COMPUTER LINGUISTICS AND CORPUS LINGUISTICS IN AN AI-DRIVEN WORLD

Ergasheva Laziza Bahodir qizi
NRU TIIAME
Email: ergasheva_laziza_1998@email.com

Qabul qilindi: 09.10.2025
Ko'rib chiqildi: 10.10.2025
Tasdiqlandi: 11.10.2025

Abstract. This paper examines the intersection of computer linguistics and corpus linguistics in an AI-driven world, focusing on how artificial intelligence (AI), natural language processing (NLP), and big data analytics are transforming linguistic research and language technologies. The study highlights the integration of computational tools for large-scale corpus analysis, sentiment detection, and machine translation. It also explores how modern AI models, particularly deep learning and transformer-based architectures, enhance linguistic prediction and semantic understanding. While these advancements offer unprecedented analytical power, the paper emphasizes the ethical implications, data bias, and the necessity of linguistic transparency in AI systems. Findings suggest that the future of linguistics lies in the symbiosis between computational efficiency and human linguistic insight.

Introduction. The rapid development of artificial intelligence (AI) has redefined how language is analyzed, processed, and understood. Computer linguistics, traditionally concerned with the computational modeling of language, now merges seamlessly with corpus linguistics, which relies on extensive linguistic datasets to study patterns and meaning in authentic language use (McEnery & Hardie, 2019). Together, these fields underpin modern applications such as speech recognition, automatic translation, and conversational agents like ChatGPT. The integration of deep learning, natural language processing (NLP), and big data analytics allows computers to process language not merely as text, but as a dynamic and context-aware form of communication (Jurafsky & Martin, 2023). This paper explores how AI-driven corpus methodologies are reshaping

linguistic inquiry and how ethical and methodological frameworks must evolve alongside technological progress.

Literature Review. Corpus linguistics has traditionally provided empirical evidence for linguistic theories by analyzing authentic language use across large datasets (Biber et al., 2021). With AI integration, corpus analysis has expanded to include computational models capable of automatic tagging, parsing, and semantic interpretation. NLP techniques, particularly those using transformer architectures like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), enable machines to identify syntactic structures and pragmatic nuances with high precision. Moreover, computational linguistics bridges theoretical linguistics with real-world applications such as automated essay scoring, sentiment analysis, and cross-linguistic modeling (Bird, 2022). However, scholars caution against overreliance on AI-generated corpora, which may reproduce data bias and reduce linguistic diversity (Bender et al., 2021). Ethical AI research thus requires critical engagement with corpus design and model interpretability.

Methodology. This study employs a comparative analytical approach, synthesizing current research in computational linguistics and corpus-based methodologies. Peer-reviewed journal articles, conference proceedings, and reports published between 2018 and 2024 were reviewed using qualitative content analysis (Creswell & Poth, 2018). The selection criteria prioritized studies addressing NLP, corpus annotation, and AI model training for linguistic purposes. The analysis focuses on identifying major trends in the application of big data to linguistic inquiry, the evolution of corpus construction techniques, and the role of AI in advancing linguistic theory. Theoretical perspectives from both computational linguistics and sociolinguistics were integrated to ensure a balanced interpretation of technological and human-centered dimensions.

Findings and Discussion. The findings indicate that the convergence of AI and corpus linguistics has accelerated the automation of linguistic analysis. Large-scale corpora now support real-time language modeling, enabling advanced applications like automatic translation, emotion detection, and speech synthesis. For instance, transformer-based systems such as GPT-4 and BERT have demonstrated human-like capabilities in text generation and comprehension (Jurafsky & Martin, 2023). At the same time, big data analytics allows linguists to study sociolinguistic variation and discourse patterns across billions of words—something previously impossible without AI tools (Biber et al., 2021). Nevertheless, challenges persist. AI systems may inherit biases present in training data, leading to ethical concerns and skewed representations of minority

languages (Bender et al., 2021). Transparency in data collection and explainability in model outputs remain essential to maintaining linguistic integrity.

Another key finding is the increasing importance of **multimodal corpora**, which integrate text, audio, and visual elements to model human communication more holistically (McEnery & Hardie, 2019). Such developments push corpus linguistics beyond written text, aligning it with the broader goals of computational linguistics and cognitive science. AI-driven corpus tools are now capable of capturing context-sensitive meaning, facilitating more accurate discourse and translation analysis. These innovations suggest that future linguistics research will rely on collaborative frameworks combining computational algorithms with human interpretive skills.

Conclusion and Implications. In conclusion, the AI-driven era has redefined the boundaries of both computer linguistics and corpus linguistics. The integration of NLP, deep learning, and big data analytics has enabled unprecedented progress in the modeling and understanding of human language. However, this transformation also calls for new ethical and methodological standards that preserve linguistic diversity and transparency. The future of linguistics lies in a hybrid paradigm—where human interpretive insight complements the analytical power of AI. For linguists, educators, and AI developers alike, this collaboration promises deeper insights into language as both a computational and cultural phenomenon.

REFERENCES

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
2. Biber, D., Reppen, R., & Friginal, E. (2021). The Routledge handbook of corpus linguistics. Routledge.
3. Bird, S. (2022). Decolonising speech and language technology. Computer Speech & Language, 74, 101412.
4. Brown, T., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
5. Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches. Sage Publications.

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT, 4171–4186.
7. Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Draft version. Stanford University.
8. McEnery, T., & Hardie, A. (2019). Corpus linguistics: Method, theory and practice. Cambridge University Press.