European International Journal of Multidisciplinary Research and Management Studies (ISSN: 2750-8587)

Received: 38 July 2026 | Accepted: 25 August 2026 | Published: 18 September 2026

DOI: 10.55640/eijmrms-06-09-02

RESEARCH ARTICLE

General Concepts on The Integration of Corpus Linguistics and Artificial Intelligence

Rajapova Malika Akhmadali kizi

PhD in philological sciences, Kokand state university, Uzbekistan

VOLUME: Vol.06 Issue09 2026

PAGE: 8-12

Abstract

This article analyzes the processes of integration between corpus linguistics and artificial intelligence technologies. It examines the use of modern artificial intelligence capabilities, particularly machine learning and deep neural networks, at the stages of compiling, collecting, tagging, and annotating linguistic corpora. The aim of the research is to theoretically and practically substantiate the interconnection between corpus linguistics methodology and artificial intelligence algorithms, and to determine their place within modern computational linguistics.

KEYWORDS

Corpus linguistics, artificial intelligence, machine learning, neural networks, natural language processing, national corpus, tagged annotation, computational linguistics.

INTRODUCTION

Stress Over the past decade, the rapid development of digital technologies has given rise to a distinct and increasingly significant field within linguistics — corpus linguistics. Corpus linguistics offers a methodology for studying language on the basis of authentic texts, that is, large, systematically structured collections of texts stored in electronic form. This approach enables linguists to analyze lexical, grammatical, and stylistic phenomena objectively, on the basis of statistical data rather than intuition alone. At the same time, advances in artificial intelligence, particularly in machine learning, deep learning, and natural language processing (NLP), have opened new possibilities for corpus linguistics. The automatic compilation of large linguistic corpora, the morphological and syntactic tagging of texts, semantic annotation, and the training of language models on corpus data are processes that can no longer be imagined without artificial intelligence algorithms.

The relevance of the topic lies in the fact that, in world linguistics, national corpora developed for English, Russian, German, Chinese, and other languages already operate in close integration with artificial intelligence tools (for example, projects such as the British National Corpus and the Corpus of Contemporary American English). In Uzbek linguistics, however, work on creating a national corpus and enriching it with modern artificial intelligence technologies is still at an early stage, which makes the development of theoretical and methodological foundations in this direction of considerable scientific and practical importance.

The purpose of this study is to theoretically analyze the processes of mutual integration between corpus linguistics and artificial intelligence technologies, to determine their place within the system of modern computational linguistics, and to develop scientifically grounded proposals for applying artificial intelligence capabilities to the development of the National Corpus of the Uzbek language.

The objectives of the study are as follows: 1) to outline the theoretical and methodological foundations of corpus linguistics; 2) to determine the functional role of artificial intelligence technologies in the process of corpus compilation; 3) to classify the artificial intelligence methods used in corpus-based linguistic research; 4) to identify prospects for the development of the Uzbek language corpus.

LITERATURE REVIEW

The theoretical foundations of corpus linguistics have been thoroughly elaborated in Western linguistics in the works of J. Sinclair, D. Biber, T. McEnery, and A. Wilson, who define a corpus as "a collection of naturally occurring texts, selected according to explicit criteria, compiled for the purpose of studying language in actual use." In Russian linguistics, scholars such as V. Plungian, V. Rykov, and A. Baranov have developed the methodological foundations of corpus linguistics, including the principles of corpus compilation and annotation systems.

In the field of artificial intelligence and natural language processing, the works of T. Mitchell, Y. Bengio, C. Manning, and D. Jurafsky occupy an important place. The emergence of language models based on deep learning — most notably transformer-based architectures such as BERT and GPT — has further strengthened the connection between corpus linguistics and artificial intelligence, since training such models requires large volumes of high-quality, annotated linguistic corpora.

In Uzbek linguistics, corpus linguistics remains a relatively new direction. The general linguistic methodology developed by scholars such as A. Nurmonov, N. Mahmudov, and Sh. Rahmatullayev serves as a theoretical foundation, and in recent years practical projects aimed at creating a national corpus of the Uzbek language have also begun. Nevertheless, the specific question of integrating the Uzbek language corpus with artificial intelligence technologies has not yet been sufficiently studied as an independent research object, which determines the scientific novelty of the present article.

METHODOLOGY

The study employed a complex of general scientific and specific linguistic methods:

— the descriptive method — to systematize the conceptual and terminological apparatus of corpus linguistics and artificial intelligence technologies;

— the comparative-typological method — to compare the structure of national corpora in different languages (English, Russian, Uzbek) and the degree to which artificial intelligence tools are used in them;

— the content-analysis method — for the systematic analysis of relevant scholarly sources and technical documentation;

— statistical processing of corpus materials — used to evaluate the frequency distribution of linguistic units and the accuracy of automatic annotation.

Open-access national corpora (the British National Corpus, the Corpus of Contemporary American English, and the National Corpus of the Russian Language), as well as preliminary data collected within the framework of the Uzbek national corpus project, were selected as the material for analysis. In addition, the operating principles of artificial-intelligence-based text-processing tools — automatic tagging systems, morphological analyzers, and neural-network-based tokenizers — were examined.

RESULTS

1. Basic concepts of corpus linguistics

A language corpus is a collection of texts selected in accordance with a specific purpose, stored in electronic form, and equipped with specialized software tools for search and analysis. The main characteristics of a corpus include its representativeness (the extent to which the selected texts accurately reflect the state of the language), size, balance (the proportion of different genres and styles), and machine-readability. Corpora are typically classified into the following types: general (national) corpora, specialized corpora, parallel corpora, learner corpora, and diachronic corpora.

2. Functions of artificial intelligence in the corpus-compilation process

The results of the study show that artificial intelligence technologies are applied at nearly every stage of corpus linguistics. This process can be classified in tabular form as follows:

No.	Stage of corpus compilation	Artificial intelligence technology applied
1	Text collection (web corpus)	Web-scraping algorithms, automatic language identification models
2	Tokenization and lemmatization	Neural-network-based tokenizers, deep learning models
3	Morphological and syntactic tagging	Part-of-Speech (POS) taggers, dependency parsers
4	Semantic annotation	Word Sense Disambiguation, Named Entity Recognition
5	Training language models on corpus data	Transformer architectures (BERT-, GPT-type models), transfer learning
6	Automatic error detection and cleaning	Classification algorithms, anomaly-detection models

3. Advantages and practical outcomes of integration

The analysis showed that integrating artificial intelligence tools into corpus linguistics yields a number of significant advantages. First, compared with traditional manual annotation, automatic tagging systems increase text-processing speed by dozens, or even hundreds, of times. Second, modern neural network models achieve high accuracy — in some studies exceeding 90 percent — even for morphologically rich, agglutinative languages such as Uzbek. Third, artificial-intelligence-based corpus systems enable users to automatically identify collocations and syntactic constructions, thereby raising the overall quality of linguistic research.

At the same time, the study identified a number of problematic aspects: the shortage of large, high-quality annotated datasets for low-resource languages, including Uzbek; the high computational requirements needed to train artificial intelligence models; and the continued necessity of linguistic verification and validation of automatically generated annotation results.

DISCUSSION

The findings obtained are broadly consistent with conclusions recognized in world linguistics: artificial intelligence technologies do not replace corpus linguistics but rather enrich and extend it methodologically. As D. Jurafsky and C. Manning note, the essential precondition for building high-quality language models is the availability of large, linguistically well-annotated corpora; corpus linguistics therefore serves as the "data source" for artificial intelligence systems, while artificial intelligence, in turn, plays the role of a "methodological instrument" by automating the processes of corpus compilation and analysis. This interaction between the two fields is recursive in nature: a high-quality corpus produces better language models, and better language models, in turn, make it possible to annotate new corpora with even greater precision.

At the same time, within the context of Uzbek linguistics, this integration process faces certain specific challenges. The agglutinative nature of the Uzbek language creates additional complexity for automatic morphological analysis algorithms. In addition, the parallel use of the Latin and Cyrillic scripts, dialectal diversity, and the limited availability of high-quality annotated training data are among the key factors that must

be taken into account when developing artificial-intelligence-based corpus tools for the Uzbek language.

The results further indicate that transfer learning together with the use of multilingual language models, represents a promising direction for low-resource languages, as this approach makes it possible to leverage the knowledge embedded in models trained on resource-rich languages for the benefit of low-resource languages such as Uzbek.

CONCLUSION

Based on the results of the study conducted, the following conclusions can be formulated:

1. Corpus linguistics and artificial intelligence technologies are developing in modern linguistics as two closely interconnected, mutually complementary directions.

2. Artificial intelligence tools are effectively applied at every stage of corpus compilation — from text collection to tagging, annotation, and the training of language models — thereby significantly increasing both the speed and the accuracy of linguistic analysis.

3. In order to introduce artificial intelligence technologies into the development of the Uzbek national corpus, it is advisable, first of all, to build a high-quality annotated initial dataset, to adapt morphological analyzers to the agglutinative features of the Uzbek language, and to apply transfer learning methods.

4. Future research should give particular attention to developing neural-network-based morphological analyzers and POS-tagging systems specifically for the Uzbek language, as well as to increasing the representation of Uzbek within multilingual language models.

REFERENCES

1. Sinclair J. Corpus, Concordance, Collocation. — Oxford: Oxford University Press, 1991. — 179 p.
2. McEnery T., Wilson A. Corpus Linguistics: An Introduction. — Edinburgh: Edinburgh University Press, 2001. — 235 p.
3. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. — Cambridge: Cambridge University Press, 1998. — 300 p.
4. Jurafsky D., Martin J.H. Speech and Language Processing. — 3rd ed. — Stanford: Stanford University, 2023. — 636 p.
5. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. — Cambridge, MA: MIT Press, 1999. — 680 p.
6. Plungian V.A. Why Do We Need the National Corpus of the Russian Language? An Informal Introduction // National Corpus of the Russian Language: 2003–2005. — Moscow: Indrik, 2005. — P. 6–20.
7. Rykov V.V. Corpus Linguistics: A Study Guide. — Moscow: Russian State University for the Humanities, 2019. — 112 p.
8. Nurmonov A. History of Uzbek Linguistics. — Tashkent: O'qituvchi, 2002. — 296 p.
9. Mahmudov N. Language and Its Social Nature. — Tashkent: Fan, 2019. — 188 p.
10. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. — 2019. — P. 4171–4186.
11. Bengio Y., Ducharme R., Vincent P., Jauvin C. A Neural Probabilistic Language Model // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 1137–1155.
12. State Department for the Development of the State Language under the Cabinet of Ministers of the Republic of Uzbekistan. The National Corpus of the Uzbek Language Project. — Tashkent, 2022.
13. Ganiyevich, M. T., & Saminovna, M. D. (2023). Factors in The Formation of a Healthy Lifestyle Among University Students. Journal of Advanced Zoology, 44(5).
14. Dilfuza, M. (2023). Exploring the benefits of bilingualism in early childhood.". In CANADA" INTERNATİONAL CONFERENCE ON DEVELOPMENTS İN EDUCATİON, SCİENCESAND HUMANİTİES (Vol. 9, No. 1).
15. Ahmedova, H. T., Mamajonova, M. N., Mirzaliyev, T., & Mirzaliyeva, D. S. (2022). The Role And Importance Of The East Individual Creators In The Development Of Rhetoric Subject. Journal of Positive School Psychology, 6(11).
16. Saminovna, M. D., & Solijanovna, M. D. (2021). Integrated Approach to Russian Language Lessons.

Academicia Globe, 2(04), 179-182.

17. Mirzaliyeva, D. (2022). UNCONVENTIONAL METHODS OF TEACHING IN RUSSIAN LANGUAGE LESSONS. ASIA PACIFIC JOURNAL OF MARKETING & MANAGEMENT REVIEW ISSN: 2319-2836 Impact Factor: 8.071, 11(11), 149-151.

18. Mirzalieva, D. S. (2022). Principles of formation of the culture of speech of students in the Russian language classes at the university. INTERNATIONAL JOURNAL OF SOCIAL SCIENCE & INTERDISCIPLINARY RESEARCH ISSN: 2277-3630 Impact factor: 8.036, 11(12), 303-305.