corpus linguistics is anchored in a theoretical paradigm characterised by an empiricist approach and as well as by a conception of language as a probabilistic system. In linguistics, empiricism Empiricism is an approach that grants primordial status to data coming from the observation of language, generally grouped together in a corpus, as opposed to rationalism Rationalism. Rationalism is based on the study of language through introspection, which is regarded as a way of assessing models of structural functioning and the formation of the cognitive process of language. [6]As a result, there is a chasm between the philosophical perspectives characteristic of the empiricist and rationalist conceptions of language, represented by its main contributors.

[4] Francis, W. Nelson; Kučera, Henry (1 June 1967). Computational Analysis of Present-Day American English. Providence: Brown University Press.
[5] Anke Lüdeling, Kytö Merja Mouton de Gruyter, Corpus Linguistics 2008
[6] Carlos Assunção, Carla Sofia Araújo Linha D'Água 32 (1), 39-57, 2019

*Methodology*

- Corpus Compilation: The first step is to compile a corpus of real-world language data. This can be done by collecting texts from a variety of sources, such as books, newspapers, magazines, websites, and social media.
- Corpus Annotation: Once the corpus has been compiled, it may be necessary to annotate the data. This involves adding additional information to the corpus, such as part-of-speech tags, syntactic annotations, or semantic tags.
- Corpus Analysis: The next step is to analyze the corpus using a variety of techniques. This can involve using concordance software to find all the occurrences of a particular word or phrase, or using statistical software to analyze the frequency and distribution of words and phrases.
- Interpretation: The final step is to interpret the results of the corpus analysis. This involves drawing conclusions about the structure, meaning, and use of language.[7]

*Key Techniques in Corpus Linguistics:*

- Concordance: A concordance is a list of all the occurrences of a particular word or phrase in a corpus. Concordances can be used to study the different contexts in which a word or phrase is used.
- Collocation: Collocation is the study of the way that words tend to occur together. Corpus linguistics can be used to identify collocations and to study their frequency and distribution.
- Frequency: The frequency of a word or phrase in a corpus is a measure of how often it occurs. Corpus linguistics can be used to study the frequency of words and phrases and to identify the most common words and phrases in a language.
- Distribution: The distribution of a word or phrase in a corpus is a measure of how it is used across different contexts. Corpus linguistics can be used to study the distribution of words and phrases and to identify the different contexts in which they are used.

*Advantages of Using Corpora:*

- Objectivity: Corpora offer a sizable and representative sample of real language usage.

- Empirical Evidence: Data from corpora can either support or counter linguistic theories.

- Insights on Usage: Corpora provide insights on how language is genuinely utilized in diverse scenarios.

- New discoveries: Corpora may lead to fresh revelations about language and its application.

[7] https://www.ucl.ac.uk/english-usage/resources/ftfs/method..

Corpora can be used to study language variation. By comparing corpora from different dialects, registers, or time periods, researchers can identify the ways in which language varies.

Corpora can be used to study language change. By comparing corpora from different time periods, researchers can track the changes that have occurred in a language over time. Corpora can be used to develop language resources. Corpora can be used to create dictionaries, grammars, and other language resources. These resources can be used by language teachers, learners, and researchers. Corpora can be used to evaluate language models. Corpora can be used to test the accuracy of language models, such as those used in machine translation and natural language processing.

## Conclusion

Corpus linguistics is a relatively new discipline, and a fast-changing one. As computer resources, particularly web-based ones, develop, sophisticated corpus investigations come within the reach of the ordinary translator, language learner, or linguist. Our understanding of the ways that types of language might vary from one another, and our appreciation of the ways that words pattern in language, have been immeasurably improved by corpus studies. [8]Even more significant, perhaps, is the development of new theories of language that take corpus research as their starting point.

When discussing Corpus Linguistics, it is important to understand its various types and applications in the field of linguistics. Since Corpus Linguistics is a methodology rather than a subfield, it can be applied in numerous ways to investigate different aspects of language, such as phonetics, syntax, semantics, pragmatics, and sociolinguistics, among others. In this section, we will explore some of the commonly distinguished types of Corpus Linguistics research. Corpus linguistics is a powerful tool for studying language. It is a rapidly growing field of study with a bright future. Corpus linguistics is used by linguists, language teachers, lexicographers, and other researchers to study a wide range of linguistic phenomena and to develop language resources.

## References

- G. Kennedy, in International Encyclopedia of the Social & Behavioral Sciences, 2001
- Francis, W. Nelson; Kučera, Henry (1 June 1967). Computational Analysis of Present-Day American English. Providence: Brown University Press.
- Anke Lüdeling, Kytö Merja Mouton de Gruyter, Corpus Linguistics 2008
- Carlos Assunção, Carla Sofia Araújo Linha D'Água 32 (1), 39-57, 2019

[8] S. Hunston, in Encyclopedia of Language & Linguistics (Second Edition), 2006

- S. Hunston, in Encyclopedia of Language & Linguistics (Second Edition), 2006
- https://en.m.wikipedia.org/wiki/Corpus_language
- www.studysmarter.co.uk
- https://www.ucl.ac.uk/english-usage/resources/ftfs/method..