Website:https://worldsiencepub.com/index.php/Imc Volume:5,Issue 02,September-2026

THE MAIN PRINCIPLES AND METHODS OF CORPUS LINGUISTICS

Shermamatova Madina Obidjon qizi
Teacher and Researcher, "Behbudiy Vorislari" Private School
https://doi.org/10.5281/zenodo.22938347

ANNOTATION

This article discusses the main principles and research methods of corpus linguistics. Corpus linguistics is analyzed as a modern branch of linguistics aimed at the systematic and empirical study of language phenomena based on language corpora, which are large collections of naturally occurring written and spoken texts. The study reveals the content of such fundamental principles as corpus compilation, text selection and annotation, representativeness, balance, authenticity, and statistical reliability. It also examines the application of concordance, frequency analysis, collocation, n-gram analysis, keyword analysis, and statistical analysis in the study of corpus data.

Keywords: corpus linguistics, language corpus, corpus analysis, corpus data, representativeness, authenticity, balance, concordance, frequency analysis, collocation, computational linguistics, electronic corpus.

INTRODUCTION

At present, the rapid development of information and communication technologies has made the widespread introduction of modern digital tools at all levels of the education system one of the most relevant and priority tasks. In this regard, integrating the traditional educational process with corpus linguistics technologies and enriching it with modern pedagogical innovations presents important scientific, methodological, and practical tasks for specialists. Such an approach not only modernizes the educational process technically but also makes it possible to update its content and develop students' competencies in independent thinking and working with information. As a result, the teaching process is becoming more interactive, effective, and individualized.

The formalization of linguistic data is the process of coding and systematizing the basic units of natural language, namely words, word combinations, sentences, and texts, on the basis of precise rules, mathematical apparatus, and algorithmic systems that can be processed by computers. This process is one of the most important theoretical and methodological foundations of corpus linguistics and constitutes a fundamental condition for the automatic analysis, understanding, and generation of natural language. During the formalization process, linguistic phenomena are transformed into a precise, consistent, and unambiguous system, free from subjective human interpretation. In this process, the complex and multilayered structure of natural language is represented through simplified but formally consistent models. As a result, linguistic units are transformed into defined structures and codes understandable to computer systems, making automatic processing possible.

The widespread introduction of information and communication technologies into the educational process is currently one of the most important directions in the modernization of the higher education system. Digital technologies, particularly corpus linguistics methods, make it possible to improve the quality of education, optimize the learning process, and create more convenient and effective forms of learning for students. Therefore, organizing the educational process in higher education institutions on the basis of modern information technologies, strengthening interactivity in education, and developing students' independent learning skills have become important tasks. Today, the educational process is no longer limited to classroom activities; it also requires the integration of extracurricular activities into the digital learning environment. This necessitates the widespread use of electronic language corpora, digital textbooks, virtual laboratories, simulation platforms, and Internet-based learning systems. Such an approach contributes to making the educational process more flexible, individualized, and effective. In addition, the development of electronic teaching materials, improvement of corpus-based (Data-

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

Driven Learning) educational platforms, and effective use of Internet resources are becoming increasingly important in the modern education system. These tools not only facilitate the acquisition of educational materials but also help students develop independent thinking, information analysis, and practical skills.

The development of distance education significantly expands opportunities for learning. It eliminates geographical limitations and provides broad segments of the population with access to quality education. In addition, electronic language corpora and distance education contribute to increasing the competitiveness of the national education system in the international arena and to the development of educational service exports. In this regard, the integration of information technologies into education not only improves the pedagogical process but also has a positive impact on the strategic development of the entire education system.

In Uzbekistan, large-scale reforms are being implemented to consistently develop computerized teaching systems and improve them in accordance with modern requirements. In particular, the widespread introduction of information and communication technologies into the education system, the creation and practical implementation of electronic educational resources, as well as the fundamental modernization of the personnel training system have been identified as priority areas of state policy. This process contributes to improving the quality of education, digitalizing the learning process, and training competitive specialists with modern knowledge and skills.

A number of important decisions and regulatory documents are particularly significant in establishing the legal and regulatory foundations of these reforms. In particular, the Resolution of the First President of the Republic of Uzbekistan dated May 30, 2002, “On Measures for Further Development of Computerization and Introduction of Information and Communication Technologies,” is one of the important documents that defined the strategic directions in this field. This resolution created a strong legal basis for the widespread introduction of information technologies into the economy and social sectors, particularly the education system.

The resolution identified such priority tasks as modernizing the education sector, introducing modern pedagogical and information technologies into the educational process, and improving the system of training qualified and highly capable personnel. The document also considered increasing the intellectual potential of the country, bringing the education system closer to international standards, and developing civil society based on democratic principles and a market economy as important factors. In this regard, the resolution has served as an important programmatic and legal foundation for the formation and development of the digital education system in Uzbekistan. At present, as a result of the rapid development of computer hardware and software, their capabilities have expanded significantly. This contributes to making processes in education, science, production, and other fields more efficient and faster. In particular, the formation and rational use of knowledge bases in computer-based teaching systems are of particular importance. This situation contributes to the gradual improvement of traditional automated education systems and their transition to a higher level, namely intelligent teaching systems. Such systems not only provide information but also analyze students’ levels of knowledge and implement individualized instruction.

The integration of computer technologies with linguistics is currently one of the most important areas. Linguistics developed as an independent science beginning in the nineteenth century, and over time many new branches emerged. These include sociolinguistics, psycholinguistics, ethnolinguistics, neurolinguistics, mathematical linguistics, and computational linguistics, all of which contribute to the study of language from different perspectives. This demonstrates the growing importance of interdisciplinary integration and cooperation in modern linguistics.

From the middle of the twentieth century, the emergence of concepts such as “machine translation” and “machine linguistics” contributed to the development of this field and its transition to a new stage. During this period, ideas of formal language modeling developed, and attempts to express linguistic structures on a mathematical basis became more intensive. As a result, mathematical

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

linguistics emerged as an independent field, developing approaches for analyzing linguistic phenomena through precise mathematical and logical methods.

The European structuralist school, particularly the Copenhagen School of Linguistics and the theories of its representative Louis Hjelmslev, played an important role in the formation of mathematical linguistics. He emphasized the possibility of expressing language mathematically and even proposed the idea of a "language algebra." This approach became a basis for studying language as a formal system [1]. In addition, the theory of formal and transformational grammar developed by the American scholar Noam Chomsky had a significant influence on the development of mathematical linguistics [2]. His ideas made it possible to analyze language on the basis of strict rules and structural models. Thus, by the middle of the twentieth century, mathematical linguistics had developed as an independent scientific field and subsequently created the theoretical foundations of computational linguistics.

Formalization is one of the most important and initial stages of corpus linguistics. It refers to the process of transforming natural language phenomena into a precise and consistent system that can be understood and processed by computers. Although natural language is intuitive and ambiguous for humans, it must be represented for computers through precise rules, symbols, and algorithms. Therefore, linguistic units—sounds, words, word combinations, and sentences—are described on the basis of formal models.

During the formalization process, grammatical rules of language are represented in the form of mathematical and logical formulas. For example, sentence structure can be described using syntactic trees, systems of rules, or formal grammars such as HPSG or LFG. The structure and transformations of words are represented through morphological models. This process makes it possible to algorithmize language, that is, to create a system of step-by-step operations that can be performed by a computer. Formalization also helps organize complex aspects of language such as polysemy, synonymy, and homonymy. This enables computers to interpret texts correctly, determine context, and generate appropriate responses. Without formalization, it would be almost impossible to develop systems for natural language processing (NLP), high-quality national corpora, machine translation, and speech recognition.

The main areas and important problems of mathematical linguistics arise from the need to study language on a formal and mathematical basis. One of the most important tasks facing this field is to organize linguistic phenomena into a precise, consistent, and logical system, that is, to express them using mathematical apparatus. In this regard, mathematical linguistics seeks to address the following major issues: first, the creation of an axiomatic theory of language, that is, explaining the language system on the basis of initial rules and axioms; second, modeling the structure of language in the form of formal grammar; and third, developing mathematical models of natural languages and transforming them into a form suitable for computer processing.

CONCLUSION

In conclusion, the main objective of mathematical linguistics is to create a mathematical model of natural languages and provide its theoretical and practical justification. This objective involves not only describing language but also representing it as a precise algorithmic system. To achieve this objective, the field performs a number of important tasks. These include developing formal models of natural and artificial languages and representing linguistic structures as systems based on precise rules; evaluating linguistic phenomena using mathematical parameters and determining their quantitative characteristics; and conducting an in-depth analysis of linguistic units based on mathematical methods. Thus, unlike traditional descriptive approaches to language, mathematical linguistics serves to study language as a formal, precise, and computable system.

REFERENCES

1.....Hjelmslev, L. Prolegomena to a Theory of Language. Madison: University of Wisconsin Press, 1961. 240 p.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

2.....Chomsky, N. Syntactic Structures. The Hague: Mouton, 1957. 160 p.
3.....Muhammedova, S. Harakat Fe'llari Asosida Kompyuter Dasturlari Uchun Lingvistik Ta'min Yaratish. Methodological manual. Tashkent, 2006. 280 p.
4.....Muhamedova, S. Kompyuter Lingvistikasi (Methodological Manual). Tashkent: ToshDPU, 2007.
5.....Mengliyev, B. O'zbek Tili Korpusshunosligining Dolzarb Masalalari. Tashkent, 2022. 12 p.
6.....Kadirova, M. Korpus Tilshunosligining Paydo Bo'lishi va Rivojlanishi Tarixi. "Iqro" Journal, 2023. 87 p.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.