DIGITAL CORPORA, ARTIFICIAL INTELLIGENCE AND LINGUISTIC MODELLING IN LINGUISTIC RESEARCH

Mamanov Elmurod Faxriddin o`g`li
Faculty N1 Foreign Language and Literature, 3rd year student
elmurodmamanov9@gmail.com
Research supervisor: Alikulova Shaxnoza Abdullo qizi:
O‘zbekiston davlat jahon tillari universiteti

Annotatsiya. Ushbu maqolada raqamli korpuslar, sun’iy intellekt va lingvistik modellashtirishning zamonaviy tilshunoslikdagi o‘rni tahlil qilinadi. So‘nggi yillarda bu texnologiyalar tilni o‘rganish usullarini sezilarli darajada o‘zgartirib, uni an’anaviy yondashuvlardan ko‘ra ko‘proq ma’lumotlarga asoslangan va avtomatlashtirilgan jarayonga aylantirdi. Raqamli korpuslar real til namunalarini taqdim etsa, sun’iy intellekt ushbu ma’lumotlarni tez va samarali qayta ishlash imkonini beradi. Lingvistik modellashtirish esa til tuzilishini tushuntirish va ayrim til hodisalarini oldindan bashorat qilishga yordam beradi. Tadqiqot natijalari ushbu yondashuvlarning tilshunoslik tadqiqotlarining aniqligi va ko‘lamini oshirishini ko‘rsatadi. Shu bilan birga, ma’lumotlar nomutanosibligi, algoritmik xatoliklar hamda kam resursli tillar uchun imkoniyatlarning cheklanganligi kabi muammolar saqlanib qolmoqda.

Kalit so‘zlar: Raqamli korpuslar, sun’iy intellekt, lingvistik modellashtirish, korpus lingvistikasi, tabiiy tilni qayta ishlash NLP, hisoblash lingvistikasi, til texnologiyasi.

Аннотация. В данной статье рассматривается роль цифровых корпусов, искусственного интеллекта и лингвистического моделирования в современной лингвистике. В последние годы эти технологии существенно изменили подход к изучению языка, сделав его более основанным на данных и автоматизированным. Цифровые корпусы предоставляют реальные языковые примеры, а искусственный интеллект обеспечивает их быструю и эффективную обработку. Лингвистическое моделирование, в свою очередь, помогает объяснять языковые структуры и прогнозировать языковые явления. Результаты исследования показывают, что использование этих инструментов повышает точность и масштаб лингвистических исследований. Однако сохраняются такие проблемы, как дисбаланс данных, алгоритмическая предвзятость и недостаток ресурсов для языков с ограниченной поддержкой.

Ключевые слова: Цифровые корпуса, искусственный интеллект, лингвистическое моделирование, корпусная лингвистика, обработка естественного языка NLP, компьютерная лингвистика, языковые технологии.

Abstract. This paper explores the role of digital corpora, artificial intelligence and linguistic modelling in modern linguistic research. In recent years, these technologies have significantly transformed language studies, making them more data-driven and automated compared to traditional approaches. Digital corpora provide authentic language data, while artificial intelligence enables fast and efficient data processing. Linguistic modelling helps explain language structures and predict certain linguistic patterns. The findings show that these tools improve the accuracy and scope of linguistic research. However, challenges such as data imbalance, algorithmic bias and limited resources for low-resource languages still remain.

DOI: 10.5281/zenodo.20377865

Keywords: Digital corpora, artificial intelligence, linguistic modelling, corpus linguistics, NLP, computational linguistics, language technology.

Introduction

In the past, linguistic research mostly depended on small text samples and the intuition of researchers. Today, this situation has changed significantly due to technological development. Researchers now rely more on digital corpora and computational tools to study language in a more objective and systematic way.

Digital corpora are large collections of real-life language data taken from sources such as books, newspapers, websites and spoken communication. They make it possible to analyze language as it is actually used in real contexts.

Artificial intelligence has become an essential part of linguistic research. Through natural language processing (NLP), AI systems can analyze, interpret and even generate human language. Applications such as machine translation, chat-bots and automated text analysis clearly demonstrate this.

Linguistic modelling plays a connecting role by providing structured and often mathematical ways to represent language patterns. However, not all languages benefit equally from these developments. For example, the Uzbek language still lacks sufficient digital resources, which limits research and technological progress.

Literature Review

Corpus linguistics is widely considered one of the foundations of modern linguistic research. McEnery (2016) emphasizes that corpora provide more objective and reliable data compared to traditional methods.

Dunn (2022) notes that NLP technologies allow the automatic processing of large text datasets, which improves both efficiency and accuracy. Large datasets such as S2ORC (Lo et al., 2019) also demonstrate how important big data is for training AI systems.

At the same time, some researchers point out existing limitations. Warstadt et al. (2018) show that even advanced neural models still struggle with deep grammatical understanding and contextual reasoning.

Overall, previous studies indicate strong progress in the field, while also highlighting several challenges that still need attention.

Methods

1.This research is based on a qualitative literature review. Academic books, journal articles, and online resources related to corpus linguistics and computational linguistics were carefully analyzed.

2.The collected information was grouped into three main categories:

digital corpora and their linguistic function

artificial intelligence in language processing

linguistic modelling approaches

3.Each category was examined in terms of its role, advantages, and limitations..

Results

The analysis shows that digital corpora are essential for studying real language usage. They provide valuable data such as frequency, contextual patterns, and authentic examples.

Artificial intelligence significantly improves the speed and scale of linguistic analysis. It is widely applied in translation, sentiment analysis, text classification, and language generation.

Linguistic modelling offers a structured way to represent language mathematically and supports predictive systems in NLP.

However, several challenges remain. One of the main issues is the lack of sufficient linguistic resources for many languages, including Uzbek. Another issue is bias in AI systems, which can affect fairness and accuracy.

Discussion

The findings show that digital corpora, artificial intelligence, and linguistic modelling are closely interconnected. In simple terms, corpora provide the data, AI processes it, and models help interpret the results.

Despite these advantages, some important challenges still exist. One major issue is the unequal distribution of language resources. While English and other major languages have large corpora, many smaller languages remain underrepresented.

Another issue is the limited transparency of AI systems. Although they are effective, their decision-making processes are often difficult to understand.

Finally, human linguistic knowledge remains essential, as machines still struggle to fully understand cultural meanings, idiomatic expressions, and deeper contextual nuances.

Conclusion

In conclusion, digital corpora, artificial intelligence, and linguistic modelling have significantly transformed modern linguistics. They enable more accurate, efficient, and large-scale language analysis.

At the same time, challenges such as data imbalance, AI bias and limited resources for low-resource languages continue to exist. Future research should focus on expanding linguistic corpora, improving model transparency and supporting underrepresented languages.

References

1. McEnery, T. (2016). Corpus Linguistics.
2. Dunn, J. (2022). Natural Language Processing for Corpus Linguistics.
3. Lo, K. et al. (2019). S2ORC Dataset.
4. Warstadt, A. et al. (2018). Neural Network Acceptability Judgments.
5. Dash, N. S. (2024). Corpus Linguistics and Language Technology.