ETHIOPIAN INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 12, Issue 06, June-2025 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

CORPUS LINGUISTICS AND ITS ADVANTAGES

Muxudullayev Farrux Farxod o'g'li
Lecturer at GulDU
muxudullayevf@gmail.com

Abstract: This article provides information about corpora, corpus linguistics, its formation and development stages, as well as its current advantages and authorial corpora.

Keywords: Corpus, corpus linguistics, card index, computer era, electronic library, authorial corpus, Navoi corpus.

Introduction.

A corpus is a collection of texts or words used in linguistic research, artificial intelligence, translation systems, and other fields related to linguistics. It is a set of linguistic units forming a body of texts collected for a specific purpose, stored electronically in written or spoken form, and accessible via software-based search systems either online or offline. It differs significantly from an electronic library, which stores literary and journalistic works reflecting the socio-political, spiritual, and economic life of society. Since texts in electronic libraries are not processed from lexical, morphological, grammatical, or semantic perspectives, they pose several challenges for linguistic research. Electronic libraries aim to collect national and spiritual heritage, not to prepare materials for scientific research.

Corpus linguistics is a field of linguistics that studies and analyzes language based on electronic collections of texts (corpora). It is a component of computational linguistics, dealing with the creation of language corpora and their application through computer technologies. The subject of corpus linguistics is the linguistic corpus, also known as a 'linguistic corpus' or 'text corpus' in English. Corpus linguistics emerged in the 1960s with the creation of corpora and significantly evolved in the 1980s due to advances in computing. The term 'corpus linguistics' was first used in 1984. Though not an ancient field, it has quickly become a leading area in modern linguistics.

Milestones include:

1. The Brown Corpus (1961–1963) – Created by Henry Kucera and Nelson Francis in the USA. It included one million words across various genres.
2. The London-Lund Corpus (1970) – One of the first resources for studying spoken English.
3. The British National Corpus (BNC, 1990) – A large corpus of 100 million English words.

Current Developments.

Today, corpus linguistics is closely related to artificial intelligence, natural language processing (NLP), and big data. Large electronic corpora such as Google Books Corpus, Wikipedia Corpus, and Common Crawl are used for language learning. Multilingual corpora such as OPUS and WMT parallel corpora developed by the European Union help build translation models. Historically, corpus linguistics has evolved in conjunction with lexicography, statistical analysis, and computational linguistics.

In Uzbekistan,

the development of Uzbek language corpora is still in its early stages. Research efforts are currently focused on creating theoretical foundations for Uzbek corpus linguistics and building initial corpus samples. A national center or laboratory for Uzbek corpus linguistics, along with qualified specialists, is essential. Despite limitations, some progress has been made in creating authorial corpora. Authorial

eISSN: 2349-5715 pISSN: 2349-5707
Volume: 12, Issue 06, June-2025 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

corpora development has become one of the most advanced areas of modern corpus linguistics, even enabling the identification of authors of anonymous works.

Authorial corpora:

These corpora allow for a detailed and objective analysis of an author's language, making them superior to other information banks. They serve as a foundation, resource, and tool for various types of research. For instance, the Navoi Authorial Corpus enables researchers to quickly and accurately study linguistic features, observe changes over time, understand obsolete or emerging words, and analyze linguistic phenomena. It facilitates the creation of large-scale dictionaries and automated text processing.

The Alisher Navoi Authorial Corpus

contains thousands of lexemes from the poet's ghazals. Semantic tagging – assigning lexical meanings to ambiguous words – helps users understand contextual meanings, analyze lexical compatibility and combinability, and assess syntactic structures.

Conclusion.

Corpus linguistics is a rapidly developing field. It is useful not only for linguists and literary scholars but also for researchers and independent learners across various disciplines.

References:

1. V. Zakharov, B. Mengliyev, Sh. Hamrayeva, 'Corpus Linguistics', Textbook. Tashkent, 2023.
2. Sh. M. Hamrayeva, 'Glossary of Corpus Linguistics Terms'. Tashkent, 2018.
3. A. Polatov, 'Computational Linguistics'. Tashkent, 2011.
4. A. Norov, 'Fundamentals of Computational Linguistics'. Karshi, 2017.
5. Sh. Hamrayeva, 'Linguistic Basis for Creating Uzbek Authorial Corpus', Monograph. Germany, 2020.
6. http://v1.alishernavoicorpus.uz
7. Muxudullayev, F. (2023). 'Perspectives on Linguopoetics'. Journal of Universal Science Research, 1(5), 31–37.
8. Muhudullaev, F. (2023). 'Inversion as a Syntactic Opportunity'. Science and Innovation, 2(B1), 402–405.
9. Muxudullayev, F. (2023). 'Perspectives on Linguopoetics'. Journal of Universal Science Research, 1(5), 31–37.
10. Sayyora, Y., Shoxistaxon, X., & Farrux, M. (2024). 'Paradigmatic Relations of Diminutive-Endearing Forms'. Central Asian Journal of Multidisciplinary Research and Management Studies, 1(3), 70–73.
11. Farrux, M., Sayyora, Y., & Ma'rufjon, Q. (2024). 'The Role of Phonographic Tools in Erkin Vahidov's Works'. Central Asian Journal of Multidisciplinary Research and Management Studies, 1(3), 79–82.
12. Muxidillayev, F. (2022). 'Views on the Study of Linguopoetics'. Academic Research in Educational Sciences, 3(7), 315–320.