THE IMPORTANCE AND PRACTICAL IMPLICATIONS OF CORPUS LINGUISTICS IN MODERN LINGUISTICS

Yaxshilikova Gulsum
English language teacher at the middle school №28 in the Republic of Uzbekistan, Tashkent
DOI:https://doi.org/10.5281/zenodo.23194028

Abstract

This thesis explores one of the most active directions in modern linguistics—the fundamental principles, theoretical, and practical significance of corpus linguistics. The article investigates how text corpora differ from traditional electronic libraries. Furthermore, it analyzes the importance of text tagging (annotation) processes, as well as the role of this field in developing lexicography, translation theory, and artificial intelligence systems. Additionally, issues regarding the enrichment of the Uzbek national corpus and its role in increasing the prestige of our language on an international scale are highlighted.

Keywords: corpus linguistics, text corpus, linguistic annotation, lexicography, computer linguistics, Uzbek national corpus, natural language processing (NLP).

Introduction

The XXI century is an era of information technology and digitalization, and this process sets new tasks for linguistics. While traditional linguistics has relied for many years on limited examples and individual linguistic intuition, modern linguistics requires the analysis of vast databases of real texts. For this reason, corpus linguistics, which used to be considered an integral part of computational linguistics, has today turned into an independent and leading branch of science.

A corpus of texts is not simply an electronic library or a collection of texts. It is a collection of electronic texts sorted specifically for language research, organized into a specific system, and, most importantly, linguistically tagged (annotated). The importance of corpus linguistics lies in its provision of an opportunity to study language objectively based on written sources and spoken language during its actual functioning.

The importance of corpus linguistics can be viewed in several strategic directions:

1. Lexicography and the practice of dictionary making.

Compiling a dictionary using traditional methods used to take decades. With the help of corpora, however, the real frequency of words in a language (usage frequency), their typical word combinations (collocations), and their meanings in context are identified within seconds. This creates an opportunity to build modern, accurate, and vivid dictionaries.

2. Language Teaching and Methodology (Glottodidactics).

For students learning Uzbek or English as a foreign language, corpora serve as an invaluable resource. A corpus demonstrates to the user how a particular word behaves in speech across various grammatical forms and specific genres (literary, scientific, official), providing accurate statistical indicators.

3. Translation Studies and Parallel Corpora.

A parallel corpus (a database containing identical texts in two or more languages) fundamentally changes the practice of translation. Translators can easily resolve ambiguities in terminology with the

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

help of these databases; furthermore, machine translation systems like Google Translate are also trained using parallel corpora.

4. Artificial Intelligence and Neural Networks.

Today, ChatGPT and other large language models (LLMs), which have become widely popular, could not exist without language corpora. Corpora are the fundamental foundation for teaching computers to understand human language, analyze texts, and perform syntactic parsing.

The Relevance of the National Corpus of the Uzbek Language.

In an era where the Uzbek language holds the status of the state language, creating and developing the National Corpus of the Uzbek Language (uzcorpora.uz) is a task of state importance. Improving the morphological, syntactic, and semantic tagging of texts in the Uzbek language corpus strengthens the position of our language in the digital world. This serves to create programs that automatically edit Uzbek texts, spell checkers, and speech-to-text technologies.

Conclusion

In conclusion, corpus linguistics has elevated linguistics from a theoretical science to a high-tech and applied science. It is the most effective way to study language systematically, objectively, and on the basis of statistical data. In the future, developing Uzbek corpus linguistics, training specialists in the field, and expanding the volume of corpora should remain the most important priority of our national linguistics.

References

1. Dadaboyev H. Fundamentals of Computational Linguistics. – Tashkent, 2021.
2. Lando J. Corpus Linguistics as a Direction of Text Analysis. // Journal of Linguistics Issues, 2023.
3. Uzcorpora.uz – Data from the official platform of the National Corpus of the Uzbek Language.
4. 1.Abjalova M. “Korpus lingvistikasi”.- Toshkent: Bookmany print, 2022.
5. Zaxarov V., Mengliyev B., Xamroyeva Sh.- Korpus lingvistikasi: korpus tuzish va undan foydalanish. O‘quv qo‘llanma.- Toshkent, 2021.
6. Svartik J. Corpus linguistics comes of age. Directions in Corpus linguistics.- New York, 1992.
7. http://ice-corpora.net
8. http://uzschoolcorpara.uz

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.