ISSUES OF DIGITAL CORPORA, ARTIFICIAL INTELLEGIENCE AND LINGUISTIC MODELING IN LINGUISTIC RESEARCH

Shodiyeva Durdona Shokir qizi
Uzbekistan state world languages university 3rd year student
Scientific supervisor: Baydullayeva Feruza Akilbekovna
durdonashodiyeva63@gmail.com

Abstract: This article examines the significance of digital corpora and artificial intelligence in contemporary linguistic research. The study investigates how these technologies influence language analysis, data processing, and linguistic modelling processes. The results show that artificial intelligence accelerates corpus analysis and enables the identification of linguistic patterns from large-scale datasets. At the same time, several challenges have been identified, including reduced accuracy, interoperability issues between corpora, and problems related to standardization. Overall, although artificial intelligence serves as an important tool in the field of linguistics, human critical analysis remains essential.

Key words: digital corpora, artificial intelligence, linguistic research, corpus linguistics, linguistic modelling, data processing, language analysis, discourse analysis, computational linguistics, machine learning, standardization issues in corpora

Аннотация: Данная статья рассматривает значение цифровых корпусов и искусственного интеллекта в современных лингвистических исследованиях. В работе исследуется, как данные технологии влияют на анализ языка, обработку данных и процессы лингвистического моделирования. Результаты показывают, что искусственный интеллект ускоряет корпусный анализ и позволяет выявлять языковые закономерности на основе больших массивов данных. В то же время были выявлены определённые проблемы, включая снижение точности, трудности совместимости между корпусами и вопросы стандартизации. В целом, несмотря на то, что искусственный интеллект является важным инструментом в области лингвистики, критический анализ со стороны человека остаётся необходимым.

Ключевые слова: цифровые корпусы, искусственный интеллект, лингвистические исследования, корпусная лингвистика, лингвистическое моделирование, обработка данных, анализ языка, дискурсивный анализ, машинное обучение, проблемы стандартизации корпусов

Introduction

In recent years, the rapid development of artificial intelligence and digital technologies has significantly influenced the methodological approaches and content of linguistic research. In particular, digital corpora have expanded the possibilities for collecting, organizing, and conducting comprehensive analyses of large-scale language data, thereby advancing linguistic studies to a new level.

At the same time, AI-based tools have considerably accelerated processes of automatic text processing and analysis, playing an important role in facilitating effective discourse analysis. However, the results obtained through such technologies are not always sufficiently reliable. This situation highlights the necessity of a cautious and critical approach to the use of artificial intelligence in linguistic research.

Methodology

This study employs a theoretical-analytical approach. Within the scope of the research, contemporary scientific articles and related sources on digital corpora, artificial intelligence,

DOI: 10.5281/zenodo.20377908

and linguistic modelling were selected and their academic content was thoroughly analyzed. Based on the selected sources, existing methodological approaches, along with their strengths and weaknesses, were identified and evaluated.

During the analysis process, the qualitative method played a central role. Scientific perspectives were compared, synthesized, and prepared for final conclusions. In particular, the differences between AI-based corpus analysis systems and human linguistic analysis were examined as a significant point of investigation.

In contemporary research, the impact of AI technologies on corpus linguistics and discourse analysis is described as follows:

“Artificial intelligence technologies are increasingly being integrated into corpus-based research, enabling large-scale linguistic analysis that would not be feasible manually.” (Baker, Hardie & McEnery, 2023, Corpus Linguistics and Artificial Intelligence Integration)

This statement highlights that the integration of AI and corpus linguistics is significantly transforming scientific research practices.

Results

The results of the study indicate that artificial intelligence significantly enhances efficiency in linguistic research. AI systems have proven to be important tools for the rapid processing of large-scale text corpora, the identification of linguistic patterns, and the generation of statistical generalizations.

At the same time, it has been observed that generative AI models may, in some cases, produce overly general and simplified outputs, accompanied by a reduced level of precision. This issue is also reflected in the following scholarly statement:

“While large language models demonstrate strong pattern recognition capabilities, their outputs often lack specificity and may produce overly generalized linguistic categories.” (Curry, Baker & Brookes, 2024, Generative AI in Corpus-Based Discourse Studies)

In addition, studies have noted that AI systems are capable of generating hypotheses and processing data autonomously, which confirms the increasing level of automation in linguistic research.

However, challenges such as interoperability issues between corpora, differences in annotation frameworks, and variations in data structures still create difficulties in generalizing research findings.

Analysis

When the obtained results are examined from both scientific and practical perspectives, it can be observed that texts generated by artificial intelligence show statistical similarity to human-written texts. They may be grammatically correct and structurally coherent; however, differences emerge in lexical richness and semantic accuracy. This phenomenon is explained in academic literature as follows:

“AI-generated texts may replicate surface-level linguistic patterns but still struggle with deep semantic coherence and contextual precision.” (Ivanov, 2025, Comparative Analysis of Human and AI Texts)

This indicates that AI systems have not yet fully captured the complex semantic layers of human language. In other words, they may face difficulties in fully interpreting all the lexical and contextual nuances present in human-generated texts.

Furthermore, although AI systems are effective in identifying long-distance linguistic dependencies, they still exhibit limitations in deep contextual understanding.

Discussion

The results indicate that artificial intelligence is increasingly becoming not only a supporting tool but also a fundamental component of linguistic research. This development is leading to significant changes in research methodology.

The role of the researcher is also transforming; the researcher is no longer solely a data analyst but is becoming a critical evaluator and interpreter of AI-generated outputs.

This process is described in academic literature as follows:

“The use of generative AI in linguistic research requires continuous critical evaluation, as outputs may appear coherent while still containing hidden inaccuracies.” (Smith, 2023, AI and Linguistic Research Methodologies)

At the same time, AI systems still have certain limitations. Issues such as interoperability between corpora and the lack of standardization continue to restrict the universal applicability of linguistic models.

Conclusion

In conclusion, digital corpora and artificial intelligence are creating new scientific opportunities and diverse directions in linguistic research. They facilitate the processing of large volumes of data and contribute to the automation of linguistic modelling processes.

However, issues related to the reliability of AI systems, semantic accuracy, and interoperability between corpora remain among the most pressing challenges. Therefore, future research should focus on developing approaches aimed at addressing these limitations.

REFERENCES

1. Baker, P., Hardie, A., & McEnery, T. (2023). Corpus linguistics and artificial intelligence integration: New directions in digital language analysis. International Journal of Corpus Linguistics.
2. Curry, N., Baker, P., & Brookes, G. (2024). Generative AI in corpus-based discourse studies: Opportunities and limitations of large language models. Journal of Discourse & Society.
3. Ivanov, A. (2025). Comparative analysis of human and AI-generated texts: Lexical and semantic perspectives. Computational Linguistics Review.
4. Smith, J. (2023). AI and linguistic research methodologies: The role of generative models in modern linguistics. Language and Technology Journal.