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

Sobirova Nasiba
Uzbekistan State World Languages University
Supervisor: Rakhmanova Viktoriya, senior teacher
Uzbekistan State World Languages University

Annotation. The rapid development of digital technologies has significantly transformed modern linguistic research. This article examines the role of digital corpora, artificial intelligence, and linguistic modeling in the study of language. Digital corpora provide authentic language data for large-scale analysis, while artificial intelligence technologies contribute to natural language processing, machine translation, and speech recognition. Linguistic modeling helps researchers represent and predict language structures and patterns through computational methods. The article also discusses the major challenges related to data reliability, algorithmic bias, multilingual limitations, and ethical concerns. The study concludes that although these technologies greatly enhance linguistic research, careful methodological and ethical considerations remain essential for obtaining accurate and responsible research outcomes.

Keywords: digital corpora, artificial intelligence, linguistic modeling, corpus linguistics, natural language processing, computational linguistics

Аннотация. Стремительное развитие цифровых технологий значительно изменило современные лингвистические исследования. В данной статье рассматривается роль цифровых корпусов, искусственного интеллекта и лингвистического моделирования в изучении языка. Цифровые корпусы предоставляют аутентичные языковые данные для масштабного анализа, а технологии искусственного интеллекта способствуют развитию обработки естественного языка, машинного перевода и распознавания речи. Лингвистическое моделирование помогает исследователям представлять и прогнозировать языковые структуры и закономерности с помощью вычислительных методов. В статье также обсуждаются основные проблемы, связанные с надежностью данных, алгоритмической предвзятостью, многоязычными ограничениями и этическими вопросами. Исследование показывает, что, несмотря на значительные преимущества данных технологий, для получения точных и ответственных результатов необходимы тщательные методологические и этические подходы.

Ключевые слова: цифровые корпусы, искусственный интеллект, лингвистическое моделирование, корпусная лингвистика, обработка естественного языка, компьютерная лингвистика

Annotatsiya. Raqamli texnologiyalarning jadal rivojlanishi zamonaviy tilshunoslik tadqiqotlarini sezilarli darajada o‘zgartirdi. Ushbu maqolada raqamli korpuslar, sun’iy intellekt va lingvistik modellashtirishning til tadqiqotlaridagi o‘rni tahlil qilinadi. Raqamli korpuslar keng ko‘lamli tahlil uchun autentik til materiallarini taqdim etsa, sun’iy intellekt texnologiyalari tabiiy tilni qayta ishlash, mashina tarjimasi va nutqni tanish tizimlarining rivojlanishiga xizmat qiladi. Lingvistik modellashtirish esa til tuzilmalari va qonuniyatlarini hisoblash usullari orqali ifodalash hamda bashorat qilish imkonini beradi. Maqolada ma’lumotlarning ishonchliligi, algoritmik tarafkashlik, ko‘p tillilik bilan bog‘liq cheklovlar va etik masalalar kabi asosiy muammolar ham ko‘rib chiqiladi. Tadqiqot natijalari ushbu

DOI: 10.5281/zenodo.20377906

texnologiyalar lingvistik tadqiqotlarni sezilarli darajada rivojlantirishini, biroq aniq va mas’uliyatli natijalarga erishish uchun metodologik hamda etik yondashuvlar muhimligini ko‘rsatadi.

Kalit so‘zlar: raqamli korpuslar, sun’iy intellekt, lingvistik modellashtirish, korpus lingvistikasi, tabiiy tilni qayta ishlash, kompyuter lingvistikasi

Introduction

The development of information and communication technologies has had a profound impact on modern linguistic research. Traditional approaches to language analysis are increasingly being combined with computational and digital methods. Digital corpora, artificial intelligence, and linguistic modeling have become important tools for investigating language structure, usage, and evolution.[1] Modern linguistic studies require large amounts of authentic language data and advanced analytical methods. Digital corpora provide researchers with access to extensive collections of written and spoken texts, enabling more objective and evidence-based linguistic analysis. At the same time, artificial intelligence technologies improve the efficiency of natural language processing and automated language analysis.[2]

Linguistic modeling further contributes to the understanding of language systems through mathematical and computational representations. These developments have opened new possibilities for interdisciplinary research involving linguistics, computer science, and cognitive studies.

Literature Review

The significance of corpus linguistics and computational approaches has been widely discussed in academic research. Corpus-based studies allow linguists to analyze lexical frequency, grammatical structures, discourse patterns, and language variation using authentic language materials.[3]

Researchers have also emphasized the growing importance of artificial intelligence in language studies. AI technologies are widely applied in machine translation, speech recognition, text generation, and sentiment analysis. These systems rely on machine learning algorithms that process large datasets to identify linguistic patterns.[4]

Linguistic modeling is based on statistical and computational frameworks that simulate language processes. According to modern linguistic theories, computational models can help explain language acquisition, communication patterns, and language evolution.[5] However, scholars also note that language complexity and cultural context remain difficult for computational systems to fully interpret.

Methodology

This study uses a qualitative research approach based on the analysis of academic literature and existing technological applications in linguistics. Various studies related to digital corpora, artificial intelligence, and linguistic modeling were examined to identify their major functions, advantages, and limitations.

Comparative analysis was applied to evaluate the role of computational technologies in modern linguistic research. The study also considers ethical and methodological challenges associated with digital language analysis.

Results

The findings demonstrate that digital corpora significantly improve the efficiency and accuracy of linguistic research. Researchers can analyze large volumes of authentic language data in relatively short periods of time. Corpus-based analysis also increases objectivity because conclusions are based on real language usage rather than intuition.

Artificial intelligence technologies contribute to automated language processing, machine

translation, and speech recognition. AI systems are capable of identifying complex linguistic patterns and supporting multilingual communication.

Linguistic modeling provides opportunities for predicting language changes and analyzing communication patterns through computational methods. These models help researchers better understand the structure and function of language systems.

Discussion

The integration of digital corpora, artificial intelligence, and linguistic modeling has transformed the field of linguistics. One of the most important advantages of digital corpora is the accessibility of authentic linguistic materials from different genres, social contexts, and historical periods. Researchers can investigate language variation and change more effectively using corpus-based approaches.[6]

Artificial intelligence technologies also play a crucial role in modern linguistic analysis. Natural language processing systems allow computers to process and interpret human language automatically. AI applications such as chatbots, virtual assistants, and translation systems demonstrate the practical importance of computational linguistics in everyday communication.

However, several challenges remain significant. One major issue is corpus representativeness. Many digital corpora primarily contain materials from dominant languages, while minority languages remain underrepresented. This imbalance limits the inclusiveness of linguistic research.

Another important problem involves algorithmic bias. AI systems learn from existing datasets, and if those datasets contain cultural or social biases, the systems may reproduce discriminatory patterns. For instance, machine translation programs sometimes generate gender stereotypes due to biased training data.

Ethical considerations are equally important in digital linguistic research. The collection and analysis of online communication data may raise concerns regarding privacy, copyright, and informed consent. Researchers must therefore ensure responsible data management and ethical research practices.

Linguistic modeling also faces methodological limitations. Human language contains emotional, cultural, and contextual dimensions that cannot always be accurately represented through mathematical or computational models. As a result, computational systems may struggle with irony, metaphor, sarcasm, and pragmatic meaning.

Despite these challenges, digital technologies continue to expand opportunities for interdisciplinary collaboration. Linguistics increasingly intersects with computer science, artificial intelligence, psychology, and cognitive science, creating new directions for future research.

Conclusion

Digital corpora, artificial intelligence, and linguistic modeling have become essential components of contemporary linguistic research. These technologies improve analytical efficiency, support large-scale language studies, and contribute to the development of computational linguistics.

At the same time, researchers must address important challenges related to data quality, multilingual inclusion, algorithmic bias, and ethical responsibility. Future studies should focus on developing more representative corpora, improving AI transparency, and creating more context-sensitive linguistic models.

Overall, the effective integration of digital technologies into linguistics requires a balanced approach that combines technological innovation with methodological rigor and ethical awareness.

Reference list:

1. McEnery, T., & Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012.
2. Jurafsky, D., & Martin, J. H. *Speech and Language Processing*. Pearson Education, 2023.
3. Biber, D., Conrad, S., & Reppen, R. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, 1998.
4. Manning, C. D., & Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
5. Crystal, D. *Language and the Internet*. Cambridge University Press, 2006.
6. Bird, S., Klein, E., & Loper, E. *Natural Language Processing with Python*. O’Reilly Media, 2009.
7. McCarthy, J. *Artificial Intelligence: A Modern Approach*. Pearson, 2021.
8. Leech, G. *Introducing Corpus Linguistics*. Routledge, 2014.
9. Russell, S., & Norvig, P. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2020.
10. Sinclair, J. *Corpus, Concordance, Collocation*. Oxford University Press, 1991.