AI-DRIVEN APPROACHES TO CORPUS-BASED LINGUISTIC MODELING

Shomirzayeva Shakhnoza
Maxmudova Muxlisa
cvetocekangelina4@gmail.com
shomirzayevashahnoza06@gmail.com
Scientific supervisor: **Masharipov Yunus**
Teacher Faculty of English language-1
Uzbekistan State World Language University

Abstract. The most recent achievements in the development of AI technologies have radically changed the methods and techniques used for corpus-based language modeling that involve the application of large datasets and high computing capacity. By integrating artificial intelligence technologies into corpus linguistics, one can address the issues associated with the vast amount of text material and its inherent complexity in terms of finding underlying connections between different aspects to create advanced models of language behavior. Recent academic studies emphasize the role of machine learning algorithms and neural network technologies in enhancing language modeling and analyzing various lexical, grammatical, and semantic features of digital corpora.

Key words: artificial intelligence, corpus linguistics, language modeling, machine learning, digital corpora, computational linguistics, neural networks, natural language processing, text mining, big data analysis, semantic analysis, syntactic parsing, linguistic variation, data bias, algorithmic transparency, automated text processing, probabilistic models, transformer models, discourse analysis, sociolinguistics.

Аннотация. Развитие технологий искусственного интеллекта существенно изменяет подходы к корпусному лингвистическому моделированию. Использование методов ИИ позволяет анализировать большие массивы текстовых данных, выявлять скрытые языковые закономерности и создавать более точные языковые модели. Современные исследования подчеркивают роль машинного обучения и нейронных сетей в изучении лексических, синтаксических и семантических структур цифровых корпусов. Такие технологии обеспечивают автоматическую аннотацию, прогнозное моделирование и контекстно-зависимое представление языка. Вместе с тем применение ИИ связано с рядом проблем, включая предвзятость данных, прозрачность алгоритмов и воспроизводимость результатов. Несмотря на это, ИИ-ориентированные подходы открывают новые перспективы для развития теоретической и прикладной лингвистики. В статье рассматриваются современные тенденции, методологические подходы и практическое значение интеграции искусственного интеллекта в корпусные исследования.

Ключевые слова: искусственный интеллект, корпусная лингвистика, языковое моделирование, машинное обучение, цифровые корпуса, вычислительная лингвистика, нейронные сети, обработка естественного языка, анализ текста, большие данные, семантический анализ, синтаксический разбор, языковая вариативность, смещение данных, прозрачность алгоритмов, автоматическая обработка текста, вероятностные модели, трансформерные модели, дискурс-анализ, социолингвистика.

Annotatsiya. Sun’iy intellekt texnologiyalarining jadal rivojlanishi korpusga asoslangan lingvistik modellashtirish jarayonini tubdan o‘zgartirmoqda. AI usullari yordamida katta hajmdagi matn korpuslarini tahlil qilish, yashirin lingvistik qonuniyatlarni

DOI: 10.5281/zenodo.20378001

aniqlash va yanada aniq til modellari yaratish imkoniyati kengaymoqda. Zamonaviy tadqiqotlar mashinaviy o‘rganish va neyron tarmoqlar orqali leksik, sintaktik hamda semantik strukturalarni chuqurroq o‘rganish mumkinligini ko‘rsatadi. Bunday yondashuvlar avtomatik annotatsiya, bashoratli modellashtirish va kontekstga mos til tavsifini shakllantirishga xizmat qiladi. Shu bilan birga, AI texnologiyalaridan foydalanish ma’lumotlar tarafkashligi, algoritmik shaffoflik va natijalarni qayta tiklash kabi muammolarni ham yuzaga keltiradi. Shunga qaramay, sun’iy intellekt asosidagi korpus modellashtirish lingvistikaning nazariy va amaliy rivojiga katta hissa qo‘shadi. Ushbu maqolada mazkur yo‘nalishdagi zamonaviy tendensiyalar, metodologik yondashuvlar va amaliy ahamiyat tahlil qilinadi.

Kalit so‘zlar: sun’iy intellekt, korpus lingvistikasi, til modellashtirish, mashinaviy o‘rganish, raqamli korpuslar, hisoblash lingvistikasi, neyron tarmoqlar, tabiiy tilni qayta ishlash, matn tahlili, katta ma’lumotlar (big data), semantik tahlil, sintaktik tahlil, lingvistik variativlik, ma’lumotlar tarafkashligi (bias), algoritmik shaffoflik, avtomatlashtirilgan matn qayta ishlash, ehtimollik modellari, transformer modellari, diskurs tahlili, sotsiolingvistika.

Introduction

With the speedy evolution of the field of artificial intelligence technologies, there is a new paradigm which is now being developed in the realm of linguistics, and even more specifically, in corpus modeling. Being grounded on the idea of manual or partially automated text analysis, the classic approach in corpus linguistics could not address some issues in terms of scale and scope in this field. However, the use of advanced technologies has led to the emergence of ways to analyze big amounts of textual material and discover correlations that could not be observed prior (Loukanova, 2026). The results of the current research indicate that the use of machine learning tools and neural language models may assist in finding probabilities of lexemes and constructions within the text (Wei, 2025).

In addition, all of this contributes to the dynamism of language as a system that changes and develops depending on how the language works instead of according to strict regulations (Grieve et al., 2025). Artificial intelligence can contribute to keeping the linguistic model up-to-date with the introduction of new data (Pérez-Paredes et al., 2025). Another important factor in using artificial intelligence in corpus linguistics concerns the automation of the processes involved. Algorithms are capable of carrying out a wide range of operations in an incredibly accurate manner, namely tagging, parsing, and semantic classification. Consequently, people no longer need to be involved in this process (Angelova et al., 2025). This not only makes it possible to analyze the data without bias but also increases the speed at which it is carried out. According to specialists, results should be analyzed using the linguistic theory as well. On the other hand, the use of artificial intelligence in corpus-based language modeling has its own peculiarities. The problem of the potential data bias, for example, is significant enough not to be overlooked. If the training corpus lacks diverse features of language structures, then the model will display the biasness, instead of analyzing the data. To avoid this problem and any other related to data bias, all data needs to be evenly distributed throughout the corpus.

Another problem arising from the inclusion of AI in language modeling is the increased level of complexity of such systems. Some modern machine learning approaches make it really hard to figure out how exactly the algorithm works. That is why the machine learning algorithms and neural networks are called "black boxes".

Thus, it turns out that quite often it becomes difficult to define whether the output provided by a particular model is precise or wrong. However, it is high time we started paying our attention to constructing more clear and understandable language models based on the

use of artificial intelligence. These issues highlight the significance of appropriate approaches to addressing a few issues connected with the construction of a corpus and evaluation of a model.

Still, working with the mentioned technologies may result in a number of issues. In the first place, one should pay particular attention to the quality and ethical aspects of data used for constructing linguistic models, since these aspects play an essential role (Shormani, 2024). This means that it is highly important to discuss the opportunities offered by AI within the scope of language research by discussing the methodologies used by various scholars.

In the present work, the primary emphasis is made on the application of modern machine learning algorithms, neural models, etc. for linguistic corpus processing. Indeed, it is especially important at the moment when scholars do not only analyze a considerable amount of data through the computer but also find novel linguistic phenomena. The fact is that artificial intelligence models have greatly helped in modeling the linguistic data obtained through corpora. Machine learning algorithms help to find complex patterns while neural models provide contextually-sensitive models. Among other important aspects that should be mentioned nowadays, it should be stated that the role of big corpora of linguistic data has become more and more prominent. With the help of big corpora, language models can be very diverse because various versions can be analyzed depending on the representatives of particular social groups. Artificial intelligence poses additional problems, as well. One of them includes the issue of bias. In case a model based on the biased data is created, false results will be inevitable. Another issue associated with the application of artificial intelligence is its complexity. Finally, artificial intelligence affects the roles of linguists.

Methods. Methodologically, the current research employs a mixed-methodology approach, involving corpus analysis and computational modeling approaches. In particular, the objective here will be to explore the possibility of employing artificial intelligence software in order to enhance the examination of linguistic features through corpus analysis. The corpus used in this study consists of balanced texts taken from various sources and registers, including academics, journalism, and online communication. Particular attention has been paid to constructing a representative corpus of texts, as recommended in certain corpus studies recently conducted (Pérez-Paredes). As regards data analysis, both machine learning approaches and neural language models capable of finding patterns and modeling linguistic features via probabilistic and contextual approaches have been employed respectively. Certain such approaches have been reviewed in relevant literature sources (cf. Angelova et al., 2025). Furthermore, statistics were utilized to uncover recurrent trends and relationships within the corpus. This methodological approach also integrates criteria for critical assessment to address any bias present in the data set, as highlighted by previous studies conducted in the field of artificial intelligence and linguistics (Shormani, 2024).

Results. From the findings, it can be seen that using artificial intelligence for analyzing the corpus has increased effectiveness and accuracy greatly. It was possible to identify lexical and syntactic features, which could not have been detected manually. In addition, neural models have shown to be effective in capturing the dependencies of linguistic elements in the process and understanding of probability relations, which is consistent with the statements made by Wei (2025). In terms of dynamics of language usage within various social and communicative contexts, the findings confirmed sociolinguistic claims about the dynamic nature of language (see, for example, Grieve et al., 2025). The analysis allowed detecting changes in usage patterns of different types of language, especially those used in modern electronic communications. It should be noted that there were certain biases in the outputs

generated by AI models when the input data was not diverse enough. As a result, there was an emphasis on the dominant linguistic patterns, which was discussed in previous research on AI bias in linguistic studies.

Discussion. The results of the current paper reveal the profound transformation potential of the use of AI in corpus linguistics. In particular, one of the key features of AI models is their ability to analyze enormous amounts of texts and detect sophisticated patterns in them. As Loukanova (2026) argues, such an approach significantly widens the scope of linguistic researches. At the same time, it becomes evident that the application of artificial intelligence requires critical thinking skills. In particular, such issues as biases in data and low interpretability of models continue to exist in practice. The fact is that many neural language models work as black boxes which makes it difficult to assess their accuracy. It can be associated with the issue of algorithmic understanding of the problem described by Ruckenstein (2023). Finally, it is important to state that artificial intelligence should enhance linguistic theories but not become an alternative to them. In other words, human knowledge remains crucial to interpret results correctly. As such, even while there are various benefits in applying artificial intelligence to language corpus modeling, care must be taken to use proper methodology and approach the issue in an ethical way. In this light, future efforts can be directed at increasing transparency and minimizing bias.

Conclusion. AI technologies may help to improve opportunities for developing corpus-based linguistics and other sciences where processing and analyzing data become extremely important for achieving success. Many possibilities appear due to using AI methods for language studies and effective problem-solving strategies can be applied. However, applying AI methods in linguistic studies should be conducted taking into account all questions raised concerning methodology and ethics because some inaccuracies or mistakes can happen. All data selection criteria, modeling methods, and evaluation procedures call for detailed analysis from a scholar's standpoint since even small methodological flaws may result in low-quality results of AI-driven linguistic research. In addition, it is important to discuss ways to manage data and make decisions based on algorithms, avoid any misinterpretations, and perform analysis with integrity. Last but not least, it is necessary to understand that AI cannot become a replacement for linguists. Artificial intelligence is just an effective research tool that helps scientists analyze the material better and draw valid conclusions. At the same time, AI cannot substitute human reasoning.

References:

1. Adeshola, I., & Adepoju, A. P. (2024). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, *32*(10), 6159–6172.
2. Angelova, G., Kunilovskaya, M., Escribe, M., & Mitkov, R. (Eds.). (2025). *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing: Natural language processing in the generative AI era*. INCOMA Ltd.
3. Exploring the future of corpus linguistics: Innovations in AI and social impact. (2025). *International Journal of Mass Communication*, *3*, 1–10.
4. Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., & Winter, B. (2025). The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, *7*, 1472411.
5. Ismayilova, F. (2025). Artificial intelligence as a new lens on linguistics. *Porta Universorum*, *1*(10), 23–27.

6. Loukanova, R. (Ed.). (2026). *Computational linguistics, information, reasoning, and AI 2024*. Springer.
7. Pérez-Paredes, P. (2026). Corpus linguistics and AI in the reconfiguration of language learning ecologies. *Language Teaching*. Cambridge University Press.
8. Pérez-Paredes, P., Curry, N., & Aguado Jiménez, P. (2025). Integrating critical corpus and AI literacies in applied linguistics: A mixed-methods study. *Computer Assisted Language Learning*.
9. Ruckenstein, M. (2023). *The feel of algorithms*. University of California Press.
10. Shormani, M. Q. (2024). What fifty-one years of linguistics and artificial intelligence research tell us about their correlation: A scientometric review. *arXiv*.
11. Wei, L. (Ed.). (2025). *Computational Linguistics, Volume 51*(1). MIT Press.