The Use Of Generative Ai Tools In Corpus Linguistics: A Comparative Analysis Based On Uzbek And Russian Language Materials

Kommuna Khursanovna Umurzakova[1], Guzalkhon Makhmudovna Yakubova[2], Nargiza Rustamovna Abdukakhkhorova[3], Anna Vitalyevna Urazkulova[4], Gayane Arturovna Tevosyan[5], Zarema Rustemovna Sherefetdinova[6]

Abstract

As Corpus linguistics has not been studied fully yet, the purpose of the article is to find out how accurately generative AI tools can assist in corpus-based linguistic analysis of Uzbek and Russian language materials, and what methodological limitations emerge in low-resource and relatively high-resource language contexts. Moreover, the study is based on a qualitative review and comparative synthesis of recent research on large language models, corpus-based AI systems, Uzbek computational linguistics, semantic tagging, tense annotation, web-corpus neologism extraction, and Russian corpus traditions. The findings show that Russian, as a relatively high-resource language with longer-standing corpus infrastructure, benefits from more stable morphological, syntactic, and semantic processing tools. Uzbek, by contrast, is undergoing active corpus development, with growing research on semantic analyzers, author corpora, grammatical annotation, and web-based lexical databases. Generative AI can support Uzbek corpus linguistics by assisting in query generation, preliminary annotation, concordance interpretation, translation alignment, metadata enrichment, and educational material generation. Nevertheless, the study emphasizes that generative AI cannot replace linguistically verified corpus methods. Problems such as hallucination, inconsistent corpus grounding, bias toward high-resource languages, inaccurate treatment of agglutinative morphology, and unreliable semantic interpretation require human expert validation. The article argues for a hybrid model in which generative AI functions as an assistant to corpus linguists rather than as an autonomous analytical authority. Such a model is especially important for Uzbek, where corpus resources are still developing and where careful integration of linguistic expertise, corpus design, and AI-based automation can accelerate the creation of reliable digital language resources.

[1]Associate Professor, Faculty of Philology, Kokand University, Kokand, Uzbekistan.
Email: umurzakovakommunaxon@gmail.com
[2]Kokand University, Kokand, Uzbekistan
[3]Kokand University, Kokand, Uzbekistan
[4]Kokand University, Kokand, Uzbekistan
[5]Kokand University, Kokand, Uzbekistan. ORCID: https://orcid.org/0000-0001-5200-8651.
Email: gayatevosyan@gmail.com
[6]Kokand University, Kokand, Uzbekistan. ORCID: https://orcid.org/0009-0004-9918-546X.
Email: zara.sherefetdinova@gmail.com

Keywords: generative AI; corpus linguistics; Uzbek language; Russian language; comparative linguistics; NLP; low-resource languages; AI-assisted analysis.

Introduction

With the development of computer technologies, the field of corpus linguistics also began to expand significantly. Opportunities for storing large volumes of texts in electronic form, searching them, and conducting statistical analysis became increasingly available. The emergence of corpus linguistics dates back to the 1960s. In particular, the Brown Corpus of the English language, created in the 1960s, can rightly be considered an important milestone in the development of this field.

Today, corpus linguistics is closely connected with a number of areas, including artificial intelligence, translation software, various AI-based language models, automatic text editing, speech recognition, sentiment analysis, electronic dictionaries, and educational technologies.

The development of corpus linguistics has also been influenced by the growing needs of scientific research, whether in the form of writing academic articles, conducting small-scale theoretical studies, or carrying out large-scale linguistic investigations. These processes, together with fundamental changes in linguistic research, have contributed to the emergence and establishment of corpus linguistics as an important branch of modern linguistics.

Large language models and other generative AI systems are now capable of doing various activities like producing summaries, classifying texts, identifying semantic patterns, generating translations, and assisting in the interpretation of large-scale linguistic data. These functions are particularly relevant to corpus linguistics, where researchers work with extensive collections of written or spoken texts and attempt to identify recurrent lexical, grammatical, semantic, and discourse patterns. From a linguistic perspective, the emergence of computational linguistics has created new opportunities for addressing a number of current problems in the field. The application of computer technologies to scientific, theoretical, and philosophical issues in linguistics such as

- language and speech
- system and structure
- semiotics
- syntagmatics and paradigmatics
- typology, and lexicography has demonstrated significant advantages in modern linguistic research.

On the basis of these developments, the field of corpus linguistics gradually emerged. There is also a common view that computational linguistics and corpus linguistics are similar, and in some cases even identical, fields. However, despite certain shared features, there are substantial differences between them.

In this field, a considerable number of studies have been conducted in the Russian language, and research in this area is still ongoing. However, in Uzbek linguistics, this direction remains one of the areas that requires further systematic investigation. Academic English version:

The aim of this small-scale study is to compare the development and scholarly investigation of corpus linguistics in Uzbek and English. In doing so, the article seeks to examine several key aspects, including the use of artificial intelligence in linguistic research, its impact on linguistics, and the influence of AI-generated language on youth speech. These issues constitute the main objectives of the present study.

In addition, the article discusses the relationship between generative AI and corpus linguistics, as well as the linguistic analysis of Uzbek and Russian texts from various perspectives. It also examines issues related to AI adaptation to these languages and the extent to which generative AI tools can be effectively applied in the corpus-based study of Uzbek and Russian language materials. Also, the study addresses to these research questions: How do generative AI tools perform in the corpus-based analysis of Uzbek and Russian language materials? What types of linguistic errors occur when AI tools analyze Uzbek and Russian corpus data?

Literature Review

As mentioned, corpus linguistics has become one of the most influential methodological directions in contemporary language studies. Its strength lies in the systematic analysis of authentic language data through frequency counts, concordance lines, collocation patterns, semantic annotation, grammatical tagging, and discourse-level investigation. In the digital era, corpus linguistics has moved beyond simple text collection and has become closely connected with natural language processing, computational linguistics, machine translation, lexicography, linguistic typology, and language education.

The emergence of generative artificial intelligence, especially large language models, has significantly changed how researchers interact with textual data. Unlike traditional corpus tools, which usually require knowledge of formal query languages, part-of-speech tags, regular expressions, or corpus-specific interfaces, generative AI tools allow researchers to ask natural-language questions, generate hypotheses, create preliminary annotations, summarize concordance patterns, and transform corpus output into pedagogical or analytical materials. Recent studies on AI-assisted corpus systems demonstrate that generative AI can help users convert research questions into corpus queries, generate corpus-based exercises, support semantic classification, and facilitate linguistic interpretation. However, these possibilities are accompanied by serious methodological concerns, including hallucination, lack of source transparency, weak grounding in actual corpus data, and uneven performance across languages.

The aim of the article is to identify the major possibilities, limitations, and methodological requirements of using generative AI tools in Uzbek and Russian corpus linguistics.

Large language models are trained on massive text datasets and are capable of generating fluent, contextually relevant language. Their main advantage for corpus linguistics is their ability to process natural-language instructions and

produce explanations, classifications, summaries, and reformulations. Recent research has shown that AI-based corpus assistants can translate ordinary research questions into formal corpus queries and help users interpret corpus output. This is particularly important because many corpus interfaces remain difficult for non-specialists, especially when they require knowledge of Corpus Query Language or advanced search syntax.

The AI Corpus Linguist model, for example, has been described as a system that allows researchers to interact with corpora using natural language rather than formal query languages. Such tools lower the technical barrier to corpus-based research and allow wider use of corpora in linguistics, language teaching, discourse analysis, and lexicography. Similarly, research on the automatic generation of corpus-based exercises demonstrates that generative AI can use collocation and translation patterns extracted from tagged corpora to create pedagogical materials. This shows that AI can mediate between corpus data and educational application.

Nevertheless, studies also emphasize that AI-generated corpus outputs require critical evaluation. Generative AI may produce plausible but inaccurate explanations, invent unsupported examples, misinterpret corpus frequency, or use corpus data inconsistently. Therefore, the integration of generative AI into corpus linguistics should be understood as a form of assisted analysis rather than independent scientific interpretation.

Large language models are increasingly used in linguistic research for classification, annotation, semantic analysis, discourse analysis, and language comparison. They can assist in identifying lexical patterns, classifying semantic roles, generating hypotheses about grammatical constructions, and comparing linguistic features across languages. However, they are not neutral linguistic instruments. Their performance depends on the training data, the language represented in that data, the quality of prompts, and the availability of external grounding resources.

Research on multilingual large language models shows that performance is uneven across languages. High-resource languages tend to benefit from richer training data, stronger evaluation benchmarks, and better-developed NLP tools. Low-resource or underrepresented languages may experience weaker performance, more frequent errors, and stronger dependence on translation through dominant languages. This is directly relevant to Uzbek, which is less represented in global AI training datasets than Russian or English.

Another important issue is linguistic bias. Large language models may reproduce structural, cultural, and ideological biases present in their training data. In multilingual contexts, they may also privilege high-resource languages and impose patterns from those languages onto lower-resource languages. For corpus linguistics, this means that AI-generated analysis must be checked against actual corpus evidence and language-specific grammatical knowledge.

Uzbek corpus linguistics has developed significantly in recent years. Earlier stages of Uzbek computational linguistics focused on lexicography, automatic translation, morphological analysis, text editing, and linguistic modeling. More recent studies have shifted toward the creation of national and specialized corpora, semantic analyzers, author corpora, annotation systems, and web-based lexical databases.

Research on Uzbek sentence semantic analysis has emphasized the importance of developing models, algorithms, and information systems capable of analyzing Uzbek linguistic structures. Such studies show that Uzbek requires specialized computational solutions because its lexical, grammatical, and semantic features cannot be fully processed by tools designed for Indo-European languages. Semantic analysis is especially important for resolving homonymy, polysemy, and multifunctional word forms.

The creation of Alisher Navoi's author corpus and its semantic tag database represents another important direction in Uzbek corpus linguistics. This research demonstrates the relevance of corpus methods not only for contemporary language analysis but also for the digital preservation and interpretation of classical literary heritage. Semantic tagging of literary texts allows researchers to study lexical meaning, imagery, thematic fields, and cultural concepts in a structured digital environment.

The annotation of tense forms for the Uzbek language corpus is also significant. Uzbek tense forms are complex and context-dependent, and their automatic annotation requires grammatical, semantic, and contextual levels of analysis. This is especially relevant for generative AI, because AI systems may misinterpret tense if they rely only on surface forms and fail to consider context.

Recent research on Uzbek web corpora and neologism databases shows the importance of collecting data from news platforms, social media, and digital publications. Such resources make it possible to track new lexical units, measure frequency, analyze distribution, and observe semantic change. These developments are highly relevant to AI-assisted corpus linguistics because generative AI can support the extraction, classification, and preliminary explanation of neologisms, while human experts verify the results.

Russian corpus linguistics has a longer institutional and technological history than Uzbek corpus linguistics. Russian benefits from large national corpus resources, established morphological annotation practices, syntactic tagging, and extensive computational tools. Russian is also more strongly represented in multilingual language models than Uzbek. As a result, generative AI tools generally perform better with Russian lexical, grammatical, and semantic material than with Uzbek.

However, Russian corpus linguistics also faces challenges. Russian morphology is complex, with rich inflection, aspectual distinctions, case marking, and free word order. Automatic analysis of Russian requires robust morphological disambiguation, syntactic parsing, and semantic interpretation. Generative AI can assist in these tasks but should not replace corpus-based verification.

The comparison between Uzbek and Russian is therefore methodologically useful. Russian represents a relatively resource-rich Slavic language with developed corpus infrastructure, while Uzbek represents a developing corpus environment for a Turkic agglutinative language. The contrast helps reveal how generative AI performs differently depending on corpus availability, grammatical typology, and resource maturity.

Methods

This study uses a qualitative comparative and analytical methodology. The research is based on the review and synthesis of scholarly materials on generative AI, large language models, corpus linguistics, Uzbek computational linguistics, semantic annotation, tense tagging, web corpora, neologism extraction, and Russian corpus traditions.

The materials were grouped into four categories: Generative AI and corpus interaction which studies on AI corpus assistants, corpus-based exercise generation, prompt-based corpus analysis, and AI-supported language learning. Next, Large language models and linguistic research which studies on LLM capabilities, limitations, multilingual behavior, linguistic bias, and ethical risks. Also, Uzbek corpus linguistics: studies on semantic analysis of Uzbek sentences, Alisher Navoi's author corpus, tense annotation for Uzbek corpora, computational linguistics, and Uzbek web-corpus neologisms. Last, Comparative Uzbek–Russian corpus perspective: materials discussing corpus structure, annotation standards, morphological and semantic processing, and the different levels of resource development in Uzbek and Russian.

The study applies comparative content analysis. The comparison is organized around five analytical criteria: corpus infrastructure, morphological and grammatical processing, semantic annotation and interpretation, AI-assisted corpus querying and analysis, methodological and ethical reliability.

The article does not present a quantitative experiment with newly compiled Uzbek and Russian corpora. Instead, it offers a conceptual and methodological framework for AI-assisted corpus linguistics based on existing research. This approach is appropriate because the main goal is to identify possibilities and limitations before implementing a large-scale empirical AI-corpus experiment.

Results And Discussions

The first major finding is that the effectiveness of generative AI in corpus linguistics depends on the maturity of the corpus infrastructure. Russian is more AI-ready because it has larger corpora, more stable annotation traditions, and stronger integration with existing NLP tools. Generative AI can therefore be used more easily for Russian corpus querying, concordance interpretation, collocation analysis, and grammatical explanation.

Uzbek is developing quickly, but its corpus infrastructure is still in a formative stage. The creation of semantic analyzers, author corpora, tense annotation models, and web-based neologism databases shows strong progress. However, Uzbek still needs more standardized annotation schemes, larger balanced corpora, reliable morphological analyzers, and open-access computational tools. For this reason, generative AI should be used carefully in Uzbek corpus research. It can accelerate analysis, but it must be grounded in verified Uzbek corpora and evaluated by Uzbek linguists.

Both Uzbek and Russian require advanced morphological processing, but the nature of the challenge differs. Russian morphology is inflectional and involves case, gender, number, aspect, tense, and agreement. Uzbek morphology is agglutinative and involves suffix chains, derivational complexity, tense-aspect-modality markers, possessive markers, case markers, and word-form ambiguity.

For Uzbek, tense annotation is particularly important because a single grammatical form may express different temporal meanings depending on context. The distinction between grammatical, semantic, and contextual levels of tense annotation is essential. Generative AI can assist by suggesting possible interpretations, but it may fail when it treats suffixes mechanically or ignores contextual meaning.

For Russian, generative AI may be more accurate due to richer training data, but it still faces challenges with aspect, verbal government, free word order, and syntactic ambiguity. Therefore, in both languages, AI-assisted annotation should be treated as preliminary and subject to expert validation.

Semantic annotation is one of the most promising areas for generative AI in corpus linguistics. AI tools can help classify semantic fields, identify possible meanings of polysemous words, group lexical items by conceptual domains, and generate explanations of concordance patterns.

In Uzbek, semantic annotation is especially important for developing semantic analyzers and author corpora. The semantic tagging of Alisher Navoi's works demonstrates how corpus linguistics can be used to preserve and analyze cultural heritage. Generative AI may assist in preliminary semantic grouping, but classical literary texts require deep philological knowledge. AI systems may misunderstand archaic vocabulary, metaphorical usage, religious-cultural references, and poetic semantics.

In Russian, semantic annotation can benefit from larger digital resources and existing lexical databases. However, Russian literary and historical texts also contain semantic shifts, archaisms, and cultural references that require expert interpretation. Thus, for both Uzbek and Russian, generative AI is useful as a semantic assistant but not as an independent interpreter.

Generative AI is highly useful for transforming natural-language research questions into corpus queries. This is one of its most practical advantages. Many researchers and students find corpus query systems difficult because they require technical knowledge. AI tools can help formulate search strategies, propose keywords, generate regular expressions, explain part-of-speech tags, and suggest possible collocation searches.

For Russian, this function can be implemented more easily because of the availability of mature corpus platforms. For Uzbek, AI-assisted querying could be especially valuable because it may help students and researchers access developing corpora more easily. However, Uzbek corpus query systems must be designed with Uzbek morphology in mind. For example, search tools should recognize suffixal variation, lemmatized forms, and orthographic variation.

Web corpora are important for tracking new lexical items. Uzbek research on sectoral neologisms shows that web-based materials, especially news platforms and social media, can be used to identify new words, analyze their frequency, classify them by field, and study their distribution across time and context.

Generative AI can support this process by suggesting semantic classifications, explaining possible origins of new words, grouping neologisms by domain, and generating preliminary lexicographic descriptions. However, AI may also invent etymologies or incorrectly classify borrowed words. Therefore, neologism analysis must combine frequency-based corpus extraction, dictionary comparison, contextual analysis, and human verification.

Russian web corpora also provide rich material for neologism studies, especially in technology, politics, youth language, business, and media discourse. Compared with Uzbek, Russian AI-based neologism analysis is likely to benefit from more available digital data. Yet both languages require diachronic tracking to distinguish stable neologisms from short-lived trends.

Generative AI can also be used in data-driven learning. It can transform corpus examples into exercises, create gap-fill tasks, generate collocation activities, and explain authentic usage. This is especially useful for Uzbek and Russian language teaching because learners often need contextual examples rather than isolated rules.

However, corpus-based educational materials generated by AI must be checked carefully. AI may choose unnatural examples, simplify authentic data too much, or provide incorrect translation equivalents. Therefore, the best model is a combination of authentic corpus data, structured prompts, teacher evaluation, and learner-oriented adaptation.

Discussion

The comparison between Uzbek and Russian shows that generative AI does not function equally across languages. Russian benefits from a stronger digital ecosystem, while Uzbek requires more language-specific development. This difference reflects a broader problem in multilingual AI: high-resource languages are more accurately represented, while lower-resource languages are more vulnerable to error, simplification, and cross-linguistic interference.

For Uzbek, generative AI can be especially useful in four directions. First, it can assist in corpus construction by helping organize metadata, classify texts, and propose annotation categories. Second, it can support morphological and semantic annotation by generating preliminary labels. Third, it can help researchers interpret concordance lines and identify patterns. Fourth, it can support language teaching by converting corpus data into exercises.

However, Uzbek also presents risks for AI systems. Agglutinative morphology, polyfunctional suffixes, spelling variation, Arabic-Persian and Russian borrowings, dialectal variation, and limited training data can reduce AI accuracy. Therefore, any AI-based Uzbek corpus analysis must be verified through linguistic expertise.

For Russian, generative AI can be used more confidently for basic corpus tasks, but it still cannot replace corpus evidence. Russian aspect, case government, syntactic flexibility, and discourse-level meaning require careful analysis. AI-generated explanations must be checked against actual corpus examples.

The study supports a hybrid model of AI-assisted corpus linguistics. In this model, generative AI performs supportive tasks: query generation, preliminary annotation, clustering, summarization, explanation, exercise generation, and translation alignment. Human researchers remain responsible for corpus design, annotation standards, interpretation, validation, and theoretical conclusions.

This model is important for Scopus-level research because international journals require methodological transparency. Articles using generative AI in corpus linguistics should clearly state which AI tools were used, what prompts were applied, what corpus data were analyzed, how outputs were verified, and how errors were handled. Without such transparency, AI-assisted corpus research may lack scientific reliability.

A reliable model for using generative AI in Uzbek and Russian corpus linguistics may include the following stages:

1. Corpus selection which chooses verified Uzbek and Russian corpora or clearly defined web-corpus datasets.
2. Pre-processing that normalizes orthography, remove duplicates, tokenize texts, and prepare metadata.
3. Linguistic annotation which applies morphological, syntactic, and semantic annotation using available tools.
4. AI-assisted query formulation that uses generative AI to transform research questions into search strategies.
5. Corpus evidence extraction collects concordance lines, frequency lists, collocations, and semantic patterns.
6. AI-assisted interpretation asks AI to summarize patterns, but only after providing actual corpus output.
7. Human validation which verifies AI interpretations through expert linguistic analysis.
8. Error classification records hallucinations, mistranslations, wrong annotations, and unsupported generalizations.
9. Comparative analysis comparing Uzbek and Russian results according to typological and corpus-based criteria.
10. Transparent reporting which includes prompts, corpus details, validation procedure, and limitations.

This framework can reduce the risk of unsupported AI conclusions and improve the quality of corpus-based research.

Conclusion

Generative AI tools offer important opportunities for corpus linguistics, especially in query generation, preliminary annotation, semantic classification, corpus-based teaching, and the interpretation of large textual datasets. However, their value depends on corpus grounding, linguistic expertise, and methodological control.

The comparative analysis of Uzbek and Russian language materials shows that Russian currently has stronger corpus infrastructure and greater AI-readiness, while Uzbek is undergoing active development in semantic analysis, author corpus construction, tense annotation, and web-corpus neologism studies. This does not mean that Uzbek is unsuitable for AI-assisted corpus linguistics. On the contrary, generative AI can accelerate Uzbek corpus development if it is integrated with reliable linguistic models and expert validation.

The article concludes that generative AI should not be treated as a replacement for corpus linguists. It should be used as a research assistant that supports technical and analytical stages while human specialists control interpretation, accuracy, and theoretical validity. For Uzbek and Russian corpus linguistics, the most promising future lies in hybrid systems combining verified corpora, linguistic annotation, transparent AI prompting, and expert evaluation. Such an

approach can contribute to more reliable multilingual corpus research and support the development of digital resources for both high-resource and developing-resource languages.

Reference

1. Abdullayeva, O. (2026). Dissertation abstract on Uzbek language and digital-linguistic research. Andijan.
2. Axmedova, X. I. (2023). *Development of a model, algorithms and information system for semantic analysis of sentences in the Uzbek language*. PhD dissertation abstract. Tashkent.
3. Baumgart, J., Bohle-Jurok, U., & Mandl, T. (2024). Ink of innovation: Exploring ChatGPT and other AI-writing tools in an English as Medium of Instruction context. In *Exploring Artificial Intelligence in Applied Linguistics*. Iowa State University Digital Press.
4. G‘ulomova, N. S. (2023). *Creation of the author’s corpus of Alisher Navoi and its database of semantic tags based on “Badoye’ul-vasat”*. PhD dissertation abstract. Tashkent.
5. Koeva, S. (2024). Large language models in linguistic research: The pilot and the copilot. *Proceedings of CLIB 2024*.
6. Laakso, A., Kemell, K.-K., & Nurminen, J. K. (2024). Ethical issues in large language models: A systematic literature review. *TETHICS 2024*.
7. Milička, J., & Machálek, T. (2026). AI Corpus Linguist: More than a year of experience. *Proceedings of SIGHUM 2026*, 305–310.
8. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). Large language models: A survey.
9. Opitz, J., Wein, S., & Schneider, N. (2025). Natural Language Processing RELIES on linguistics. *Computational Linguistics*, 51(3).
10. Resnik, P. (2025). Large language models are biased because they are large language models. *Computational Linguistics*, 51(3).
11. Sousa-Silva, R. (2024). Fighting cyber-malice: A forensic linguistics approach to detecting AI-generated malicious texts. *Proceedings of the 1st International Conference on NLP & AI for Cyber Security*, 164–174.
12. Ungless, E. L., Vitsakis, N., Talat, Z., Garforth, J., Ross, B., Onken, A., Kasirzadeh, A., & Birch, A. (2025). The only way is ethics: A guide to ethical research with large language models. *Proceedings of the 31st International Conference on Computational Linguistics*, 8992–9005.
13. Xolmanova, Z. T. (2019). *Kompyuter lingvistikasi*. Tashkent: Alisher Navoiy Tashkent State University of Uzbek Language and Literature.
14. Xonnazarov, E. G. (2023). *Annotation of grammatical tense forms for the Uzbek language corpus*. PhD dissertation abstract. Tashkent.
15. Yulbarsov, O. O. (2025). *Creation of a sectorial neologism database using the Uzbek language web corpus*. PhD dissertation abstract. Tashkent.
16. Zasina, A. J. (2025). Automatic generation of corpus-based exercises using generative AI. *Proceedings of ROCLING 2025*, 80–86.
17. Zhao, W. (2025). Reconstructing stance in EFL doctoral thesis writing through generative artificial intelligence. *Humanities and Social Sciences Communications*, 12.