THE ROLE OF ARTIFICIAL INTELLIGENCE AND DIGITAL CORPORA IN MODERN LINGUISTIC RESEARCH: OPPORTUNITIES AND CHALLENGES

Saburova Saodat
Student, 1st English faculty, Uzbekistan State World Language University
Scientific advisor: **Yunus Masharipov**
Teacher, 1st English faculty, Uzbekistan State World Language University.

Abstract: This article analyzes the role of artificial intelligence, digital corpora, and linguistic modeling in modern linguistic research. It examines the development of linguistic approaches, starting from traditional theories, particularly those proposed by Chomsky, to contemporary methods based on artificial intelligence. In particular, the effectiveness of technologies such as Natural Language Processing (NLP), machine learning, and deep learning in language analysis is discussed. The article also describes linguistic modeling processes using modern models such as Word2Vec, Transformer, and BERT, and highlights their advantages in processing large amounts of textual data. In addition, special attention is given to the importance of digital corpora in developing empirical research in linguistics.

However, the limitations of artificial intelligence technologies are also considered from a critical perspective. Issues such as the problem of “understanding” language, dependence on data, model bias, and ethical concerns are discussed. These aspects show that, along with its significant potential, artificial intelligence also has certain limitations in linguistic research.

The results of the study indicate that artificial intelligence and digital corpora are taking linguistics to a new level. However, their effective use requires a critical approach and scientifically grounded integration. In this article, the author analyzes existing approaches, compares their advantages and disadvantages, and presents several suggestions for more effective use of artificial intelligence in linguistic research.

Keywords: Artificial Intelligence, Natural Language Processing, Corpus Linguistics, Linguistic Modeling, Machine Learning, Deep Learning, Word Embedding, Transformer Model, BERT.

Annotatsiya: Ushbu maqolada zamonaviy tilshunoslik tadqiqotlarida sun’iy intellekt, raqamli korpuslar va lingvistik modellashtirishning o‘rni kompleks tarzda tahlil qilinadi. An’anaviy lingvistik yondashuvlar, xususan, Chomsky tomonidan ilgari surilgan nazariy modellardan boshlab, bugungi kunda qo‘llanilayotgan sun’iy intellektga asoslangan metodlargacha bo‘lgan rivojlanish jarayoni ko‘rib chiqiladi. Xususan, tabiiy tilni qayta ishlash (NLP), mashinaviy o‘rganish va chuqur o‘rganish texnologiyalarining tilni tahlil qilishdagi samaradorligi tahlil etiladi. Maqolada Word2Vec, Transformer va BERT kabi zamonaviy modellar misolida lingvistik modellashtirish jarayonlari yoritilib, ularning katta hajmdagi matnlar bilan ishlashdagi ustunliklari ko‘rsatib beriladi. Shu bilan birga, raqamli korpuslarning tilshunoslikdagi empirik tadqiqotlarni rivojlantirishdagi ahamiyati alohida ta’kidlanadi.

Biroq, sun’iy intellekt texnologiyalarining cheklovlari ham tanqidiy nuqtai nazardan ko‘rib chiqiladi. Xususan, tilni “tushunish” masalasi, ma’lumotlarga bog‘liqlik, model tarafkashligi (bias) va etik muammolar mavjudligi asoslab beriladi. Ushbu jihatlar sun’iy intellektning lingvistik tadqiqotlardagi imkoniyatlari bilan bir qatorda uning cheklangan tomonlarini ham ko‘rsatadi.

Tadqiqot natijalari shuni ko‘rsatadiki, sun’iy intellekt va raqamli korpuslar tilshunoslikni yangi bosqichga olib chiqmoqda, biroq ularni samarali qo‘llash uchun tanqidiy

DOI: 10.5281/zenodo.20377418

yondashuv va ilmiy asoslangan integratsiya zarur. Mazkur maqolada muallif tomonidan mavjud yondashuvlar tahlil qilinib, ularning afzallik va kamchiliklari solishtiriladi hamda sun’iy intellektdan lingvistik tadqiqotlarda yanada samarali foydalanish bo‘yicha ayrim taklif va mulohazalar ilgari suriladi.

Kalit so‘zlar: Sun’iy intellekt, Tabiiy tilni qayta ishlash, Korpus lingvistikasi, Lingvistik modellashtirish, Mashinaviy o‘rganish, Chuqur o‘rganish, Word Embedding, Transformer modeli, BERT.

1.Introduction: In the digital era, linguistics is shifting from traditional theoretical approaches to a data-driven scientific paradigm. In recent years, these changes have accelerated due to the rapid development of artificial intelligence and digital technologies. While traditional linguistic research has mainly relied on theoretical models [1], modern studies are increasingly based on large-scale linguistic data and digital corpora. As a result, new methodologies and approaches for language analysis are emerging.

Artificial intelligence technologies, particularly Natural Language Processing (NLP), machine learning, and deep learning methods, play an important role in linguistics [2]. These approaches make it possible to automatically analyze language units, identify semantic and syntactic relationships, and efficiently process large volumes of text. Modern models such as Word2Vec, Transformer, and BERT have significantly improved linguistic modeling [3][4][5] and contributed to a deeper understanding of language phenomena. In addition, digital corpora strengthen the empirical basis of linguistic research by enabling the analysis of language phenomena using real data [7]. However, these technologies also present certain challenges. In particular, issues such as the ability of AI systems to truly “understand” language, dependence on data, model bias, and ethical concerns remain important problems [6].

The main aim of this study is to analyze the role of artificial intelligence, digital corpora, and linguistic modeling in modern linguistics, to identify their advantages and limitations, and to provide scientifically grounded conclusions on their effective application.

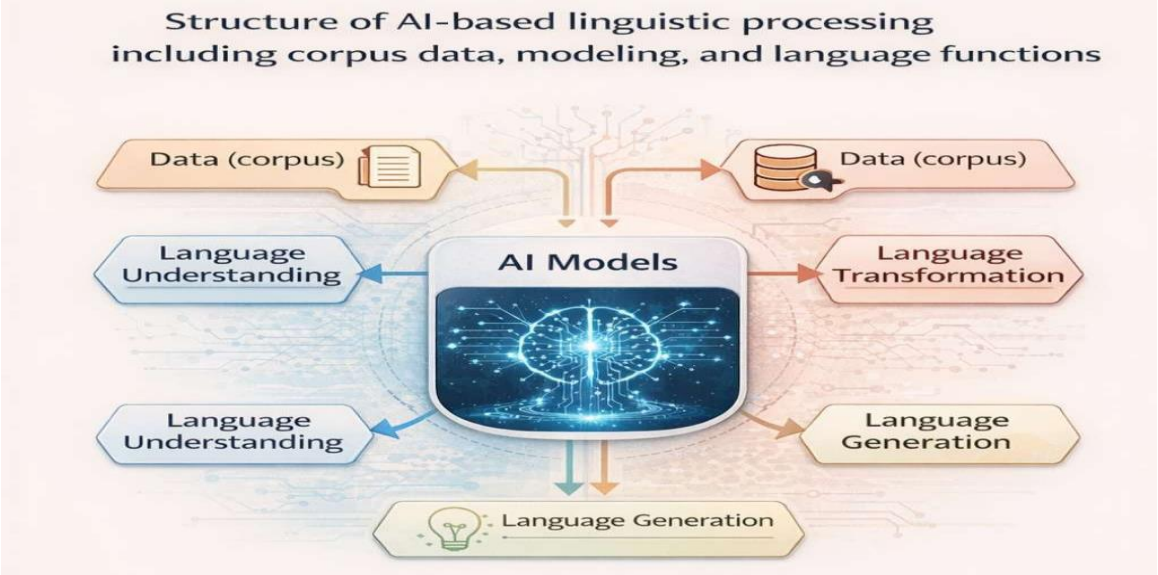
This study seeks to answer the following question: how effectively can artificial intelligence improve linguistic research while addressing its limitations?

2. Literature Review: In recent years, the integration of artificial intelligence and linguistics has become one of the important directions in scientific research. In particular, AI models based on digital corpora have brought the processes of language understanding, processing, and generation to a new level. To explain these processes, modern research is often based on the principle of data – model – output (language functions). First of all, the theoretical foundations of linguistic research go back to the theory of generative grammar developed by Noam Chomsky. This approach aims to explain language as a system of the human mind and later became the basis for the development of computational linguistics. However, in modern research, language study is not only theoretical but also carried out based on large-scale data.

Digital corpora play an important role as the main source of linguistic analysis. According to Tony McEnery and Andrew Hardie, corpus linguistics allows researchers to systematically study real language samples and draw conclusions based on empirical results. Therefore, modern AI systems rely on large corpora as their primary source of data.The next main stage after data is AI models. The Word2Vec model proposed by Tomas Mikolov made it possible to represent semantic relationships between words in vector form. This approach

opened the way to understanding language as a mathematical model. Later, the development of deep learning methods expanded the ability to identify complex linguistic structures.

The Transformer architecture created a major breakthrough in this field. This model, developed by Ashish Vaswani and his co-authors, allows deep contextual analysis of text. Based on this, the BERT model developed by Jacob Devlin achieved high results in NLP tasks by understanding language in a bidirectional context.



In addition, these models perform three main linguistic functions: language understanding, language transformation, and language generation. As Daniel Jurafsky and James H. Martin point out, Natural Language Processing technologies help automate these processes. For example, language understanding means identifying the meaning of a text, transformation refers to translation or paraphrasing, and generation involves creating new text. These processes can also be seen in everyday use. For instance, machine translation systems are based on large corpus data. First, the system learns from this data, and then AI models are used to translate text from one language into another. In this way, the transformation function of language is clearly demonstrated in practice.

At the same time, the performance of AI models largely depends on data. If the corpus is not sufficiently large or diverse, the results may not be accurate. As Emily Bender and her colleagues note, such systems do not truly understand language but rely on existing data patterns. In addition, model bias and ethical issues remain important challenges.

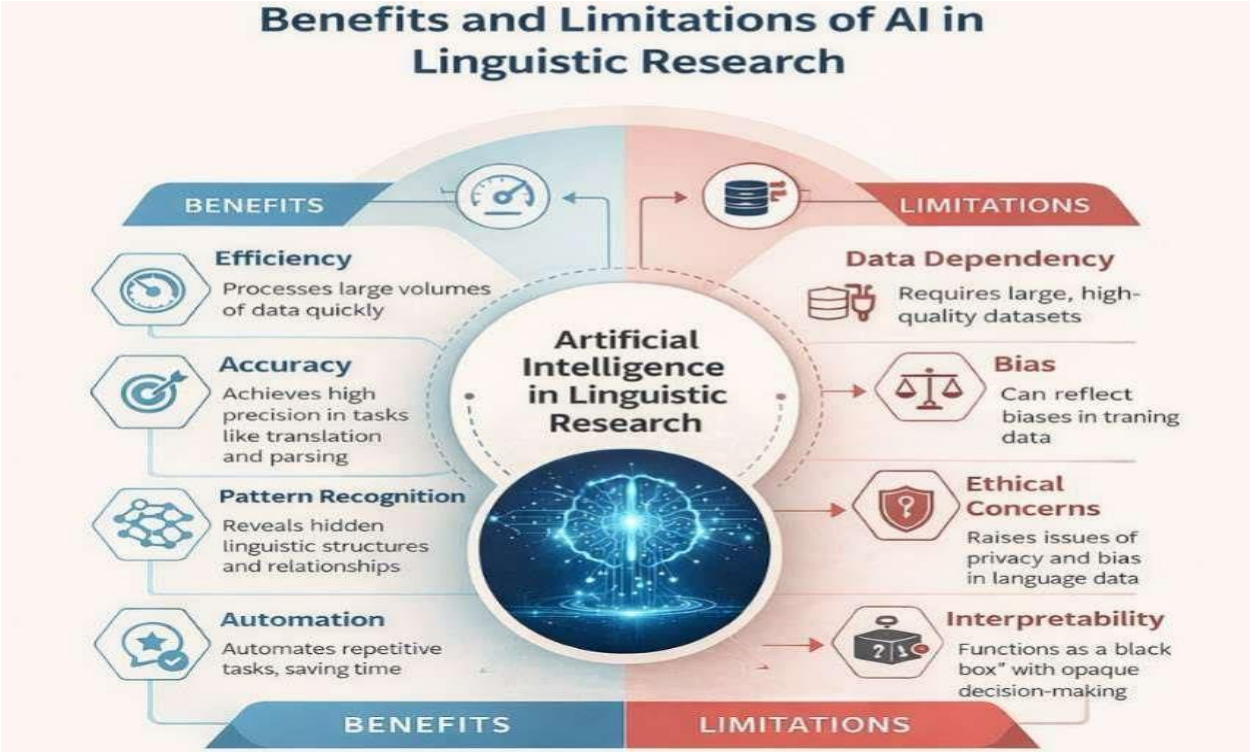
In general, modern linguistic research is based on the relationship between data, AI models, and language functions. This approach allows a deeper understanding of language, but its effective use requires a critical approach.

3.Benefits and Limitations of AI in Linguistic Research: Benefits of AI in Linguistics - One of the main advantages of artificial intelligence in linguistic research is its efficiency in processing large volumes of data. According to Daniel Jurafsky and James H. Martin (2019), Natural Language Processing systems allow automatic analysis of linguistic data, which significantly reduces the time required for manual work. For example, tasks such as part-of-speech tagging and syntactic parsing can now be performed with high speed and accuracy. Another important benefit is the ability to achieve high accuracy and precision. Research by Tomas Mikolov et al. (2013) shows that word embedding models can capture semantic relationships between words based on context. This allows AI systems to better

understand meaning and improve performance in tasks such as translation and text classification.

In addition, AI helps uncover hidden linguistic patterns. As noted by Yoav Goldberg (2017), deep learning models are capable of identifying complex structures in language that are difficult to detect using traditional methods. This makes linguistic analysis more detailed and data-driven.

In conclusion, automation is another key advantage. AI-powered tools can perform repetitive tasks such as translation, summarization, and speech recognition. This increases productivity and allows researchers to focus on more theoretical aspects of linguistics.



Limitations of AI in Linguistics: Despite these advantages, artificial intelligence also has several important limitations. One of the main issues is the dependence on data. As highlighted by Emily M. Bender et al. (2021), AI models rely heavily on large datasets and do not truly understand language. Instead, they learn statistical patterns, which can lead to incorrect interpretations. Another serious limitation is model bias. According to Timnit Gebru (2021), biases present in training data can be reflected in AI outputs. This can result in unfair or misleading conclusions, especially in sensitive linguistic contexts.

Ethical concerns also play a significant role in AI-based linguistic research. Luciano Floridi (2019) emphasizes that issues such as data privacy, consent, and responsible use of technology must be carefully considered when working with language data. In addition, interpretability remains a challenge. As noted by Zachary Lipton (2018), many AI models, especially deep learning systems, function as “black boxes,” making it difficult to understand how decisions are made. This limits transparency and trust in AI-based linguistic analysis.

3.Methodology: This study is based on the analysis of existing academic literature related to artificial intelligence and linguistics. Various sources, including books, journal articles, and conference papers, were selected to examine the role of AI technologies and digital corpora in linguistic research.

The research follows a qualitative and descriptive approach. Different models and methods were analyzed and compared in order to identify their advantages and limitations. For example, the effectiveness of AI models such as Word2Vec and BERT was evaluated based on the findings of previous researchers, including Mikolov et al. (2013) and Devlin et al. (2019), who showed that these models improve semantic understanding and contextual analysis. In addition, some observations were made based on practical examples. For instance, machine translation systems demonstrate how AI models can transform text from one language into another using large corpora. Based on personal observation, such systems work efficiently in general contexts, but sometimes fail to capture deeper meanings or cultural nuances. This shows a difference between theoretical results and real-world application.

The study also compares traditional linguistic approaches with modern AI-based methods. While traditional models, such as those proposed by Chomsky, focus on rules and structure, AI approaches rely more on data and statistical patterns. This comparison helps to better understand both the strengths and limitations of each approach.

Overall, the methodology combines theoretical analysis with practical observation, allowing for a more balanced and realistic evaluation of artificial intelligence in linguistic research.

4. Conclusion: This study has shown that artificial intelligence and digital corpora play an increasingly important role in modern linguistic research. The findings confirm that AI technologies, including Natural Language Processing, machine learning, and deep learning, significantly improve the efficiency, accuracy, and scope of language analysis. Models such as Word2Vec, Transformer, and BERT demonstrate strong capabilities in capturing semantic and contextual relationships in language.

At the same time, the study also highlights several important limitations. As noted by researchers such as Bender et al. (2021), AI systems do not truly understand language but rely on statistical patterns derived from data. Issues such as data dependency, model bias, and ethical concerns remain key challenges that cannot be ignored. Based on both previous research and personal observation, it can be concluded that AI technologies are highly effective in practical tasks such as machine translation and text analysis. However, they may still struggle with deeper meaning, context, and cultural nuances. This shows that human linguistic knowledge still plays an important role.

Therefore, rather than replacing traditional linguistic approaches, artificial intelligence should be used as a complementary tool. A balanced integration of theoretical knowledge and data-driven methods can lead to more reliable and meaningful results in linguistic research.

Future studies should focus on improving model transparency, reducing bias, and developing more advanced systems that better reflect the complexity of human language.

References:

1. Chomsky, N. (1957). Syntactic Structures. Mouton.
2. Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).
4. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. https://arxiv.org/abs/1706.03762

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019).
7. BERT: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? https://dl.acm.org/doi/10.1145/3442188.3445922
9. Goldberg, Y. (2017). Neural network methods for natural language processing. https://www.jair.org/index.php/jair/article/view/11198
10. Floridi, L. (2019). Establishing the rules for building trustworthy AI. https://link.springer.com/book/10.1007/978-3-030-04447-8
11. Lipton, Z. C. (2018). The mythos of model interpretability. https://queue.acm.org/detail.cfm?id=3241340